On the Supermarket Model with Memory

Derek Yiu-Chung Wan

October 2012

A thesis submitted for the degree of Doctor of Philosophy

Department of Mathematics, London School of Economics and Political Science

Abstract

The *supermarket model with memory* consists of n single-server, infinite-capacity, first-in-first-out queues with service rate 1. The service times are independent. At all times, exactly one queue is distinguished as the *memory queue*. Customer arrivals form a Poisson process of rate λn, where $0 < \lambda < 1$. Upon arrival, each customer chooses an ordered list of d queues uniformly at random with replacement, adds the memory queue to the end of the list, and then joins the first shortest queue in the list. With the updated queue lengths, the first shortest queue in the list is then saved as the new memory queue. Our main contributions are to show that the system is rapidly mixing, and that with probability tending to 1 as $n \to \infty$, the maximum queue length in equilibrium is concentrated on two consecutive values which are $\frac{\ln \ln n}{\ln \alpha} + O(1)$, where $\alpha := d + \frac{1}{2} + \sqrt{d^2 + \frac{1}{4}}$.

Declaration

Acknowledgements

First and foremost, I would like to thank my supervisors Prof. Malwina Luczak and Prof. Graham Brightwell for their continual guidance and help. Without them, this project would not at all have been possible. Second, I would like to thank the Department of Mathematics for funding this research. Last but not least, I would like to thank my family for their unconditional support, especially my brother Michael who has always been here for me: you own.

Contents

Chapter 1

Introduction

1.1 The model

Throughout this thesis, let $d \geq 1$ be a fixed integer and let $0 < \lambda < 1$ be a fixed constant.

The *standard supermarket model* consists of n single-server, infinite-capacity, first-in-first-out queues with service rate 1. The service times are independent. Customer arrivals form a Poisson process of rate λn. Upon arrival, each customer chooses an ordered list of d queues uniformly at random with replacement, and then joins the first shortest queue in the list.

The *supermarket model with memory* distinguishes, at all times, exactly one queue as the *memory queue*. Upon arrival, each customer chooses an ordered list of d queues as above, then adds the memory queue to the end of the list, before joining the first shortest queue in the list. With the updated queue lengths, the first shortest queue in the list is then saved into memory. This model has been studied before in [21, 25, 13].

1.2 Statement of results

In this thesis, we have two main results, both of which are analogues of results proved by Luczak and McDiarmid [10] for the standard supermarket model. Let us introduce some notation so we may state these results.

Let $\mathcal{Q}_n := \mathbb{Z}_+^n \times \{1, \dots, n\}$, where $\mathbb{Z}_+ := \{0, 1, 2, \dots\}$. We will call the elements of $\mathcal{Q}_n$ *lengths vectors*. Although lengths vectors depend on n, we will only stress this dependence by adding a superscript n when we need to. For a lengths vector $x \in \mathcal{Q}_n$, we will write

$$x = \left(\left(x\left(1\right), \dots, x\left(n\right) \right), \xi \right),$$

and take $x\left(i\right)$ to be the length of queue i in x, and ξ to be the index of the memory queue in x. It follows that $\left\| x \right\|_1 := \sum_{i=1}^n x\left(i\right)$ is the number of customers in x, and that $\left\| x \right\|_\infty := \max\left(x\left(1\right), \dots, x\left(n\right) \right)$ is the maximum queue length in x.

The supermarket model with memory will be described by a continuous-time Markov jump process $\mathbf{X} = \left(X_t\right)_{t \geq 0}$ with state space $\mathcal{Q}_n$ as follows. For $t \geq 0$, we will write

$$X_t = \left(\left(X_t\left(1\right), \dots, X_t\left(n\right) \right), \Xi_t \right),$$

take $X_t(i)$ to be the length of queue i at time t, and take Ξ_t to be the index of the memory queue at time t. We will call $\mathbf{X}$ a *lengths process*, and take it to be right-continuous.

The *total variation distance* between distributions μ and ν on a common measurable space $(\Sigma, \mathcal{G})$ is

$$d_{\mathrm{TV}}(\mu, \nu) := \sup_{A \in \mathcal{G}} |\mu(A) - \nu(A)|.$$

For a random variable W, let $\mathcal{L}(W)$ denote the law of W. For $t \geq 0$ and $x \in \mathcal{Q}_n$, let $\mathcal{L}_x(X_t)$ denote the law of X_t conditional on $X_0 = x$. In Section 2.2, we will show that $\mathbf{X}$ is ergodic, and thus has a unique stationary distribution Π on $\mathcal{Q}_n$, and that

$$\lim_{t \to \infty} d_{\mathrm{TV}}(\mathcal{L}_x(X_t), \Pi) = 0$$

for all $x \in \mathcal{Q}_n$. Our first main result is that, under reasonable initial conditions, the convergence to equilibrium is very fast.

Theorem 1.1. *Let $c > \frac{\lambda}{1-\lambda}$. Then there exists $\eta = \eta(c) > 0$ such that the following holds. Let $n \geq 1$, and let $\mathbf{X}^n$ have any initial distribution. Then*

$$d_{\mathrm{TV}}(\mathcal{L}(X_t^n), \Pi^n) \leq ne^{-\eta t} + 2e^{-\eta\sqrt{n}} + \mathbb{P}(\|X_0^n\|_1 > cn) + \mathbb{P}(\|X_0^n\|_\infty > \eta t)$$

for all $t \geq 0$.

This result is directly analogous to Theorem 1.1 in [10] by Luczak and McDiarmid. Now define sequences $(a_i)_{i=0}^\infty$ and $(b_i)_{i=0}^\infty$ by setting $a_0 = b_0 = 1$ and

$$a_i := \lambda a_{i-1}^d b_{i-1}, \quad b_i := \frac{a_i^d b_{i-1}}{1 - d(a_{i-1} - a_i) a_i^{d-1}}, \tag{1.1}$$

for all $i \geq 1$. For $n \geq 1$, let

$$i_n^* := \min\left\{ i \geq 1 : a_i \leq \frac{\ln^2 n}{\sqrt{n}} \right\}. \tag{1.2}$$

We will show that

$$i_n^* = \frac{\ln \ln n}{\ln \alpha} + O(1), \quad \text{as } n \to \infty,$$

where

$$\alpha := d + \tfrac{1}{2} + \sqrt{d^2 + \tfrac{1}{4}}. \tag{1.3}$$

Note that $2d < \alpha < 2d + 1$. Our second main result is that with probability tending to 1 as $n \to \infty$, the equilibrium maximum queue length is concentrated on the two consecutive values $i_n^* - 1$ and i_n^*. Note that

$$\min\left(\tfrac{1}{2}d, 1\right) + \tfrac{1}{2}d - 1 = \min\left(d - 1, \tfrac{1}{2}d\right) \geq 0 \tag{1.4}$$

for all $d \geq 1$.

Theorem 1.2. *There exists $c > 0$ such that the following holds. Let $n \geq 1$, and let X^n have the equilibrium distribution for the lengths process. Then*

$$\mathbb{P}(\|X^n\|_\infty \neq i_n^* - 1 \text{ or } i_n^*) \leq \frac{c \ln^{4d+4} n}{n^{\min(d/2,1)+d/2-1+d/(2\alpha)}}.$$

This result is analogous to the first part of Theorem 1.3 in [10] by Luczak and McDiarmid; the analogue to the other part is our Theorem 7.6.

1.3 Literature review

In this section, we will give a brief review of the existing literature concerning the standard supermarket model and the supermarket model with memory.

First note that the standard supermarket model with $d = 1$ is equivalent to a system of n independent $M/M/1$ queues with arrival rate λ and service rate 1, and this is a well-understood system. Of relevance to us is Theorem 1.2 in [10] by Luczak and McDiarmid, which says that the equilibrium maximum queue length is about

$$\frac{\ln n}{\ln (1/\lambda)}, \tag{1.5}$$

and is not concentrated on a bounded range of values. More precisely, if $m = m(n)$, then with probability tending to 1 as $n \to \infty$, the equilibrium maximum queue length is at least (resp., at most) $m(n)$ if and only if $m(n) - \frac{\ln n}{\ln(1/\lambda)}$ tends to $-\infty$ (resp., $+\infty$) as $n \to \infty$.

The earliest work we know of on the standard supermarket model with $d \geq 2$ is by Mitzenmacher [23, 17] and Vvedenskaya, Dobrushin and Karpelevich [27], independently. For $i \geq 1$ and $t \geq 0$, let $u_i(t)$ denote the proportion of queues of length at least i at time t. Mitzenmacher heuristically argues that the $u_i(t)$ evolve in an almost deterministic fashion, and that in the limiting system as $n \to \infty$, they should satisfy the differential equations

$$\frac{du_i(t)}{dt} = \lambda \left[u_{i-1}(t)^d - u_i(t)^d \right] - \left[u_i(t) - u_{i+1}(t) \right], \tag{1.6}$$

for all $i \geq 1$. We explain his reasoning below. Mitzenmacher does show that

$$\mu = (\mu_i)_{i=1}^{\infty}, \quad \mu_i = \lambda^{1+d+\cdots+d^{i-1}}, \tag{1.7}$$

is a unique, attracting fixed point for (1.6). Mitzenmacher then heuristically argues that the μ_i should also be the expected proportion of queues of length at least i for the finite system (i.e., the standard supermarket model). This is based on the principle that the $u_i(t)$ in the limiting and in the finite systems have similar transition rates if they are near each other. Thus, if the two systems have the same initial state, then the trajectories of the $u_i(t)$ should not diverge by much over a short period of time, whence their difference over any period of time can be bounded by induction. Mitzenmacher then heuristically argues that the equilibrium queuing time of a customer is at most

$$\sum_{i=1}^{\infty} \lambda^{1+d+\cdots+d^{i-1}} + o(1),$$

and that the equilibrium maximum queue length is

$$\frac{\ln \ln n}{\ln d} + O(1), \tag{1.8}$$

with high probability. This technique of analysing a system through an idealised system

defined by differential equations is commonly known as the technique of *fluid limits*. The heuristic result (1.8) is later proved by Luczak and McDiarmid [10]. Hence, in conjunction with (1.5), one sees that taking $d = 2$ instead of $d = 1$ yields an exponential improvement in the maximum queue length, whilst taking larger values of $d \geq 2$ only yields constant factor improvements in the maximum queue length. This phenomenon is commonly known as the *power of two choices*.

Mitzenmacher's reasoning behind (1.6) is as follows. In the finite system, the expected change in $u_i(t)$ over a period of length Δt should be

$$\Delta u_i(t) = \tfrac{1}{n} \cdot \lambda n \Delta t \left[u_{i-1}(t)^d - u_i(t)^d \right] - \tfrac{1}{n} \cdot n \Delta t \left[u_i(t) - u_{i+1}(t) \right].$$

This is because there is an arrival with probability $\lambda n \Delta t$, and this customer joins a queue of length i with probability $u_{i-1}(t)^d - u_i(t)^d$, since he/she must select only queues of length at least $i-1$ but not only queues of length at least i; such an arrival increases $u_i(t)$ by $\frac{1}{n}$. On the other hand, there is a departure with probability $n \Delta t$, and this comes from a queue of length i with probability $u_i(t) - u_{i+1}(t)$; such a departure decreases $u_i(t)$ by $\frac{1}{n}$. Dividing by Δt and letting $\Delta t \to 0$ then yields (1.6).

In [27], Vvedenskaya, Dobrushin and Karpelevich also arrive at the differential equations (1.6). Let $\mathbf{Z} = (Z_t)_{t \geq 0}$ be a lengths process (as appropriately defined for the standard supermarket model) and let $w(0) = (w_i(0))_{i=1}^{\infty}$ be a sequence such that $u_i(Z_0) \to w_i(0)$ in probability as $n \to \infty$, for all $i \geq 1$. Let $w(t) = (w_i(t))_{i=1}^{\infty}$ denote the unique solution to (1.6) with initial state $w(0)$. Vvedenskaya et al. then show that

$$u_i(Z_t) \to w_i(t)$$

in probability as $n \to \infty$ uniformly on bounded time intervals. They also show that μ is a unique, attracting fixed point for (1.6).

In [7], Graham shows that the standard supermarket model is *chaotic* if it starts close to a suitable deterministic state, or is in equilibrium. That is, the queues in any finite subset of queues are asymptotically independent of each other, uniformly on bounded time intervals.

In [12], Luczak and Norris show three approximation theorems for the standard supermarket model: a law of large numbers, a jump process approximation, and a central limit theorem.

In [10], Luczak and McDiarmid show that the standard supermarket model is rapidly mixing, and that with probability tending to 1 as $n \to \infty$, the equilibrium maximum queue length is concentrated on two consecutive values which are

$$\frac{\ln \ln n}{\ln d} + O(1).$$

Thus, unlike the case $d = 1$, there is concentration on a bounded range of values. More precisely, let

$$\hat{i}_n := \min \left\{ i \geq 1 : \lambda^{1+d+\cdots+d^{i-1}} \leq \frac{\ln^2 n}{\sqrt{n}} \right\}.$$

Then the maximum queue length is concentrated on $\left\{ \hat{i}_n, \hat{i}_n + 1 \right\}$ if $d = 2$, and $\left\{ \hat{i}_n - 1, \hat{i}_n \right\}$

if $d \geq 3$. Since $\hat{i}_n = \frac{\ln \ln n}{\ln d} + O(1)$, this yields a rigorous proof of (1.8). We remark that these results, and the arguments used (which we will outline below), are similar to ours because we have based our work on [10].

To show rapid mixing of the lengths process, Luczak and McDiarmid show that two lengths processes with certain pairs of initial states can be coupled to coalesce rapidly. By using *path coupling* arguments (e.g., see [2]), a coupling only needs to be constructed for pairs of initial states which constitute the edge set of a certain graph structure on the state space. This is a considerably easier task than having to consider all possible pairs of initial states. The aforementioned coalescence is shown to occur rapidly by analysing a suitable random walk.

To show the result on the maximum queue length, Luczak and McDiarmid first establish some concentration of measure results for lengths processes. That is, if Z has the equilibrium distribution for the lengths process, then Lipschitz functions of Z are tightly concentrated around their means. This is done by using the bounded differences approach on two lengths processes: one initially empty and the other in equilibrium. Luczak and McDiarmid then apply these results to the functions which give the number of queues of length at least i, for all $i \geq 1$, which are Lipschitz.

Next, Luczak and McDiarmid derive (1.6). Using this and the concentration of measure results, they deduce that the equilibrium means $\mathbb{E}[u_i(Z)]$ closely follow a family of recurrence relations, in that there exists a constant $c_1 > 0$ such that

$$\sup_{i \geq 1} \left| \mathbb{E}[u_i(Z)] - \lambda \mathbb{E}[u_{i-1}(Z)]^d \right| \leq \frac{c_1 \ln^2 n}{\sqrt{n}}.$$

This suggests that $\mathbb{E}[u_i(Z)]$ should be close to $\hat{\mu}_i$, where the sequence $(\hat{\mu}_i)_{i=0}^{\infty}$ satisfies

$$\hat{\mu}_i - \lambda \hat{\mu}_{i-1}^d = 0$$

for all $i \geq 1$. This is easily solved to give $\hat{\mu}_i = \lambda^{1+d+\cdots+d^{i-1}}$, so $\mathbb{E}[u_i(Z)]$ should be close to $\lambda^{1+d+\cdots+d^{i-1}}$. Indeed, it is shown that there exists a constant $c_2 > 0$ such that

$$\sup_{i \geq 1} \left| \mathbb{E}[u_i(Z)] - \lambda^{1+d+\cdots+d^{i-1}} \right| \leq \frac{c_2 \ln^2 n}{\sqrt{n}}.$$

Using this and the concentration of measure results, it follows that $u_i(\cdot)$ is close to $\lambda^{1+d+\cdots+d^{i-1}}$. More precisely, it is shown that if $\mathbf{Z} = (Z_t)_{t \geq 0}$ is in equilibrium and $z, r > 0$, then

$$\mathbb{P}\left(\sup_{i \geq 1} \left| u_i(Z_t) - \lambda^{1+d+\cdots+d^{i-1}} \right| \geq \frac{z \ln^2 n}{\sqrt{n}} \text{ for some } 0 \leq t \leq n^r \right) = e^{-\Omega\left(\ln^2 n\right)}. \tag{1.9}$$

Here, we say that $f(n) = e^{-\Omega(g(n))}$ if there exists a positive constant $\eta > 0$ such that $f(n) \leq e^{-\eta g(n)}$ for all sufficiently large n.

Finally, Luczak and McDiarmid show two-point concentration of the equilibrium maximum queue length as follows. From (1.9) it easily follows that $\mathbb{P}\left(\|Z\|_{\infty} \leq \hat{i}_n - 2 \right) = e^{-\Omega\left(\ln^2 n\right)}$. By analysing an equilibrium lengths process and using (1.9) to control the

proportion of very long queues, Luczak and McDiarmid then show that

$$\mathbb{P}\left(\|Z\|_\infty \geq \hat{i}_n + z\right) = O\left(\left(\frac{\ln^{2d+2} n}{n^{d/2-1}}\right)^z\right),$$

for all $z \geq 1$. Hence, for $d \geq 3$, we have $\mathbb{P}\left(\|Z\|_\infty \geq \hat{i}_n + 1\right) \to 0$ as $n \to \infty$, and the proof for this case is complete. More complex arguments are needed for the case $d = 2$.

In [11], Luczak and McDiarmid quantify the rate of convergence of the equilibrium distribution of a typical queue length to its limiting distribution as $n \to \infty$. They also quantify the result that the standard supermarket model is chaotic by showing that the total variation distance between the joint law of a fixed set of queue lengths and the corresponding product law is essentially of order at most $\frac{1}{n}$.

There is much literature on variations of the standard supermarket model. In the survey [22], Mitzenmacher, Richa and Sitaraman reference the following variations: one where there are also *low-priority arrivals* which only join uniformly random queues, one where the queues have non-exponential service times (e.g., see [18, 28]), one where there are *thresholds* so that arriving customers who select a queue longer than the threshold will reselect (e.g., see [18, 28]), one where there is *load-stealing* so that any empty queue will find a non-empty queue to steal a customer from (e.g., see [23]), one where serviced customers recirculate into the system (e.g., see [4, 19]), and one where arriving customers join queues based on *stale* queue length information which is only updated periodically (e.g., see [3, 16, 20]). Mitzenmacher et al. also reference *Jackson networks* (e.g., see [14, 26]), where there are m nodes of n queues each, and arriving customers select a uniformly random node and then a shortest queue from within. In [1], Brightwell and Luczak study a variation where $d = d(n) \to \infty$ and $\lambda = \lambda(n) \uparrow 1$ are no longer fixed. Brightwell and Luczak identify, for suitable triples (n, d, λ), a subset $\mathcal{N}$ of the state space where the process remains for a long time in equilibrium, and show that the process is rapidly mixing when started from $\mathcal{N}$.

The supermarket model with memory is defined in [25] by Prabhakar and Shah, and in [21] by Mitzenmacher, Prabhakar and Shah. For $i \geq 1$ and $t \geq 0$, let $p_i(t)$ denote the probability that the memory queue has length at least i at time t. Mitzenmacher et al. heuristically argue that the length of the memory queue evolves so much faster than the $u_i(t)$ do, that the $p_i(\cdot)$ almost appear to be in equilibrium, and thus should satisfy

$$\frac{du_i(t)}{dt} = \lambda\left[u_{i-1}(t)^d p_{i-1}(t) - u_i(t)^d p_i(t)\right] - \left[u_i(t) - u_{i+1}(t)\right], \tag{1.10}$$

$$p_i(t) = u_i(t)^d p_{i-1}(t) + d\left(u_{i-1}(t) - u_i(t)\right) u_i(t)^{d-1} p_i(t), \tag{1.11}$$

for all $i \geq 1$. Mitzenmacher et al. do show that if $u = (u_i)_{i=1}^\infty$ is a fixed point for (1.10), then there exists $0 < c < 1$ such that

$$u_i \sim c^{\alpha^i},$$

where α is as defined in (1.3).

In [13], Luczak and Norris show how to approximate certain Markov chains with a fast, rapidly oscillating component alongside a slower, essentially deterministic component, by the solutions of differential equations. This includes the supermarket model with memory,

and the application of their method yields a rigorous derivation of (1.10) and (1.11). They also prove a natural monotonicity property of the supermarket model with memory.

As mentioned, our main contributions (and arguments) are analogous to those of [10]: we show that the supermarket model with memory is rapidly mixing, and that with probability tending to 1 as $n \to \infty$, the equilibrium maximum queue length is concentrated on two consecutive values which are

$$\frac{\ln \ln n}{\ln \alpha} + O(1).$$

1.4 Basic notation and results

In this section, we will outline the basic notation and results which we will assume the reader is familiar with. This material is adapted from [9, 24].

First we will discuss *discrete-time Markov chains*. Let Σ denote a countable set. A *discrete-time stochastic process* with *state space* Σ is a sequence of random variables $\mathbf{W} = (W_i)_{i=0}^{\infty}$, all defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in Σ.

A *stochastic matrix* on Σ is a real matrix $P = (p_{vw})_{v,w \in \Sigma}$ such that

1. $p_{vw} \geq 0$ for all $v, w \in \Sigma$, and
2. $\sum_{w \in \Sigma} p_{vw} = 1$ for all $v \in \Sigma$.

If

$$\mathbb{P}(W_{i+1} = w_{i+1} \mid W_0 = w_0, \ldots, W_i = w_i) = p_{w_i w_{i+1}}$$

for all $i \geq 0$ and $w_0, \ldots, w_{i+1} \in \Sigma$, then $\mathbf{W}$ is a *discrete-time Markov chain* with *transition matrix* P. Thus, if $\mathbf{W}$ is at a state $v \in \Sigma$, then it goes to the state $w \in \Sigma$ with probability p_{vw}, regardless of its history. A *stationary distribution* for $\mathbf{W}$ is a distribution $\Pi = (\pi_w)_{w \in \Sigma}$ on Σ such that $\Pi = \Pi P$, that is, such that $\pi_w = \sum_{v \in \Sigma} \pi_v p_{vw}$ for all $w \in \Sigma$.

Let $\mathbf{W}$ be a discrete-time Markov chain with state space Σ. For $w \in \Sigma$, let $\mathbb{P}_w(\cdot)$ and $\mathbb{E}_w[\cdot]$ denote the probability and expectation conditional on $W_0 = w$, respectively. If, for all $v, w \in \Sigma$, there exists $i = i(v, w) \geq 1$ such that $\mathbb{P}_v(W_i = w) > 0$, then $\mathbf{W}$ is *irreducible*. If $\gcd\{i \geq 1 : \mathbb{P}_w(W_i = w) > 0\} = 1$ for all $w \in \Sigma$, then $\mathbf{W}$ is *aperiodic*. For $A \subseteq \Sigma$, the *hitting time* of A is

$$H_A := \min\{i \geq 1 : W_i \in A\}.$$

For $w \in \Sigma$, we will write H_w instead of $H_{\{w\}}$. If $\mathbb{E}_w[H_w] < \infty$ for all $w \in \Sigma$, then $\mathbf{W}$ is *positive recurrent*. If $\mathbf{W}$ is irreducible, aperiodic and positive recurrent, then it is *ergodic*. It is well-known (e.g., see [9], Proposition 21.11) that if $\mathbf{W}$ is irreducible, then it is positive recurrent if and only if $\mathbb{E}_w[H_w] < \infty$ for some $w \in \Sigma$. Moreover (e.g., see [9], Theorem 21.14), if $\mathbf{W}$ is ergodic, then there exists a unique stationary distribution Π_W on Σ, and

$$\lim_{i \to \infty} d_{\mathrm{TV}}(\mathcal{L}(W_i, w), \Pi_W) = 0$$

for all $w \in \Sigma$; here $\mathcal{L}(W_i, w)$ is the law of W_i conditional on $W_0 = w$.

Next we will discuss *continuous-time Markov jump processes*. A *continuous-time stochastic process* with *state space* Σ is a family of random variables $\mathbf{W} = (W_t)_{t \geq 0}$, all defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in Σ. The processes we will study

are *right-continuous jump processes*, that is, processes such that for all $\omega \in \Omega$ and $t \geq 0$, there exists $\varepsilon = \varepsilon(\omega, t) > 0$ such that

$$W_t(\omega) = W_u(\omega)$$

for all $t \leq u \leq t + \varepsilon$. It is well-known (e.g., see [24], Section 6.6) that the probabilities concerning a right-continuous jump process may be determined from its *finite-dimensional distributions*, that is, the probabilities

$$\mathbb{P}\left(\bigcap_{k=0}^{i} \{W_{t_k} = w_k\}\right),$$

where $i \geq 0$, $0 \leq t_0 \leq \cdots \leq t_i$ and $w_0, \ldots, w_i \in \Sigma$. The *jump times* of $\mathbf{W}$ are the times $J_0 := 0$ and

$$J_i := \inf\left\{t > J_{i-1} : W_t \neq W_{J_{i-1}}\right\},$$

for all $i \geq 1$, where $\inf \emptyset = \infty$, and the *holding times* of $\mathbf{W}$ are the durations

$$D_i := \begin{cases} J_i - J_{i-1}, & \text{if } J_{i-1} < \infty, \\ \infty, & \text{if } J_{i-1} = \infty, \end{cases}$$

for all $i \geq 1$. Furthermore, the jump processes we will study are *non-explosive*, that is, processes such that

$$\mathbb{P}\left(\sup_{i \geq 0} J_i = \infty\right) = 1.$$

The *jump process* of $\mathbf{W}$ is the discrete-time stochastic process $\mathbf{W_J} = (W_{J_i})_{i=0}^{\infty}$.

A *Q-matrix* on Σ is a real matrix $Q = (q_{vw})_{v,w \in \Sigma}$ such that

1. $-\infty < q_{vv} \leq 0$ for all $v \in \Sigma$,
2. $q_{vw} \geq 0$ for all distinct $v, w \in \Sigma$, and
3. $\sum_{w \in \Sigma} q_{vw} = 0$ for all $v \in \Sigma$.

For $v \in \Sigma$, let $q_v := -q_{vv}$, and suppose that

1. whenever $\mathbf{W}$ is at a state $v \in \Sigma$ such that $q_v > 0$, it waits for an exponential time of rate $q_v > 0$ and then goes to the state $w \in \Sigma$ with probability $0 \leq \frac{q_{vw}}{q_v} \leq 1$, and
2. if $\mathbf{W}$ is at a state $v \in \Sigma$ such that $q_v = 0$, then it stays there forever.

Then $\mathbf{W}$ is a *continuous-time Markov jump process* with *generator matrix* Q. It follows that the jump process $\mathbf{W_J}$ of $\mathbf{W}$ is a discrete-time Markov chain with transition matrix $P = (p_{vw})_{v,w \in \Sigma}$, where for distinct $v, w \in \Sigma$,

$$p_{vw} := \begin{cases} \frac{q_{vw}}{q_v}, & \text{if } q_v > 0, \\ 0, & \text{if } q_v = 0, \end{cases} \qquad p_{vv} := \begin{cases} 0, & \text{if } q_v > 0, \\ 1, & \text{if } q_v = 0. \end{cases}$$

(It is straightforward to check that P is indeed a transition matrix.) In this case, we will call $\mathbf{W_J}$ the *jump chain* of $\mathbf{W}$. Moreover, it follows that for all $i \geq 1$ and $w_0, \ldots, w_{i-1} \in \Sigma$,

conditional on $W_{J_0} = w_0, \ldots, W_{J_{i-1}} = w_{i-1}$, the holding times $D_1, \ldots, D_i$ are independent exponential random variables with rates $q_{w_0}, \ldots, q_{w_{i-1}}$, respectively.

Let $\mathbf{W}$ be a continuous-time Markov jump process with state space Σ. For $w \in \Sigma$, let $\mathbb{P}_w(\cdot)$ and $\mathbb{E}_w[\cdot]$ denote the probability and expectation conditional on $W_0 = w$, respectively. If the jump chain $\mathbf{W_J}$ is irreducible, then $\mathbf{W}$ is *irreducible*. For $A \subseteq \Sigma$, the *hitting time* of A is

$$H_A := \min\{t \geq J_1 : W_t \in A\}.$$

For $w \in \Sigma$, we will write H_w instead of $H_{\{w\}}$. If $\mathbb{E}_w[H_w] < \infty$ for all $w \in \Sigma$, then $\mathbf{W}$ is *positive recurrent*. If $\mathbf{W}$ is irreducible and positive recurrent, then it is *ergodic*. It is well-known (e.g., see [9], Proposition 21.11) that if $\mathbf{W}$ is irreducible, then it is positive recurrent if and only if $\mathbb{E}_w[H_w] < \infty$ for some $w \in \Sigma$. Moreover (e.g., see [24], Theorem 3.8.1), if $\mathbf{W}$ is ergodic, then there exists a unique stationary distribution Π_W on Σ, and

$$\lim_{t\to\infty} d_{\mathrm{TV}}(\mathcal{L}(W_t, w), \Pi_W) = 0$$

for all $w \in \Sigma$; here $\mathcal{L}(W_t, w)$ is the law of W_t conditional on $W_0 = w$.

Next we will discuss *couplings*. A *coupling* of distributions μ and ν on a common measurable space $(\Sigma, \mathcal{G})$ is a pair of random variables (V, W), both defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, both taking values in Σ, and such that $\mathcal{L}(V) = \mu$ and $\mathcal{L}(W) = \nu$. It is well-known (e.g., see [9], Proposition 4.7) that

$$d_{\mathrm{TV}}(\mu, \nu) = \inf\{\mathbb{P}(V \neq W) : (V, W) \text{ is a coupling of } \mu \text{ and } \nu\}. \tag{1.12}$$

Now let P and P' be transition matrices on Σ and Σ', respectively. A *coupling* of discrete-time Markov chains with initial state $(v, w) \in \Sigma \times \Sigma'$ and transition matrices P and P' is a discrete-time stochastic process $(\mathbf{V}, \mathbf{W}) = ((V_i, W_i))_{i=0}^{\infty}$ with state space $\Sigma \times \Sigma'$ such that $\mathbf{V} = (V_i)_{i=0}^{\infty}$ is a discrete-time Markov chain with initial state v and transition matrix P, and $\mathbf{W} = (W_i)_{i=0}^{\infty}$ is a discrete-time Markov chain with initial state w and transition matrix P'. The couplings we will use are *Markovian couplings*, that is, couplings which are themselves Markov chains. We will also need to couple Markov chains where the initial states are random: in this case, we first sample the initial states of the chains, and then proceed as above. Furthermore, the couplings we will use, when coupling two copies of the same Markov chain, will keep the two processes together once they meet. That is, if $(\mathbf{V}, \mathbf{W})$ satisfies $V_k = W_k$ for some $k \geq 0$, then $V_i = W_i$ for all $i \geq k$. In this case, we will say that $\mathbf{V}$ and $\mathbf{W}$ have *coalesced*.

Analogous definitions are to be made with 'discrete-time Markov chain' replaced by 'continuous-time Markov jump process', and 'transition matrix' replaced by 'generator matrix'.

Finally, we will state three elementary results which we will be using several times in this thesis. The first concerns concentration of measure for sums of i.i.d. random variables.

Lemma 1.3 ([8], Theorem 5.11)**.** *Let W be an $\mathbb{R}$-valued random variable such that $\mathbb{E}[W] < \infty$ and $\mathbb{E}\left[e^{tW}\right] < \infty$ in some neighbourhood around $t = 0$, and let $c > \mathbb{E}[W]$. Then there exists $\eta > 0$ such that the following holds. Let $n \geq 1$, and let $W_1, \ldots, W_n$ be i.i.d. random*

variables each distributed like W. Then

$$\mathbb{P}\left(\sum_{i=1}^{n} W_i > cn\right) \leq e^{-\eta n}.$$

The second concerns concentration of measure for Poisson random variables. For $\mu > 0$, let $\mathrm{Po}(\mu)$ denote the Poisson distribution with mean μ.

Lemma 1.4 ([15], Theorem 2.3)**.** *Let $W \sim \mathrm{Po}(\mu)$. Then*

$$\mathbb{P}(|W - \mu| \geq \varepsilon\mu) \leq 2e^{-\frac{1}{3}\varepsilon^2\mu}$$

for all $0 \leq \varepsilon \leq 1$, and

$$\mathbb{P}(W \geq w) \leq 2^{-w}$$

for all $w \geq 2e\mu$.

The third is an easy algebraic result.

Lemma 1.5. *Let $0 \leq x \leq y$.*

1. *If $d \geq 2$, then $d(y-x)x^{d-1} \leq \frac{1}{2}y^d$.*
2. *We have $d(y-x)x^{d-1} \leq y^d - x^d$, with strict inequality if $0 < x < y$ and $d \geq 2$.*

1.5 Outline of thesis

The rest of this thesis is organised as follows. In Chapter 2, we will show that lengths processes are ergodic, and then establish some results about the equilibrium distribution for the lengths process. We will also prove an important random walk lemma which will be used to show rapid mixing of the lengths process. In Chapter 3, we will begin our proof of rapid mixing of the lengths process. We stop in Chapter 4 to establish some concentration of measure results, and then in Chapter 5 to analyse the equilibrium proportion of queues of length at least i, for all $i \geq 1$, and the equilibrium memory queue length. We will then complete the proof of rapid mixing of the lengths process in Chapter 6. In Chapter 7, we will analyse the equilibrium maximum queue length. In Chapter 8, we end with some concluding remarks and further ideas.

Chapter 2

Preliminary results

2.1 Notation for lengths processes

In this section, we will introduce some additional notation for lengths processes. In this thesis, at any time we will be referring to at most two lengths processes in detail. We have already introduced the notation

$$\mathbf{X} = (X_t)_{t\geq 0}\,, \quad X_t = ((X_t(1),\ldots,X_t(n)),\Xi_t)\,,$$

to denote a lengths process. A second lengths process, when we need it, will be denoted

$$\mathbf{Y} = (Y_t)_{t\geq 0}\,, \quad Y_t = ((Y_t(1),\ldots,Y_t(n)),\Theta_t)\,.$$

Similarly, we have already specified $x = ((x(1),\ldots,x(n)),\xi)$ to denote a lengths vector. A second lengths vector, when we need it, will be denoted $y = ((y(1),\ldots,y(n)),\theta)$.

We will use the following construction of a lengths process. Let a Poisson process $\mathbf{T}^{\mathrm{a}} = (T_i^{\mathrm{a}})_{i=1}^{\infty}$ of rate λn give the *arrival times*, and let $\mathbf{C}^{\mathrm{a}} = (C_i^{\mathrm{a}})_{i=1}^{\infty}$ be a corresponding sequence of ordered lists of d queues chosen uniformly at random with replacement, which we will call *choices*. For each arrival time T_i^{a}, we will take $C_i^{\mathrm{a}} = (C_i^{\mathrm{a}}(1),\ldots,C_i^{\mathrm{a}}(d))$ as the ordered list of d queues chosen by the arriving customer. The *candidates list* is the ordered list

$$\left(C_i^{\mathrm{a}}(1),\ldots,C_i^{\mathrm{a}}(d),\Xi_{T_i^{\mathrm{a}}-}\right),$$

where $\Xi_{T_i^{\mathrm{a}}-}$ is the memory queue immediately before T_i^{a}. We then add a customer to the first shortest queue in the candidates list, and with the updated queue lengths, save the first shortest queue in the candidates list into memory. Let a Poisson process $\mathbf{T}^{\mathrm{d}} = \left(T_i^{\mathrm{d}}\right)_{i=1}^{\infty}$ of rate n give the *potential departure times*, and let $\mathbf{S}^{\mathrm{d}} = \left(S_i^{\mathrm{d}}\right)_{i=1}^{\infty}$ be a corresponding sequence of queues selected uniformly at random, which we will call *selections*. For each potential departure time T_i^{d}, we will take S_i^{d} as the queue completing its service of any current customer. Thus, we remove a customer from S_i^{d} if it is currently non-empty. It follows that a potential departure time is not necessarily a jump time, since nothing happens if the selection is already empty.

The four processes $\mathbf{T}^{\mathrm{a}}$, $\mathbf{C}^{\mathrm{a}}$, $\mathbf{T}^{\mathrm{d}}$ and $\mathbf{S}^{\mathrm{d}}$ are defined on a common probability space $(\Omega,\mathcal{F},\mathbb{P})$ and are independent. Since the selections for each potential departure time are uniformly random, $\mathbf{T}^{\mathrm{d}}$ splits into n Poisson processes of rate 1, and the n queues have

independent exponential service times with rate 1. Thus, this construction of a lengths process is equivalent to that given in Section 1.1.

Let $T_0 := 0$, then enumerate all the arrival and potential departure times as $T_1, T_2, \ldots$. We will call $\mathbf{T} = (T_i)_{i=0}^{\infty}$ the *event times*. Let $(\mathcal{F}_t)_{t\geq 0}$ denote the natural filtration of $\mathcal{F}$ with respect to $\mathbf{X}$, and for a stopping time $T > 0$, let

$$\mathcal{F}_{T-} := \sigma\left\{\{A \cap \{t < T\} : t \geq 0, A \in \mathcal{F}_t\} \cup \mathcal{F}_0\right\}$$

denote the σ-field generated by all events before T. By definition, we have $\mathcal{F}_{0-} = \mathcal{F}_0$.

Suppose we have a procedure that, when given two lengths vectors, returns a pairing of the n queues in one lengths vector to the n queues in the other. Then we may construct a coupling of two lengths processes, $\mathbf{X}$ and $\mathbf{Y}$, using only this pairing procedure as follows. Let $\mathbf{X}$ and $\mathbf{Y}$ share the same arrival and potential departure times. For an event time T, pair the queues in X_{T-} and Y_{T-} using the given pairing procedure.

1. If T is an arrival time, let $C = (C(1), \ldots, C(d))$ be an ordered list of d queues chosen uniformly at random with replacement, then define $C' = (C'(1), \ldots, C'(d))$ by setting $C'(i)$ to be the queue paired with $C(i)$, for all $1 \leq i \leq d$. We will take C and C' as the choices for the arriving customer in $\mathbf{X}$ and in $\mathbf{Y}$, respectively.

2. If T is a potential departure time, let S be a queue in X_{T-} selected uniformly at random, then set S' to be the queue in Y_{T-} paired with S. We will take S and S' as the selections in $\mathbf{X}$ and in $\mathbf{Y}$, respectively.

It is easy to see that for an arrival time, C' is an ordered list of d queues chosen uniformly at random with replacement, and that for a potential departure time, S' is a queue in Y_{T-} selected uniformly at random. Thus, $\mathbf{Y}$ does have the distribution of a lengths process.

Note that $\mathbf{X}$ and $\mathbf{Y}$ necessarily jump together at each arrival time, but not necessarily together at each potential departure time, since a potential departure time is not necessarily a jump time.

For an arrival time, we will refer to the arriving customer in $\mathbf{X}$ as the $\mathbf{X}$-*customer*, his/her choices as the $\mathbf{X}$-*choices*, and the candidates list in $\mathbf{X}$ as the $\mathbf{X}$-*candidates list*. For a potential departure time, we will refer to the selection in $\mathbf{X}$ as the $\mathbf{X}$-*selection*. Analogous definitions are to be made for $\mathbf{Y}$.

2.2 Ergodicity and results about the equilibrium distribution

In this section, we will couple the supermarket model with memory with the standard supermarket model to show a certain stochastic domination result. We will then use this to establish that the former model is ergodic, and then extend some results about the equilibrium distribution for the standard supermarket model to the equilibrium distribution for the supermarket model with memory.

Let us introduce some additional notation for the standard supermarket model first. We will also call elements of $\mathbb{Z}_+^n$ *lengths vectors*. For a lengths vector $z \in \mathbb{Z}_+^n$, we will write

$$z = (z(1), \ldots, z(n)),$$

and take $z(i)$ to be the length of queue i in z. It follows that $\|z\|_1 := \sum_{i=1}^n z(i)$ is the number of customers in z, and that $\|z\|_\infty := \max(z(1), \ldots, z(n))$ is the maximum queue length in z.

The standard supermarket model will be described by a continuous-time Markov jump process $\mathbf{Z} = (Z_t)_{t \geq 0}$ with state space $\mathbb{Z}_+^n$ as follows. For $t \geq 0$, we will write

$$Z_t = (Z_t(1), \ldots, Z_t(n)),$$

and take $Z_t(i)$ to be the length of queue i at time t. We will call $\mathbf{Z}$ a *standard lengths process*, and take it to be right-continuous. It is well-known (e.g., see [6], Lemma 2.4) that $\mathbf{Z}$ is ergodic.

We may think of the memory queue as offering each arriving customer an additional opportunity to join a shorter queue, relative to the standard supermarket model. This implies that we should expect the former model to have more balanced queues, and thus shorter queues. In particular, we should also expect it to have fewer customers and a shorter maximum queue length. The same conclusions should also hold if each arriving customer in the standard supermarket model only makes d' choices, where $1 \leq d' \leq d$ is a fixed constant. All in all, we are looking to show that $\|X_t\|_1$ and $\|X_t\|_\infty$ are stochastically dominated by $\|Z_t\|_1$ and $\|Z_t\|_\infty$, respectively. This would allow us to extend the following results about the equilibrium distribution for the standard supermarket model, by Luczak and McDiarmid [10].

Lemma 2.1 ([10], Lemmas 2.4—2.6)**.** *Let $c > \frac{\lambda}{1-\lambda}$. Then there exist $\eta_1 = \eta_1(c) > 0$ and $\eta_2 > 0$ such that the following holds. Let $n \geq 1$.*

1. *Let Z have the equilibrium distribution for the standard lengths process. Then*

$$\mathbb{E}[\|Z\|_1] \leq \frac{\lambda n}{1-\lambda}, \quad \mathbb{P}(\|Z\|_1 > cn) \leq e^{-\eta_1 n},$$

and

$$\mathbb{P}(\|Z\|_\infty > r) \leq n e^{-\eta_2 r}$$

for all $r \geq 0$.

2. *Let $\mathbf{Z}$ have initial state $z \in \mathbb{Z}_+^n$, where $\|z\|_1 \leq cn$. Then*

$$\mathbb{P}(\|Z_t\|_1 > 2cn \text{ for some } 0 \leq t < e^{\eta_1 n}) \leq 2e^{-\eta_1 n}.$$

We will be coupling the supermarket model with memory with d choices and the standard supermarket model with d' choices, where $1 \leq d' \leq d$ is a fixed constant. However, we will only make use of the case where $d' = d$.

Let us *rank* the elements in a set of queues by length (in ascending order), and then if necessary, by queue index (also in ascending order). That is, given a set of k queues, we let the shortest queue with the least index have rank 1, and then repeat this for the ranks $2, \ldots, k$.

Definition 2.2. The *rank coupling* is the following coupling of a lengths process $\mathbf{X}$ and a standard lengths process $\mathbf{Z}$. Let $\mathbf{X}$ and $\mathbf{Z}$ share the same arrival and potential departure

times. For an event time T, pair the queues in X_{T-} and Z_{T-} by rank (from 1 to n).

1. If T is an arrival time, let the **X**-choices $C = (C(1), \ldots, C(d))$ be an ordered list of d queues chosen uniformly at random with replacement, then define the **Z**-choices $C' = (C'(1), \ldots, C'(d'))$ by setting $C'(i)$ to be the queue paired with $C(i)$, for all $1 \leq i \leq d'$.

2. If T is a potential departure time, let the **X**-selection be a queue in X_{T-} selected uniformly at random, then set the **Z**-selection to be the queue in Z_{T-} paired with the **X**-selection.

Remark. It is easy to see that for an arrival time, the **Z**-choices is an ordered list of d' queues chosen uniformly at random with replacement, and that for a potential departure time, the **Z**-selection is a queue in Z_{T-} selected uniformly at random. Thus, **Z** does have the distribution of a standard lengths process. This coupling was introduced by Graham in [7] to couple two standard lengths processes together.

Observe that, for a state w of either process (that is, for $w \in \mathcal{Q}_n \cup \mathbb{Z}_+^n$) and $i \geq 0$,

$$l_i(w) := \sum_{k=1}^{n} \mathbf{1}_{w(k) \geq i}, \quad f_i(w) := \sum_{k=i+1}^{\infty} l_k(w), \tag{2.1}$$

are the number of queues in w of length at least i, and the number of customers in w with at least i customers in front, respectively.

Lemma 2.3. *Let* **X** *and* **Z** *have initial states $x \in \mathcal{Q}_n, z \in \mathbb{Z}_+^n$, respectively, where $f_i(x) \leq f_i(z)$ for all $i \geq 0$, and let* **X** *and* **Z** *be coupled by the rank coupling. Then*

$$f_i(X_t) \leq f_i(Z_t)$$

for all $t \geq 0$ and $i \geq 0$. We have $\|X_t\|_1 \leq \|Z_t\|_1$ and $\|X_t\|_\infty \leq \|Z_t\|_\infty$ for all $t \geq 0$.

Remark. This proof is essentially the same as the proof of Theorem 4.1 in [7]. The only difference is that here, we also have a memory queue to deal with at each arrival time.

Proof. Clearly it suffices to show that $f_i(X_T) \leq f_i(Z_T)$ for the first event time $T > 0$ and all $i \geq 0$, so assume that

$$f_j(X_T) > f_j(Z_T) \tag{2.2}$$

for some $j \geq 0$. For a state w of either process and $1 \leq i \leq n$, let $r_i(w)$ denote the rank (from 1 to n) of queue i in w. There are now two cases to consider.

Case 1 T is an arrival time.

Let K and K' denote the length of the queue joined by the **X**- and **Z**-customer, respectively. Then (2.2) gives

$$f_j(x) + \mathbf{1}_{K \geq j} = f_j(X_T) > f_j(Z_T) = f_j(z) + \mathbf{1}_{K' \geq j}.$$

The hypothesis $f_j(x) \leq f_j(z)$ gives

$$f_j(x) = f_j(z), \quad K' < j \leq K.$$

These results, and the hypothesis $f_{j-1}(x) \leq f_{j-1}(z)$ (noting that $j > 0$), give

$$l_K(x) \leq l_j(x) = f_{j-1}(x) - f_j(x) \leq f_{j-1}(z) - f_j(z) = l_j(z) \leq l_{K'+1}(z). \tag{2.3}$$

Now let H denote the highest ranked queue in the **X**-candidates list $C = (C(1), \ldots, C(d+1))$, and let H' denote the highest ranked queue in the **Z**-choices $C' = (C'(1), \ldots, C'(d'))$. Recalling that the first d' coordinates of C and C' have the same rank, we have

$$r_H(x) = \min_{1 \leq i \leq d+1} r_{C(i)}(x) \leq \min_{1 \leq i \leq d'} r_{C'(i)}(z) = r_{H'}(z).$$

As highest ranked queues, H and H' are necessarily shortest queues in C and C', respectively, and thus have lengths K and K', respectively. The queues in x of length at least K have the ranks $n+1-l_K(x), \ldots, n-1, n$, so $r_H(x)$ must be one of these integers. Similarly, the the queues in z of length at least $K'+1$ have the ranks $n+1-l_{K'+1}(z), \ldots, n-1, n$, so $r_{H'}(z)$ must be strictly less than all these integers. Hence

$$l_{K'+1}(z) < n+1-r_{H'}(z) \leq n+1-r_H(x) \leq l_K(x).$$

This contradicts (2.3).

Case 2 T is a potential departure time.

Let K and K' denote the length of the **X**- and **Z**-selections, respectively. Then (2.2) gives

$$f_j(x) - \mathbf{1}_{K>j} = f_j(X_T) > f_j(Z_T) = f_j(z) - \mathbf{1}_{K'>j}.$$

The hypothesis $f_j(x) \leq f_j(z)$ gives

$$f_j(x) = f_j(z), \quad K \leq j < K'.$$

These results, and the hypothesis $f_{j+1}(x) \leq f_{j+1}(z)$, give

$$l_{K'}(z) \leq l_{j+1}(z) = f_j(z) - f_{j+1}(z) \leq f_j(x) - f_{j+1}(x) = l_{j+1}(x) \leq l_{K+1}(x). \tag{2.4}$$

Now let S and S' denote the **X**- and **Z**-selections, respectively. Recall that S and S' have lengths K and K', respectively. The queues in x of length at least $K+1$ have the ranks $n+1-l_{K+1}(x), \ldots, n-1, n$, so $r_S(x)$ must be strictly less than all these integers. Similarly, the queues in z of length at least K' have the ranks $n+1-l_{K'}(z), \ldots, n-1, n$, so $r_{S'}(z)$ must be one of these integers. Hence

$$l_{K+1}(x) < n+1-r_S(x) = n+1-r_{S'}(z) \leq l_{K'}(z).$$

This contradicts (2.4).

The last two inequalities in the statement of the lemma follow from the fact that

$$\|w\|_1 = f_0(w), \quad \|w\|_\infty = \min\{i \geq 0 : f_i(w) = 0\},$$

for all states w of either process. □

Lemma 2.3 implies that $\|X_t\|_1$ and $\|X_t\|_\infty$ are stochastically dominated by $\|Z_t\|_1$ and $\|Z_t\|_\infty$, respectively, as claimed. Now we will show that lengths processes are ergodic.

Lemma 2.4. $\mathbf{X}$ *is ergodic.*

Proof. Let $\mathbf{J} = (J_i)_{i=0}^\infty$ denote the jump times of $\mathbf{X}$, and for $1 \leq i \leq n$, let $\mathbf{0}_i := ((0,\dots,0),i)$ denote the empty state with memory queue i. We must show that $\mathbf{X}$ is irreducible and positive recurrent.

Part 1 Irreducibility.

Recall that $\mathbf{X}$ is irreducible if and only if its jump chain $\mathbf{X_J} = (X_{J_i})_{i=0}^\infty$ is. Instead of the jump chain, we will first consider $\mathbf{X_T} = (X_{T_i})_{i=0}^\infty$, the lengths process at the event times $\mathbf{T} = (T_i)_{i=0}^\infty$.

For $1 \leq j \leq n$, we will say that an event time is a j-*arrival* if it is an arrival time with choices $(j,\dots,j)$, and a j-*departure* if it is a potential departure time with selection j. For $1 \leq j,k \leq n$, we will also say that four consecutive event times are a (j,k)-*switcher* if they consist of a j-arrival, a k-arrival, a j-departure, and then a k-departure.

A (j,k)-switcher is a sequence of event times which, if $x(j) \geq x(k)$, will change the memory queue from j to k without changing the queue lengths. Thus, we are claiming that if

$$X_{T_i} = ((x(1),\dots,x(n)),j), \quad x(j) \geq x(k),$$

and if the event times $T_{i+1},\dots,T_{i+4}$ form a (j,k)-switcher, then $X_{T_{i+4}}$ will have memory queue k and the exact same queue lengths. To see this claim, note that a j-arrival gives the candidates list as $(j,\dots j,j)$, and thus queue j receives the customer and remains the memory queue. In particular, it is now strictly longer than queue k. A k-arrival then gives the candidates list as $(k,\dots,k,j)$, and thus queue k receives the customer and becomes the memory queue. Finally, a j- and k-departure will undo the changes in the queue lengths. This proves the claim.

Now let $x,y \in \mathcal{Q}_n$. We will construct a sequence of event times which will take $\mathbf{X}$ from x to y. We begin by taking $\mathbf{X}$ down to the empty state with memory queue ξ. Thus, if we have $x(j)$ j-departures, for $j = 1,\dots n$ in order, then

$$\begin{aligned} X_{T_{x(1)}} &= ((0,x(2),\dots,x(n)),\xi), \\ &\vdots \\ X_{T_u} &= ((0,\dots,0),\xi) = \mathbf{0}_\xi, \end{aligned}$$

where $u = \|x\|_1$. Next, we will take $\mathbf{X}$ to the empty state with memory queue θ. Thus, if we next have a (ξ,θ)-switcher, then

$$X_{T_{u+4}} = ((0,\dots,0),\theta) = \mathbf{0}_\theta.$$

Finally, we will restore the queue lengths one queue at a time. We will first build up queue $\theta+1$, then queue $\theta+2$, repeating this until we finish with queue θ. This will ensure that the switchers never go to a strictly longer queue. Thus, if we next have a $(j,j+1)$-switcher and then $(j+1)$-arrivals, exactly $y(j+1)$ of them, for $j = \theta,\dots,n,1\dots,\theta-1$ in order,

then

$$\begin{aligned} X_{T_{u+8+y(\theta+1)}} &= \left((0,\ldots,0,y\left(\theta+1\right),0,\ldots,0),\theta+1\right), \\ X_{T_{u+12+y(\theta+1)+y(\theta+2)}} &= \left((0,\ldots,0,y\left(\theta+1\right),y\left(\theta+2\right),0,\ldots,0),\theta+2\right), \\ &\vdots \\ X_{T_m} &= \left(\left(y\left(1\right),\ldots,y\left(n\right)\right),\theta\right) = y, \end{aligned}$$

where $m := \|x\|_1 + \|y\|_1 + 4\left(n+1\right)$.

Since each event time described here is a jump time, we have $T_i = J_i$ for all $1 \leq i \leq m$, and thus

$$\mathbb{P}_x\left(X_{J_m} = y\right) \geq \mathbb{P}_x\left(X_{T_m} = y\right) \geq \left(\frac{\lambda}{\lambda+1}\frac{1}{n^d}\right)^{\|y\|_1+2(n+1)} \left(\frac{1}{\lambda+1}\frac{1}{n}\right)^{\|x\|_1+2(n+1)} > 0.$$

This shows that the jump chain $\mathbf{X_J}$ is irreducible, and thus $\mathbf{X}$ is irreducible.

Part 2 Positive recurrence.

Let $A := \{\mathbf{0}_1,\ldots,\mathbf{0}_n\}$ denote the empty states in $\mathcal{Q}_n$, and let $\mathbf{0} := (0,\ldots,0)$ denote the empty state in $\mathbb{Z}_+^n$. Let $\mathbf{X}$ have an initial state in A, and let $\mathbf{Z}$ be a standard lengths process with initial state $\mathbf{0}$. Let $\mathbf{X}$ and $\mathbf{Z}$ be coupled by the rank coupling. Recall that $\mathbf{J} = (J_i)_{i=0}^{\infty}$ denotes the jump times of $\mathbf{X}$; let $\mathbf{J'} = (J'_i)_{i=0}^{\infty}$ denote the jump times of $\mathbf{Z}$. Then

$$H_A := \inf\left\{t \geq J_1 : X_t \in A\right\}, \quad H_{\mathbf{0}} := \inf\left\{t \geq J'_1 : Z_t = \mathbf{0}\right\},$$

are the hitting time of A by $\mathbf{X}$, and the hitting time of $\mathbf{0}$ by $\mathbf{Z}$, respectively. Since the initial states satisfy $f_i\left(x\right) = 0 = f_i\left(z\right)$ for all $i \geq 0$, Lemma 2.3 implies that

$$\|X_t\|_1 \leq \|Z_t\|_1$$

for all $t \geq 0$. But $J_1 = J'_1$, since $\mathbf{X}$ and $\mathbf{Z}$ are both initially at empty states and thus make their first jump at a common arrival time. It follows that

$$H_A \leq H_{\mathbf{0}}.$$

Using the well-known fact (e.g., see [6], Lemma 2.4) that $\mathbf{Z}$ is ergodic, we have

$$\max_{w\in A} \mathbb{E}_w\left[H_A\right] \leq \mathbb{E}_{\mathbf{0}}\left[H_{\mathbf{0}}\right] < \infty. \tag{2.5}$$

For the rest of this proof, we will be looking at the times when $\mathbf{X}$ lies in A. Enumerate $\{J_i : i \geq 0 \text{ and } X_{J_i} \in A\}$ into a sequence $(U_i)_{i=0}^{\infty}$. Now fix $a \in A$, then let

$$N_a := \inf\left\{i \geq 1 : X_{U_i} = a\right\}.$$

Then N_a is equal to the hitting time of a by the irreducible discrete-time Markov chain $\mathbf{X_U} = (X_{U_i})_{i=0}^{\infty}$, which has the finite state space A. It is well-known (e.g., see [9], Section 1.7) that an irreducible discrete-time Markov chain with a finite state space is positive

recurrent, so

$$\mathbb{E}_a[N_a] < \infty. \tag{2.6}$$

Moreover, the random variables H_a, N_a and U_i are all defined on the same probability space, and satisfy

$$H_a = U_{N_a} = \sum_{i=1}^{N_a} (U_i - U_{i-1}) = \sum_{i=1}^{\infty} (U_i - U_{i-1}) \mathbf{1}_{N_a \geq i}.$$

For $i \geq 1$, let $\mathcal{G}_i$ denote the σ-field generated by all events in $[0, U_i]$. By the Tower Rule, we have

$$\mathbb{E}_a[H_a] = \sum_{i=1}^{\infty} \mathbb{E}_a[(U_i - U_{i-1}) \mathbf{1}_{N_a \geq i}] = \sum_{i=1}^{\infty} \mathbb{E}_a[\mathbb{E}[(U_i - U_{i-1}) \mathbf{1}_{N_a \geq i} \mid \mathcal{G}_{i-1}]].$$

But $\mathbf{1}_{N_a \geq i}$ is $\mathcal{G}_{i-1}$-measurable, since $N_a \geq i$ if and only if $X_{U_1}, \ldots, X_{U_{i-1}} \neq a$. Hence, by (2.5) and (2.6), we have

$$\begin{aligned}
\mathbb{E}_a[H_a] &= \sum_{i=1}^{\infty} \mathbb{E}_a[\mathbb{E}[U_i - U_{i-1} \mid \mathcal{G}_{i-1}] \mathbf{1}_{N_a \geq i}] \\
&\leq \max_{w \in A} \mathbb{E}_w[U_1 - U_0] \, \mathbb{E}_a\left[\sum_{i=1}^{\infty} \mathbf{1}_{N_a \geq i}\right] \\
&= \max_{w \in A} \mathbb{E}_w[H_A] \, \mathbb{E}_a[N_a] < \infty.
\end{aligned}$$

Thus, $\mathbf{X}$ is positive recurrent. □

Having established that lengths processes are ergodic, we may now extend the aforementioned results for the equilibrium distribution for the standard supermarket model to the equilibrium distribution for the supermarket model with memory.

Lemma 2.5. *Let $c > \frac{\lambda}{1-\lambda}$. Then there exist $\eta_1 = \eta_1(c) > 0$ and $\eta_2 > 0$ such that the following holds. Let $n \geq 1$.*

1. *Let X have the equilibrium distribution for the lengths process. Then*

$$\mathbb{E}[\|X\|_1] \leq \frac{\lambda n}{1-\lambda}, \quad \mathbb{P}(\|X\|_1 > cn) \leq e^{-\eta_1 n},$$

and

$$\mathbb{P}(\|X\|_\infty > r) \leq n e^{-\eta_2 r}$$

for all $r \geq 0$.

2. *Let $\mathbf{X}$ have initial state $x \in \mathcal{Q}_n$, where $\|x\|_1 \leq cn$. Then*

$$\mathbb{P}(\|X_t\|_1 > 2cn \text{ for some } 0 \leq t < e^{\eta_1 n}) \leq 2e^{-\eta_1 n}.$$

Proof. This follows from Lemma 2.1 and Lemma 2.3. □

2.3 A random walk lemma

In this section, we will prove an important random walk lemma which we will use to show rapid mixing of the lengths process. First we will need the following result by McDiarmid [15] which concerns the concentration of sums of independent, bounded random variables.

Theorem 2.6 ([15], Theorem 2.5)**.** *Let $W_1, \ldots, W_n$ be independent random variables where W_i is $[b_i, c_i]$-valued for some $b_i < c_i$, for all $1 \leq i \leq n$. Then*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} W_i - \mathbb{E}\left[\sum_{i=1}^{n} W_i\right]\right| \geq w\right) \leq 2\exp\left(-\frac{2w^2}{\sum_{i=1}^{n}(c_i - b_i)^2}\right)$$

for all $w \geq 0$.

We will now prove a generalisation of Lemma 2.2 in [10] by Luczak and McDiarmid. This concerns a discrete-time random walk $\mathbf{R} = (R_i)_{i=0}^{\infty}$ on $\mathbb{R}$, with bounded increments, as follows. Let $a \leq b$, and suppose that:

1. $\mathbf{R}$ has negative drift with magnitude bounded away from 0, when it is above b, and
2. $\mathbf{R}$ will reach a (or a point below it) within a small number of steps with probability bounded away from 0, when it is at most b.

Then $\mathbf{R}$ should soon decrease to a, by the following reasoning. If $\mathbf{R}$ is above b, then it will first drift down towards b, by the first condition. At this point, $\mathbf{R}$ may make a finite sequence of steps and reach a. Since such a sequence occurs with probability bounded away from 0, after sufficiently many attempts, $\mathbf{R}$ will succeed at least once with high probability.

The events A_i in the lemma will be called the *background events*. Their relevance will be apparent when we apply the lemma to show rapid mixing of the lengths process.

Lemma 2.7. *Let $(\mathcal{G}_i)_{i=0}^{\infty}$ be a filtration, and let $(A_i)_{i=0}^{\infty}$ be a sequence of events such that $A_i \in \mathcal{G}_i$ for all $i \geq 0$. Let $\mathbf{R} = (R_i)_{i=0}^{\infty}$ be a random walk on $\mathbb{R}$ such that $Y_i := R_i - R_{i-1}$ is $\mathcal{G}_i$-measurable and $[-y, y]$-valued, for some $y > 0$ and all $i \geq 1$. Moreover, let $a < b$, $0 < p < 1$ and $q \geq 1$ be constants, and suppose that*

$$\mathbb{E}\left[Y_{i+1} \mid \mathcal{G}_i\right] \leq -p, \quad \text{on } A_i \cap \{R_i > b\}, \tag{2.7}$$

$$\mathbb{P}\left(\bigcup_{k=0}^{q} \{R_{i+k} \leq a\} \mid \mathcal{G}_i\right) \geq p, \qquad \text{on } A_i \cap \{R_i \leq b\}, \tag{2.8}$$

for all $i \geq 0$. Then there exists $\eta = \eta(a, b, p, q, y) > 0$ such that

$$\mathbb{P}\left(\bigcap_{i=1}^{m} (A_i \cap \{R_i > a\})\right) \leq 2e^{-\eta m} + \mathbb{P}(R_0 > \eta m)$$

for all $m \geq 1$.

Remark. This proof is based on the proof of Lemma 2.2 in [10], the result we are generalising. The main difference is that here we need a stronger result, namely Theorem 2.6, to show that the durations between the hitting times are not too long. For the original lemma, Luczak and McDiarmid use concentration of measure for binomial random variables.

Proof. Since the left-hand side is bounded by 1 and $\eta > 0$ may be arbitrarily small, it suffices to show the result for all sufficiently large m.

Let us ignore the background events A_i in the meantime; we shall see later that it is easy to incorporate them into the argument.

First define hitting times

$$I_0 := I_0' := \inf\{i \geq 0 : R_i \leq b\},$$

and

$$I_j := \inf\{i > I_{j-1} : R_i \leq b\}, \quad I_j' := \inf\left\{i > I_{j-1}' + q : R_i \leq b\right\},$$

for all $j \geq 1$. That is, let I_j be the first time after I_{j-1} when $\mathbf{R}$ is at most b, and let I_j' be the first time more than q time steps after I_{j-1}' when $\mathbf{R}$ is at most b. Note that

$$I_u' \leq I_{qu} \tag{2.9}$$

for all $u \geq 1$.

Now (2.8) and (2.9) give

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^{m}\{R_i > a\}\right) &\leq \mathbb{P}\left(\bigcap_{i=1}^{m}\{R_i > a\} \cap \left\{I_u' \leq m\right\}\right) + \mathbb{P}\left(I_u' > m\right) \\ &\leq (1-p)^u + \mathbb{P}\left(I_{qu} > m\right), \end{aligned} \tag{2.10}$$

for all $u \geq 1$. To see the first term in (2.10), note that if $R_1, \ldots, R_m > a$ and if $I_u' \leq m$, then

$$a < R_{I_j'} \leq b, \quad R_{I_j'+k} > a,$$

for all $0 \leq j < u$ and $1 \leq k \leq q$. That is, $\mathbf{R}$ fails to reach a within q steps from when it is at most b, at least u times. The probability of each of failure is bounded using (2.8).

Next we will show that the durations $I_{j+1} - I_j$ are not too long so that the term $\mathbb{P}\left(I_{qu} > m\right)$ is small. To do this, let $j \geq 0$ and $h \geq 2$, and suppose that $\{I_{j+1} - I_j > h\}$ holds. Now

$$\{I_{j+1} - I_j > h\} = \bigcap_{k=1}^{h}\left\{R_{I_j+k} > b\right\},$$

and the idea is that on the latter event, the increments $Y_{I_j+2}, \ldots, Y_{I_j+h}$ (and Y_{I_j+h+1}, though we will not need it) will each have negative expectation, by (2.7). Hence, for sufficiently large h, the $(j+1)^{\text{st}}$ hitting time I_{j+1} will occur within h time steps of the j^{th} hitting time I_j with high probability. Now for the details. If $R_{I_j+1}, \ldots, R_{I_j+h} > b$, then since $R_{I_j} \leq b$ and since each increment is at most y, we have

$$\sum_{k=2}^{h} Y_{I_j+k} = R_{I_j+h} - R_{I_j} - Y_{I_j+1} > b - b - y = -y \geq -\tfrac{1}{2}p\,(h-1), \tag{2.11}$$

if h is sufficiently large. Next we will relate this sum to a sum of independent random

variables with a straightforward conditional expectation on $\mathcal{G}_{I_j}$. For $j \geq 0$ and $k \geq 2$, let

$$Z_{j,k} := \begin{cases} Y_{I_j+k}, & \text{if } R_{I_j+1}, \ldots, R_{I_j+k} > b, \\ -p, & \text{otherwise.} \end{cases}$$

Then the Tower Rule gives

$$\mathbb{E}\left[Z_{j,k} \mid \mathcal{G}_{I_j}\right] = \mathbb{E}\left[\mathbb{E}\left[Z_{j,k} \mid \mathcal{G}_{I_j+k-1}\right] \mid \mathcal{G}_{I_j}\right] \leq -p,$$

for all $j \geq 0$ and $k \geq 2$. Hence, for all $h \geq 2$, $\sum_{k=2}^{h} Z_{j,k}$ is stochastically dominated by $\sum_{k=2}^{h} W_{j,k}$, a sum of independent $[-y, y]$-valued random variables where each $\mathbb{E}\left[W_{j,k} \mid \mathcal{G}_{I_j}\right] \leq -p$. Note that the $W_{j,k}$ need not be identically distributed. Let $\eta_1 = \eta_1(p, y) := \frac{1}{8}\left(\frac{p}{y}\right)^2$, then Theorem 2.6 (with $b_i = -y$, $c_i = y$ and $w = \frac{1}{2}p(h-1)$) gives

$$\begin{aligned} \mathbb{P}\left(\sum_{k=2}^{h} W_{j,k} \geq \mathbb{E}\left[\sum_{k=2}^{h} W_{j,k} \mid \mathcal{G}_{I_j}\right] + \tfrac{1}{2}p(h-1) \mid \mathcal{G}_{I_j}\right) &\leq 2\exp\left(-\frac{2\left(\frac{1}{2}p(h-1)\right)^2}{\sum_{k=2}^{h}(y+y)^2}\right) \\ &= 2e^{-\eta_1(h-1)}, \end{aligned} \tag{2.12}$$

for all $h \geq 2$. Returning to the event $\bigcap_{k=1}^{h}\left\{R_{I_j+k} > b\right\}$, on which we have $Z_{j,2} = Y_{I_j+2}, \ldots, Z_{j,h} = Y_{I_j+h}$, we see that (2.11) gives

$$\sum_{k=2}^{h} Z_{j,k} = \sum_{k=2}^{h} Y_{I_j+k} > -\tfrac{1}{2}p(h-1) \geq \mathbb{E}\left[\sum_{k=2}^{h} W_{j,k} \mid \mathcal{G}_{I_j}\right] + \tfrac{1}{2}p(h-1),$$

if h is sufficiently large. As $\sum_{k=2}^{h} Z_{j,k}$ is stochastically dominated by $\sum_{k=2}^{h} W_{j,k}$, (2.12) gives

$$\begin{aligned} \mathbb{P}\left(I_{j+1} - I_j > h \mid \mathcal{G}_{I_j}\right) &\leq \mathbb{P}\left(\bigcap_{k=1}^{h}\left\{R_{I_j+k} > b\right\} \mid \mathcal{G}_{I_j}\right) \\ &\leq \mathbb{P}\left(\sum_{k=2}^{h} W_{j,k} \geq \mathbb{E}\left[\sum_{k=2}^{h} W_{j,k} \mid \mathcal{G}_{I_j}\right] + \tfrac{1}{2}p(h-1) \mid \mathcal{G}_{I_j}\right) \\ &\leq 2e^{-\eta_1(h-1)}, \end{aligned}$$

if h is sufficiently large. Since the left-hand side is bounded by 1, for sufficiently small $\eta_2 = \eta_2(b, p, y) > 0$, we have

$$\mathbb{P}\left(I_{j+1} - I_j > h \mid \mathcal{G}_{I_j}\right) \leq 2e^{-\eta_2 h}$$

for all $h \geq 0$. Hence the durations $I_{j+1} - I_j$ are stochastically dominated by i.i.d. random variables H_j each distributed like an $\mathbb{N}$-valued random variable H such that $\mathbb{P}(H > h) = 2e^{-\eta_2 h}$ for all $h \geq 0$. Let

$$\gamma := \frac{3}{1 - e^{-\eta_2}}, \quad u := \left\lceil \frac{m}{4\gamma q} \right\rceil,$$

then we have

$$\mathbb{P}\left(I_{qu} - I_0 > \gamma qu\right) = \mathbb{P}\left(\sum_{j=1}^{qu}\left(I_j - I_{j-1}\right) > \gamma qu\right) \leq \mathbb{P}\left(\sum_{j=1}^{qu} H_j > \gamma qu\right).$$

Now H, the common distribution of the H_j, satisfies

$$\mathbb{E}\left[H\right] = \sum_{h=1}^{\infty} h\mathbb{P}\left(H = h\right) = \sum_{h=0}^{\infty} 2e^{-\eta_2 h} = \frac{2}{1 - e^{-\eta_2}} < \gamma,$$
$$\mathbb{E}\left[e^{\frac{1}{2}\eta_2 H}\right] = \sum_{h=1}^{\infty} e^{\frac{1}{2}\eta_2 h}\mathbb{P}\left(H = h\right) \leq \sum_{h=1}^{\infty} e^{\frac{1}{2}\eta_2 h} \cdot 2e^{-\eta_2(h-1)} < \infty,$$

so Lemma 1.3 (with $c = \gamma$) implies that there exists $\eta_3 = \eta_3\left(b, p, q, y\right) > 0$ such that

$$\mathbb{P}\left(I_{qu} - I_0 > \gamma qu\right) \leq e^{-\eta_3 qu} \tag{2.13}$$

for all $m \geq 1$.

Next we will show that I_0 is also not too long. The argument is similar to that used for the durations $I_{j+1} - I_j$, so we will be a little briefer here. Let $v := \left\lceil \frac{1}{4}m \right\rceil$, and suppose that $\left\{I_0 > v\right\} \cap \left\{R_0 \leq \frac{1}{16}pm\right\}$ holds. Now if $R_0 \leq \frac{1}{16}pm$ and $R_1, \ldots, R_v > b$, then we have

$$\sum_{k=2}^{v} Y_k = R_v - R_0 - Y_1 > b - \tfrac{1}{16}pm - y \geq -\tfrac{1}{2}p\left(v - 1\right), \tag{2.14}$$

if m is sufficiently large. For $k \geq 2$, let

$$Z_k := \begin{cases} Y_k, & \text{if } R_1, \ldots, R_k > b, \\ -p, & \text{otherwise.} \end{cases}$$

Then the Tower Rule gives $\mathbb{E}\left[Z_k\right] \leq -p$ for all $k \geq 2$. Hence, $\sum_{k=2}^{v} Z_k$ is stochastically dominated by $\sum_{k=2}^{h} W_k$, a sum of independent $[-y, y]$-valued random variables where each $\mathbb{E}\left[W_k\right] \leq -p$. Then Theorem 2.6 gives

$$\begin{aligned}\mathbb{P}\left(\sum_{k=2}^{v} W_k > \mathbb{E}\left[\sum_{k=2}^{v} W_k\right] + \tfrac{1}{2}p\left(v - 1\right)\right) &\leq 2\exp\left(-\frac{2\left(\frac{1}{2}p\left(v-1\right)\right)^2}{\sum_{k=2}^{v}\left(y + y\right)^2}\right) \\ &= 2e^{-\eta_1(v-1)}. \end{aligned} \tag{2.15}$$

Returning to the event $\left\{I_0 > v\right\} \cap \left\{R_0 \leq \frac{1}{16}pm\right\}$, on which we have $Z_2 = Y_2, \ldots, Z_v = Y_v$, we see that (2.14) gives

$$\sum_{k=2}^{v} Z_k = \sum_{k=2}^{v} Y_k > -\tfrac{1}{2}p\left(v - 1\right) \geq \mathbb{E}\left[\sum_{k=2}^{v} W_k\right] + \tfrac{1}{2}p\left(v - 1\right),$$

if m is sufficiently large. As $\sum_{k=2}^{v} Z_k$ is stochastically dominated by $\sum_{k=2}^{v} W_k$, (2.15) gives

$$\begin{aligned}\mathbb{P}\left(\{I_0 > v\} \cap \left\{R_0 \le \tfrac{1}{16}pm\right\}\right) &\le \mathbb{P}\left(\left\{R_0 \le \tfrac{1}{16}pm\right\} \cap \bigcap_{k=1}^{v} \{R_k > b\}\right) \\ &\le \mathbb{P}\left(\sum_{k=2}^{v} W_k > \mathbb{E}\left[\sum_{k=2}^{v} W_k\right] + \tfrac{1}{2}p\,(v-1)\right) \\ &\le 2e^{-\eta_1(v-1)},\end{aligned}$$

if m is sufficiently large. Since the left-hand side is bounded by 1, for sufficiently small $\eta_4 = \eta_4\,(b,p,y) > 0$, we have

$$\mathbb{P}\left(\{I_0 > v\} \cap \left\{R_0 \le \tfrac{1}{16}pm\right\}\right) \le 2e^{-\eta_4 m} \tag{2.16}$$

for all $m \ge 0$.

Now if $I_{qu} > m$, then

$$(I_{qu} - I_0) + I_0 = I_{qu} > m \ge \gamma q \left\lceil \frac{m}{4\gamma q} \right\rceil + \left\lceil \tfrac{1}{4}m \right\rceil = \gamma q u + v,$$

if m is sufficiently large. Hence, by (2.10), we have

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^{m} \{R_i > a\}\right) \le (1-p)^u &+ \mathbb{P}\left(I_{qu} - I_0 > \gamma q u\right) \\ &+ \mathbb{P}\left(\{I_0 > v\} \cap \left\{R_0 \le \tfrac{1}{16}pm\right\}\right) + \mathbb{P}\left(R_0 > \tfrac{1}{16}pm\right),\end{aligned}$$

if m is sufficiently large. By (2.13) and (2.16), there exists $\eta_5 = \eta_5\,(a,b,p,q,y) > 0$ such that

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^{m} \{R_i > a\}\right) &\le (1-p)^u + e^{-\eta_3 q u} + 2e^{-\eta_4 m} + \mathbb{P}\left(R_0 > \tfrac{1}{16}pm\right) \\ &\le 4e^{-\eta_5 m} + \mathbb{P}\left(R_0 > \eta_5 m\right) \\ &\le e^{-\frac{1}{2}\eta_5 m} + \mathbb{P}\left(R_0 > \tfrac{1}{2}\eta_5 m\right), \end{aligned} \tag{2.17}$$

if m is sufficiently large. Hence the result follows if $\eta \le \frac{1}{2}\eta_5$.

Now let us bring in the events A_i. For $i \ge 1$, let

$$Y_i' := Y_i \mathbf{1}_{A_{i-1}} - \max\left(b-a, p, y\right) \mathbf{1}_{\overline{A_{i-1}}}, \quad R_i' := R_0 + \sum_{j=1}^{i} Y_i'.$$

Then

$$\begin{aligned}\mathbb{E}\left[Y_{i+1}' \mid \mathcal{G}_i\right] &\le -p, \quad \text{on } \{R_i > b\}, \\ \mathbb{P}\left(\bigcup_{k=0}^{q} \left\{R_{i+k}' \le a\right\} \mid \mathcal{G}_i\right) &\ge p, \qquad \text{on } \{R_i \le b\},\end{aligned}$$

for all $i \ge 0$. These are obvious if $\overline{A_i}$ holds, and easily follow from (2.7) and (2.8) if A_i

holds. Hence, (2.17) applies to $(R'_i)_{i=0}^{\infty}$, giving

$$\mathbb{P}\left(\bigcap_{i=1}^{m}(A_{i-1}\cap\{R_i>a\})\right)\leq\mathbb{P}\left(\bigcap_{i=1}^{m}\{R'_i>a\}\right)\leq e^{-\frac{1}{2}\eta_5 m}+\mathbb{P}\left(R_0>\tfrac{1}{2}\eta_5 m\right),$$

if m is sufficiently large. To see the first inequality, note that if $A_0,\ldots,A_{m-1}$ all hold, then we have $Y'_i=Y_i$ for all $1\leq i\leq m$. It follows that $R'_i=R_i$ for all $1\leq i\leq m$. □

The following lemma is the main result of this section. It is an application of Lemma 2.7 and concerns a discrete-time random walk $\mathbf{S}=(S_i)_{i=0}^{\infty}$ on $\mathbb{Z}_+$, with increments in $\{-1,0,1\}$, as follows. At all times, $\mathbf{S}$ is to be described as either being *good* or *bad*; let the event B_i denote the event that $\mathbf{S}$ is *good* at time i. Let $\kappa\geq 1$, and suppose that:

1. $\mathbf{S}$ will either become good or increase, with probability bounded away from 0, when it is bad,
2. $\mathbf{S}$ will remain good and decrease, with probability bounded away from 0, when it is good,
3. $\mathbf{S}$ is more likely to remain good and decrease than to increase, when it is good and above κ, and
4. $\mathbf{S}$ will become good without changing value, with probability close to 1, when it is bad and above κ.

Then there should be the rapid occurrence of a time when $\mathbf{S}$ is simultaneously good and takes the value 0, by the following reasoning. Let $\mathbf{R}$ be the random walk equal to $\mathbf{S}$ if $\mathbf{S}$ is good, and $\mathbf{S}$ plus a *penalty* $0<\beta<1$ if $\mathbf{S}$ is bad. That is, let

$$R_i:=S_i+\beta\mathbf{1}_{\overline{B_i}},$$

for $i\geq 0$. Thus, $\mathbf{R}$ is 0 if and only if $\mathbf{S}$ is good and takes the value 0, and in particular, $\mathbf{R}$ should soon decrease to 0 by Lemma 2.7 and the following reasoning.

1. Condition (2.7) requires that $\mathbf{R}$ has negative drift with magnitude bounded away from 0, when it is above $b:=\kappa+\beta$. Now if $\mathbf{R}$ is above $\kappa+\beta$, then $\mathbf{S}$ is above κ. There are now two cases to consider.
   (a) If $\mathbf{S}$ is good, then the third condition implies that $\mathbf{S}$ will have negative drift, whence $\mathbf{R}$ will also have negative drift.
   (b) If $\mathbf{S}$ is bad, then the fourth condition implies that $\mathbf{S}$ will become good without changing value, with probability close to 1. This will represent a decrease in $\mathbf{R}$, as the penalty will no longer apply. Moreover, the aforementioned probability will be so close to 1 that this will represent a negative drift in $\mathbf{R}$.
2. Condition (2.8) requires that $\mathbf{R}$ will reach $a:=0$ within a small number of steps with probability bounded away from 0, when it is at most $\kappa+\beta$. Now if $\mathbf{R}$ is at most $\kappa+\beta$, then $\mathbf{S}$ is at most κ. There are two parts to the argument here.

(a) If $\mathbf{S}$ is good and at most $\kappa+1$, then the second condition implies that $\mathbf{S}$ will remain good and decrease to 0 within at most $\kappa+1$ steps, with probability bounded away from 0. That is, $\mathbf{R}$ will decrease to 0 within at most $\kappa+1$ steps, with probability bounded away from 0.

(b) If $\mathbf{S}$ is bad and at most κ, then the first condition implies that $\mathbf{S}$ will either have become good or have remained bad and increased to $\kappa+1$ within at most $\kappa+1$ steps, with probability bounded away from 0. If we are in the latter case, then using the fourth condition, we see that $\mathbf{S}$ will be good and at most $\kappa+1$ within at most $\kappa+2$ steps, with probability bounded away from 0. At this point, the first part of the argument applies.

The two parts together imply that $\mathbf{R}$ will decrease to 0 within at most $q := 2\kappa+3$ steps, with probability bounded away from 0.

The events A_i in the lemma will also be called the *background events*. Again, their relevance will be apparent when we apply the lemma to show rapid mixing of the lengths process.

Lemma 2.8. *Let $(\mathcal{G}_i)_{i=0}^{\infty}$ be a filtration, and let $(A_i)_{i=0}^{\infty}$ and $(B_i)_{i=0}^{\infty}$ be sequences of events such that $A_i, B_i \in \mathcal{G}_i$ for all $i \geq 0$. Let $\mathbf{S} = (S_i)_{i=0}^{\infty}$ be a random walk on $\mathbb{Z}_+$ such that $Z_i := S_i - S_{i-1}$ is $\mathcal{G}_i$-measurable and $\{-1,0,1\}$-valued, for all $i \geq 1$. Moreover, let $0 < \delta < \frac{1}{2}$ and $\kappa \geq 1$ be constants, and suppose that*

$$\mathbb{P}\left(B_{i+1} \cup \{Z_{i+1}=1\} \mid \mathcal{G}_i\right) \geq \delta, \qquad \text{on } A_i \cap \overline{B_i}, \tag{2.18}$$

$$\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1}=-1\} \mid \mathcal{G}_i\right) \geq \delta, \qquad \text{on } A_i \cap B_i \cap \{S_i > 0\}, \tag{2.19}$$

$$\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1}=-1\} \mid \mathcal{G}_i\right) \geq \mathbb{P}\left(Z_{i+1}=1 \mid \mathcal{G}_i\right) + \delta, \qquad \text{on } A_i \cap B_i \cap \{S_i > \kappa\}, \tag{2.20}$$

$$\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1}=0\} \mid \mathcal{G}_i\right) \geq 1 - \tfrac{1}{2}\delta, \qquad \text{on } A_i \cap \overline{B_i} \cap \{S_i > \kappa\}, \tag{2.21}$$

for all $i \geq 0$. Then there exists $\eta = \eta(\delta,\kappa) > 0$ such that

$$\mathbb{P}\left(\bigcap_{i=1}^{m} \left(A_{i-1} \cap \left(\overline{B_i} \cup \{S_i > 0\}\right)\right)\right) \leq 2e^{-\eta m} + \mathbb{P}\left(S_0 > \eta m\right)$$

for all $m \geq 1$.

Proof. Since the left-hand side is bounded by 1 and $\eta > 0$ may be arbitrarily small, it suffices to show the result for all sufficiently large m.

We have already outlined the idea behind this proof. We will take the penalty β to be $\frac{3}{4}\delta$. We will show that the hypotheses of Lemma 2.7 hold with the same filtration $(\mathcal{G}_i)_{i=0}^{\infty}$, the same sequence of events $(A_i)_{i=0}^{\infty}$, the constants

$$a = 0, \quad b = \kappa + \beta, \quad p = \delta^{2\kappa+3}, \quad q = 2\kappa+3, \quad y = 1+\beta,$$

and the random walk $\mathbf{R} = (R_i)_{i=0}^{\infty}$ where

$$R_i := S_i + \beta \mathbf{1}_{\overline{B_i}}.$$

It is easy to see that $p \leq \frac{1}{16}\delta$. As required by Lemma 2.7,

$$Y_i := R_i - R_{i-1} = Z_i + \beta\left(\mathbf{1}_{\overline{B_i}} - \mathbf{1}_{\overline{B_{i-1}}}\right)$$

is $\mathcal{G}_i$-measurable and $[-y, y]$-valued, for all $i \geq 1$. There are now two conditions to verify: (2.7) and (2.8).

1. For condition (2.7), suppose that $A_i \cap \{R_i > b\}$ holds, so that $S_i > \kappa$. There are now two cases to consider. If B_i holds, then the next increment in $\mathbf{R}$ is

   $$Y_{i+1} = Z_{i+1} + \beta\mathbf{1}_{\overline{B_{i+1}}} \in \{-1, -1+\beta, 0, \beta, 1, 1+\beta\},$$

   whence we may write

   $$\begin{aligned}\mathbb{E}\left[Y_{i+1} \mid \mathcal{G}_i\right] &\leq -\mathbb{P}\left(Y_{i+1} = -1 \mid \mathcal{G}_i\right) + \beta\mathbb{P}\left(Y_{i+1} = \beta \mid \mathcal{G}_i\right) + (1+\beta)\,\mathbb{P}\left(Y_{i+1} \geq 1 \mid \mathcal{G}_i\right) \\ &\leq -\mathbb{P}\left(Y_{i+1} = -1 \mid \mathcal{G}_i\right) + \beta + \mathbb{P}\left(Y_{i+1} \geq 1 \mid \mathcal{G}_i\right), \quad \text{on } A_i \cap B_i \cap \{R_i > b\}.\end{aligned}$$

   By (2.20), we have

   $$\begin{aligned}\mathbb{E}\left[Y_{i+1} \mid \mathcal{G}_i\right] &\leq -\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = -1\} \mid \mathcal{G}_i\right) + \beta + \mathbb{P}\left(Z_{i+1} = 1 \mid \mathcal{G}_i\right) \\ &\leq -\delta + \tfrac{3}{4}\delta \leq -p, \quad \text{on } A_i \cap B_i \cap \{R_i > b\}.\end{aligned}$$

   On the other hand, if $\overline{B_i}$ holds, then the next increment in $\mathbf{R}$ is

   $$Y_{i+1} = Z_{i+1} - \beta\mathbf{1}_{B_{i+1}} \in \{-1-\beta, -1, -\beta, 0, 1-\beta, 1\},$$

   whence we may write

   $$\begin{aligned}\mathbb{E}\left[Y_{i+1} \mid \mathcal{G}_i\right] &\leq -\beta\mathbb{P}\left(Y_{i+1} = -\beta \mid \mathcal{G}_i\right) + \mathbb{P}\left(Y_{i+1} \neq -\beta \mid \mathcal{G}_i\right) \\ &= 1 - (1+\beta)\,\mathbb{P}\left(Y_{i+1} = -\beta \mid \mathcal{G}_i\right), \quad \text{on } A_i \cap \overline{B_i} \cap \{R_i > b\}.\end{aligned}$$

   By (2.21) and the fact that $\delta < \frac{1}{2}$, we have

   $$\begin{aligned}\mathbb{E}\left[Y_{i+1} \mid \mathcal{G}_i\right] &\leq 1 - (1+\beta)\,\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = 0\} \mid \mathcal{G}_i\right) \\ &\leq 1 - \left(1 + \tfrac{3}{4}\delta\right)\left(1 - \tfrac{1}{2}\delta\right) \\ &= -\left(\tfrac{1}{4} - \tfrac{3}{8}\delta\right)\delta \leq -\tfrac{1}{16}\delta \leq -p, \quad \text{on } A_i \cap \overline{B_i} \cap \{R_i > b\}.\end{aligned}$$

   Thus (2.7) holds.

2. For condition (2.8), suppose that $A_i \cap \{R_i \leq b\}$ holds, so that $S_i \leq \kappa + \beta$. Since $\mathbf{S}$ only takes values in $\mathbb{Z}_+$, it follows that $S_i \leq \kappa$. As mentioned when we outlined the idea behind the proof, the first part of the argument deals with the case where $\mathbf{S}$ is good and at most $\kappa + 1$. For $s \geq 0$, let

   $$E_{i,s} := B_i \cap \{S_i = s\}$$

denote the event that $\mathbf{S}$ is good and takes the value at time i. By (2.19), we have

$$\delta \leq \mathbb{E}\left[\mathbf{1}_{B_{i+1}}\mathbf{1}_{Z_{i+1}=-1} \mid \mathcal{G}_i\right] \leq \mathbb{E}\left[\mathbf{1}_{E_{i+1,s-1}} \mid \mathcal{G}_i\right], \quad \text{on } A_i \cap E_{i,s}, \tag{2.22}$$

for all $i \geq 0$ and $s \geq 1$. Multiplying through by δ, bounding δ using (2.22) with i replaced by $i+1$ and s replaced by $s-1$, and then using the Tower Rule, we have

$$\begin{aligned}
\delta^2 &\leq \mathbb{E}\left[\mathbf{1}_{E_{i+1,s-1}}\delta \mid \mathcal{G}_i\right] \\
&\leq \mathbb{E}\left[\mathbf{1}_{E_{i+1,s-1}}\mathbb{E}\left[\mathbf{1}_{E_{i+2,s-2}} \mid \mathcal{G}_{i+1}\right] \mid \mathcal{G}_i\right] \\
&= \mathbb{E}\left[\mathbf{1}_{E_{i+1,s-1}}\mathbf{1}_{E_{i+2,s-2}} \mid \mathcal{G}_i\right], \quad \text{on } A_i \cap E_{i,s}.
\end{aligned}$$

Similarly, by (2.22) and induction, it is straightforward to see that

$$\delta^s \leq \mathbb{E}\left[\mathbf{1}_{E_{i+1,s-1}} \dots \mathbf{1}_{E_{i+s,0}} \mid \mathcal{G}_i\right] \leq \mathbb{E}\left[\mathbf{1}_{R_{i+s}=0} \mid \mathcal{G}_i\right], \quad \text{on } A_i \cap E_{i,s},$$

for all $s \geq 1$. In particular, we have

$$\delta^{\kappa+1} \leq \mathbb{E}\left[\mathbf{1}_{R_{i+s}=0} \mid \mathcal{G}_i\right] \leq \mathbb{E}\left[\mathbf{1}_{\bigcup_{k=0}^{\kappa+1} R_{i+k}=0} \mid \mathcal{G}_i\right], \quad \text{on } A_i \cap E_{i,s}, \tag{2.23}$$

for all $0 \leq s \leq \kappa+1$.

The second part of the argument deals with the case where $\mathbf{S}$ is bad and at most κ. For $s \geq 0$, let

$$F_{i,s} := \overline{B_i} \cap \{S_i = s\}$$

denote the event that $\mathbf{S}$ is bad and takes the value s at time i. By (2.18), we have

$$\begin{aligned}
\delta &\leq \mathbb{E}\left[\mathbf{1}_{B_{i+1}\cup\{Z_{i+1}=1\}} \mid \mathcal{G}_i\right] \\
&= \mathbb{E}\left[\mathbf{1}_{B_{i+1}} + \mathbf{1}_{\overline{B_{i+1}}}\mathbf{1}_{Z_{i+1}=1} \mid \mathcal{G}_i\right] \\
&\leq \mathbb{E}\left[\mathbf{1}_{B_{i+1}} + \mathbf{1}_{F_{i+1,s+1}} \mid \mathcal{G}_i\right], \quad \text{on } A_i \cap F_{i,s},
\end{aligned} \tag{2.24}$$

for all $i, s \geq 0$, and by (2.21), we have

$$\delta \leq 1 - \tfrac{1}{2}\delta \leq \mathbb{E}\left[\mathbf{1}_{B_{i+1}}\mathbf{1}_{Z_{i+1}=0} \mid \mathcal{G}_i\right] \leq \mathbb{E}\left[\mathbf{1}_{E_{i+1,s}} \mid \mathcal{G}_i\right], \quad \text{on } A_i \cap F_{i,s}, \tag{2.25}$$

for all $i \geq 0$ and $s > \kappa$. Multiplying (2.24) through by δ, bounding δ using (2.24) with i replaced by $i+1$ and s replaced by $s+1$, and then using the Tower Rule, we have

$$\begin{aligned}
\delta^2 &\leq \mathbb{E}\left[\mathbf{1}_{B_{i+1}} \mid \mathcal{G}_i\right] + \mathbb{E}\left[\mathbf{1}_{F_{i+1,s+1}}\delta \mid \mathcal{G}_i\right]. \\
&\leq \mathbb{E}\left[\mathbf{1}_{B_{i+1}} \mid \mathcal{G}_i\right] + \mathbb{E}\left[\mathbf{1}_{F_{i+1,s+1}}\mathbb{E}\left[\mathbf{1}_{B_{i+2}} + \mathbf{1}_{F_{i+2,s+2}} \mid \mathcal{G}_{i+1}\right] \mid \mathcal{G}_i\right] \\
&\leq \mathbb{E}\left[\mathbf{1}_{B_{i+1}} \mid \mathcal{G}_i\right] + \mathbb{E}\left[\mathbf{1}_{F_{i+1,s+1}}\mathbf{1}_{B_{i+2}} \mid \mathcal{G}_i\right] \\
&\qquad + \mathbb{E}\left[\mathbf{1}_{F_{i+1,s+1}}\mathbf{1}_{F_{i+2,s+2}} \mid \mathcal{G}_i\right], \quad \text{on } A_i \cap F_{i,s}.
\end{aligned}$$

Similarly, by (2.24) and induction, it is straightforward to see that if $r = r(s) :=$

$\kappa+1-s$, then

$$\begin{aligned}\delta^r &\le \sum_{j=1}^{r} \mathbb{E}\left[\mathbf{1}_{F_{i+1,s+1}} \dots \mathbf{1}_{F_{i+j-1,s+j-1}} \mathbf{1}_{B_{i+j}} \mid \mathcal{G}_i\right] \\ &\qquad + \mathbb{E}\left[\mathbf{1}_{F_{i+1,s+1}} \dots \mathbf{1}_{F_{i+r,\kappa+1}} \mid \mathcal{G}_i\right], \quad \text{on } A_i \cap F_{i,s},\end{aligned}$$

for all $0 \le s \le \kappa$. Multiplying through by δ, bounding δ using (2.25) with i replaced by $i+r$ and s replaced by $\kappa+1$, and then using the Tower Rule, we have

$$\begin{aligned}\delta^{r+1} &\le \sum_{j=1}^{r} \mathbb{E}\left[\mathbf{1}_{B_{i+j}} \mid \mathcal{G}_i\right] + \mathbb{E}\left[\mathbf{1}_{F_{i+r,\kappa+1}} \delta \mid \mathcal{G}_i\right] \\ &\le \sum_{j=1}^{r} \mathbb{E}\left[\mathbf{1}_{B_{i+j}} \mid \mathcal{G}_i\right] + \mathbb{E}\left[\mathbf{1}_{F_{i+r,\kappa+1}} \mathbb{E}\left[\mathbf{1}_{E_{i+r+1,\kappa+1}} \mid \mathcal{G}_{i+r}\right] \mid \mathcal{G}_i\right] \\ &= \sum_{j=1}^{r} \mathbb{E}\left[\mathbf{1}_{S_{i+j} \le s+r} \mathbf{1}_{B_{i+j}} \mid \mathcal{G}_i\right] + \mathbb{E}\left[\mathbf{1}_{F_{i+r,\kappa+1}} \mathbf{1}_{E_{i+r+1,\kappa+1}} \mid \mathcal{G}_i\right] \\ &\le \sum_{j=1}^{r+1} \sum_{l=0}^{\kappa+1} \mathbb{E}\left[\mathbf{1}_{E_{i+j,l}} \mid \mathcal{G}_i\right], \quad \text{on } A_i \cap F_{i,s},\end{aligned}$$

for all $0 \le s \le \kappa$. Multiplying through by $\delta^{\kappa+1}$, bounding $\delta^{\kappa+1}$ using (2.23) with i replaced by $i+j$ and s replaced by l, and then using the Tower Rule, we have

$$\begin{aligned}\delta^{r+\kappa+2} &\le \sum_{j=1}^{r+1} \sum_{l=0}^{\kappa+1} \mathbb{E}\left[\mathbf{1}_{E_{i+j,l}} \delta^{\kappa+1} \mid \mathcal{G}_i\right]. \\ &\le \sum_{j=1}^{r+1} \sum_{l=0}^{\kappa+1} \mathbb{E}\left[\mathbf{1}_{E_{i+j,l}} \mathbb{E}\left[\mathbf{1}_{R_{i+j+l}=0} \mid \mathcal{G}_{i+j}\right] \mid \mathcal{G}_i\right] \\ &\le \sum_{j=1}^{r+1} \sum_{l=0}^{\kappa+1} \mathbb{E}\left[\mathbf{1}_{R_{i+j+l}=0} \mid \mathcal{G}_i\right], \quad \text{on } A_i \cap F_{i,s},\end{aligned}$$

for all $0 \le s \le \kappa$. Since $j \le r+1 \le \kappa+2$ and $0 \le l \le \kappa+1$, we have

$$\delta^{2\kappa+3} \le \mathbb{E}\left[\mathbf{1}_{\bigcup_{k=1}^{2\kappa+3} R_{i+k}=0} \mid \mathcal{G}_i\right], \quad \text{on } A_i \cap F_{i,s}, \tag{2.26}$$

for all $0 \le s \le \kappa$. By (2.23) and (2.26), we have

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{R_i \le b} \mathbf{1}_{\bigcup_{k=0}^{2\kappa+3} R_{i+k}=0} \mid \mathcal{G}_i\right] &\ge \min\left(\delta^{\kappa+1}, \delta^{2\kappa+3}\right) = p, \\ &\text{on } A_i \cap \{R_i \le b\} = A_i \cap \left[\bigcup_{s=0}^{\kappa} E_{i,s} \cup \bigcup_{s=0}^{\kappa} F_{i,s}\right].\end{aligned}$$

Thus (2.8) holds.

By Lemma 2.7, there exists $\eta_1 = \eta_1(\delta, \kappa) > 0$ such that

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^{m}\left(A_{i-1} \cap \left(\overline{B_i} \cup \{S_i > 0\}\right)\right)\right) &\leq \mathbb{P}\left(\bigcap_{i=1}^{m}\left(A_{i-1} \cap \{R_i > 0\}\right)\right) \\ &\leq 2e^{-\eta_1 m} + \mathbb{P}\left(R_0 > \eta_1 m\right) \\ &\leq 2e^{-\frac{1}{2}\eta_1 m} + \mathbb{P}\left(S_0 > \tfrac{1}{2}\eta_1 m\right),\end{aligned}$$

if m is sufficiently large. Hence the result follows by taking $\eta = \eta(\delta, \kappa) := \frac{1}{2}\eta_1$. □

Chapter 3

Rapid mixing — part one

In this chapter, we will begin our proof of rapid mixing of the lengths process. This will require defining *profile-adjacency* for a pair of lengths vectors, and then the *profile coupling* of two lengths processes with profile-adjacent initial states. We will close this chapter with an intermediate result which will be used in the proof of rapid mixing of the lengths process. This chapter is based on Chapter 2 of [10] by Luczak and McDiarmid.

3.1 Profile-adjacency and distance

In this section, we will define *profile-adjacency* and then the *profile-distance* between a pair of lengths vectors. We begin with the concept of *profile-equivalence*.

Definition 3.1. We will say that $x, y \in \mathcal{Q}_n$ are *profile-equivalent*, and write $x \equiv y$, if x and y have

1. the same number of queues of length i, for all $i \geq 0$, and
2. memory queues of the same length.

Informally, we will say that two lengths vectors are *profile-adjacent* if we take a pair of profile-equivalent lengths vectors, and then either add/remove a single customer to/from one of them.

Definition 3.2. We will say that $x, y \in \mathcal{Q}_n$ are *profile-adjacent*, and write $x \sim y$, if there exists $l \geq 0$, called the *level*, such that

1. x and y have the same number of queues of length i for all $i \neq l, l+1$, and one lengths vector (the *lower lengths vector*) has one more queue of length l and one fewer queue of length $l+1$ than the other (the *higher lengths vector*), and
2. either x and y have memory queues of the same length, or the lower (resp., higher) lengths vector has memory queue of length l (resp., $l+1$).

If x is the lower lengths vector, so that y is the upper lengths vector, then we will write $x \prec y$.

Remark. As mentioned in Section 1.3, to show rapid mixing of the standard lengths process, Luczak and McDiarmid [10] show that two standard lengths processes with certain pairs of

initial states can be coupled to coalesce rapidly. A coupling only needs to be constructed for pairs of initial states which constitute the edge set of a certain graph structure on the state space $\mathbb{Z}_+^n$. The graph structure used is the following natural one: say $z, z' \in \mathbb{Z}_+^n$ are adjacent if they differ in exactly one queue by one customer.

A natural way to adapt the notion of adjacency between $z, z' \in \mathbb{Z}_+^n$ to a notion of adjacency between $x, y \in \mathcal{Q}_n$, is to also require that the memory queues in x and in y coincide. Using this graph structure, and then following the arguments in [10], we are able to show that two lengths processes can be coupled to rapidly have the same number of queues of each length and to have memory queues of the same length, that is, to rapidly become profile-equivalent. Although this does not show that the two lengths processes coalesce rapidly, it does however allow us to prove several results about the equilibrium number of queues of length i, for all $i \geq 0$, and the equilibrium memory queue length (in Chapter 4 and Chapter 5). These results will play a role in our proof of rapid mixing of the lengths process.

Hence, it suffices for our notion of adjacency to only concern queue lengths. That is, the first condition does not require that the differing customer to come from the same queue, and the second condition does not require that the memory queues coincide.

A final remark is that it is possible to carry out similar analysis using an abstract process which only contains queue length information: the lengths of the n queues and the length of the memory queue. This approach was taken by Luczak and Norris in [13].

For $x, y \in \mathcal{Q}_n$, define a *profile-path* of *length* m between x and y to be a sequence

$$x = z_0 \sim z_1 \sim \ldots \sim z_m = y.$$

The following lemma says that profile-adjacency induces a connected structure on the state space $\mathcal{Q}_n$.

Lemma 3.3. *Let $x, y \in \mathcal{Q}_n$. Then there exists a profile-path $x = z_0 \sim z_1 \sim \ldots \sim z_m = y$ of length at most $\|x\|_1 + \|y\|_1$ such that*

$$\|z_i\|_1 \leq \max\left(\|x\|_1, \|y\|_1\right), \quad \|z_i\|_\infty \leq \max\left(\|x\|_\infty, \|y\|_\infty\right),$$

for all $0 \leq i \leq m$.

Proof. For $1 \leq i \leq n$, let $\mathbf{0}_i := ((0, \ldots, 0), i)$ denote the empty state with memory queue i. By successively removing customers from x, we obtain a profile-path

$$x = z_0 \sim z_1 \sim \ldots \sim z_k = \mathbf{0}_\xi$$

of length $k := \|x\|_1$ from x to $\mathbf{0}_\xi$. The required inequalities clearly hold for all $0 \leq i \leq k$. Similarly, we obtain a path of length $\|y\|_1$ from y to $\mathbf{0}_\theta$. Now note that the empty states are all profile-equivalent. □

For $x, y \in \mathcal{Q}_n$, let the *profile-distance* $d_p(x, y)$ denote the length of the shortest profile-path between x and y. Then Lemma 3.3 gives

$$d_p(x, y) \leq \|x\|_1 + \|y\|_1 \tag{3.1}$$

for all $x, y \in \mathcal{Q}_n$. Note that $d_p(x, y) = 0$ if and only if $x \equiv y$, and that $d_p(x, y) = 1$ if and only $x \sim y$.

3.2 The profile coupling

In this section, we will define the *profile coupling* of two lengths processes with profile-adjacent initial states. We will then show that under this coupling, at each event time, the two processes either remain profile-adjacent or become profile-equivalent. We will only be concerned with the processes until they become profile-equivalent.

Recall that we rank the queues in a set of queues by length (in ascending order), and then if necessary, by queue index (also in ascending order).

Definition 3.4. The *profile coupling* is the following coupling of lengths processes $\mathbf{X}$ and $\mathbf{Y}$ with profile-adjacent initial states. Let $\mathbf{X}$ and $\mathbf{Y}$ share the same arrival and potential departure times. For an event time T, pair the queues in X_{T-} and Y_{T-} as follows: pair the memory queues together, rank the remaining queues (from 1 to $n-1$) and then pair these queues by rank.

1. If T is an arrival time, let the $\mathbf{X}$-choices $C = (C(1), \ldots, C(d))$ be an ordered list of d queues chosen uniformly at random with replacement, then define the $\mathbf{Y}$-choices $C' = (C'(1), \ldots, C'(d))$ by setting $C'(i)$ to be the queue paired with $C(i)$, for all $1 \leq i \leq d$.

2. If T is a potential departure time, let the $\mathbf{X}$-selection be a queue in X_{T-} selected uniformly at random, then set the $\mathbf{Y}$-selection to be the queue in Y_{T-} paired with the $\mathbf{X}$-selection.

Remark. It is easy to see that for an arrival time, the $\mathbf{Y}$-choices is an ordered list of d queues chosen uniformly at random with replacement, and that for a potential departure time, the $\mathbf{Y}$-selection is a queue in Y_{T-} selected uniformly at random. Thus, $\mathbf{Y}$ does have the distribution of a lengths process. This coupling is based on the coupling introduced by Luczak and McDiarmid [10] used to couple two standard lengths processes together.

Let us look at the first event time $T > 0$ and the initial states of $\mathbf{X}$ and $\mathbf{Y}$ in more detail. Suppose that $X_{T-} = x \prec y = Y_{T-}$ at level l. We claim that the pairing procedure described in Definition 3.4 will pair queues of equal length together, with the exception of one pair consisting of a queue of length l in x and a queue of length $l+1$ in y; we will call these the x- and y-*imbalances*, respectively. To see the claim, there are two cases to consider.

1. If x and y have memory queues of the same length, then pairing the memory queues leaves x and y with the same number of queues of length i for all $i \neq l, l+1$, and x with one more queue of length l and one fewer queue of length $l+1$ than y. These queues are to be ranked from 1 to $n-1$, and as we pair the queues of rank $1, 2, \ldots$ together, we will be pairing queues of equal length $0, 1, \ldots, l-1$ together, if there are any. Eventually, the x- and y-imbalances are created and the remaining pairs are queues of equal length.

2. If x and y have memory queues of length l and $l+1$, respectively, then pairing the memory queues immediately creates the pair of imbalances. Clearly, the remaining $n-1$ pairs are queues of equal length.

Now we will show that under this coupling, at each event time, the two processes either remain profile-adjacent or become profile-equivalent. Moreover, we will determine the precise conditions for each outcome, and give these conditions in terms of one process only.

Lemma 3.5. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *have initial states* $x, y \in \mathcal{Q}_n$*, respectively, where* $x \prec y$ *at level* l*, and let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be coupled by the profile coupling. Let* $T > 0$ *denote the first event time, and let* u *denote the* x*-imbalance.*

1. *If* T *is an arrival time, let* A *denote the event that the* $\mathbf{X}$*-candidates list contains a unique shortest queue and let* J *denote the index of the queue joined by the* $\mathbf{X}$*-customer. Then* $X_T \prec Y_T$ *at level* $L := l + \mathbf{1}_A \mathbf{1}_{J=u}$*.*

2. *If* T *is a potential departure time, let* S *denote the* $\mathbf{X}$*-selection. If* $S = u$ *and* $l = 0$*, then* $X_T \equiv Y_T$*. Otherwise, we have* $X_T \prec Y_T$ *at level* $L := l - \mathbf{1}_{S=u}$*.*

Proof. Throughout this proof, recall that the x-imbalance is paired with the y-imbalance, which has one more customer, and that every other queue in x is paired to a queue in y of equal length. There are now two cases to consider.

Case 1 T is an arrival time.

Let K and $K+\Delta$ denote the lengths of the two shortest queues in the $\mathbf{X}$-candidates list, respectively, where $K, \Delta \geq 0$; the argument is trivial if the $\mathbf{X}$-candidates lists only consists of one unique queue. There are now three cases to consider.

1. If A holds, then $\Delta \geq 1$. It follows that the $\mathbf{X}$-customer joins a queue of length K, and that the memory queue in X_T has length $K+1$.

   (a) If $J = u$, then the $\mathbf{Y}$-candidates list contains queues of length $K+1$ and $K + \Delta \geq K+1$. It follows that the $\mathbf{Y}$-customer joins a queue of length $K+1$, and that the memory queue in Y_T has length $K+1$ or $K+2$.

   (b) If $J \neq u$, then the $\mathbf{Y}$-candidates list contains a queue of length K and a queue of length at least $K + \Delta \geq K+1$. It follows that the $\mathbf{Y}$-customer joins a queue of length K, and that the memory queue in Y_T has length $K+1$.

2. If $\overline{A}$ holds, then $\Delta = 0$. It follows that the $\mathbf{X}$-customer joins a queue of length K, and that the memory queue in X_T has length K. At least one of these two shortest queues in the $\mathbf{X}$-candidates list is not the x-imbalance u, so the $\mathbf{Y}$-candidates list contains a queue of length K and a queue of length at least K. It follows that the $\mathbf{Y}$-customer joins a queue of length K, and that the memory queue in Y_T has length K or $K+1$.

In all cases, we have $X_T \prec Y_T$. Moreover, the level increases in the first case, and stays constant in the other two cases.

Case 2 T is a potential departure time.

Recall that the $\mathbf{X}$-selection is the x-imbalance if and only if the $\mathbf{Y}$-selection is the y-imbalance. Thus, $X_T \prec Y_T$ with the level decreasing if and only if this occurs. The only exception is if $l = 0$, in which case $\mathbf{X}$ and $\mathbf{Y}$ become profile-equivalent. □

We now extend the profile coupling of lengths processes with profile-adjacent initial states to lengths processes with arbitrary initial states.

Definition 3.6. Let $\mathbf{X}$ and $\mathbf{Y}$ have arbitrary initial states $x, y \in \mathcal{Q}_n$, respectively. Let $x = z_0 \sim z_1 \sim \ldots \sim z_m = y$ be a shortest profile-path of length $m = d_p(x, y)$ between x and y. For all $0 \leq i \leq m$, let $\mathbf{Z}^i$ be a lengths process with initial state z_i, and let $\mathbf{Z}^{j-1}$ and $\mathbf{Z}^j$ be coupled by the profile coupling, for all $1 \leq j \leq m$. This determines a coupling of $\mathbf{X}$ and $\mathbf{Y}$, which we will also call a *profile coupling*.

We then have the following result.

Lemma 3.7. *Let $\mathbf{X}$ and $\mathbf{Y}$ have arbitrary initial states and be coupled by a profile coupling. Then $d_p(X_t, Y_t)$ is non-increasing over time.*

Proof. Let m and the $\mathbf{Z}^i$ be as in Definition 3.6. Then

$$d_p(X_t, Y_t) \leq \sum_{i=1}^{m} d_p\left(Z_t^{i-1}, Z_t^i\right).$$

Each summand takes the value 1 before the first event time, and by Lemma 3.5, a value in $\{0, 1\}$ at the first event time. Hence $d_p(X_t, Y_t)$ is non-increasing across the first event time, and by induction, is non-increasing over all time. □

3.3 Rapid profile-equivalence

In this section, we will show that in a profile coupling, under reasonable initial conditions, the two lengths processes in fact rapidly become profile-equivalent.

We will begin by outlining our strategy for this section. Our strategy is to examine the level between the two lengths processes in a profile coupling, so we make the following definition.

Definition 3.8. Let $\mathbf{X}$ and $\mathbf{Y}$ have initial states $x, y \in \mathcal{Q}_n$, respectively, where $x \prec y$, and let $\mathbf{X}$ and $\mathbf{Y}$ be coupled by the profile coupling. Let

$$T_{\text{co}} := \inf\{t \geq 0 : X_t \equiv Y_t\}.$$

For $0 \leq t < T_{\text{co}}$, say $X_t \prec Y_t$ at level L_t. The *level walk* is the random walk $\mathbf{W} = (W_t)_{t \geq 0}$ on $\mathbb{Z}_+$ defined by setting

$$W_t := \begin{cases} L_t + 1, & \text{if } 0 \leq t < T_{\text{co}}, \\ 0, & \text{if } t \geq T_{\text{co}}. \end{cases}$$

We will show that T_{co} is small by showing that with high probability $\mathbf{W}$ soon decreases to 0. To do this, we will analyse $\mathbf{W}$ at some times $(J_i)_{i=0}^{\infty}$ to be defined later (these are

not the jump times as defined in Section 1.4), that is, we will analyse the random walk $\mathbf{W_J} = (W_{J_i})_{i=0}^{\infty}$.

We will apply Lemma 2.8 to $\mathbf{S} = \mathbf{W_J}$ as follows. The background events A_i will denote the event that $\mathbf{X}$ does not have too many customers for long periods of time; these events will hold with high probability by Lemma 2.5 (2). It follows that any long queue should be very unlikely to receive additional customers, and thus, its length should drift downwards towards b. At that point, we can wait for a sequence of b consecutive departures from the $\mathbf{X}$-imbalance (whose index may change after each departure) so that $\mathbf{S}$ decreases to 0.

However, whenever a queue is saved as the memory queue, it experiences an upward pressure on its length, since the next arriving customer will consider joining it. Thus, if the $\mathbf{X}$-imbalance is the memory queue, then $\mathbf{W}$ may not even drift downwards. Thus we will keep track of when the $\mathbf{X}$-imbalance is the memory queue: we will say that $\mathbf{S}$ is *good* at step i if the $\mathbf{X}$-imbalance at time J_i is not the memory queue, and that $\mathbf{S}$ is *bad* otherwise.

Now let us say a little about Lemma 2.8. For the first condition, (2.18), we must show that $\mathbf{S}$ will either become good or increase, with probability bounded away from 0, when it is bad. This requirement leads us to the following definition.

Definition 3.9. Let $\mathbf{X}$ and $\mathbf{Y}$ have profile-adjacent initial states and be coupled by the profile coupling. Let $T > 0$ be an arrival time where $X_{T-} \prec Y_{T-}$ at level l. We will say that a queue is *taboo* if it is a queue of length l, but is not the X_{T-}-imbalance. We will say that a queue is *non-taboo* if it is not taboo. We will say that T is *helpful* if the $\mathbf{X}$-customer selects exactly zero or at least two taboo queues, and *unhelpful* otherwise.

Now we will show that if $\mathbf{S}$ is bad, then given a helpful arrival time, it will either become good or increase.

Lemma 3.10. *Let $\mathbf{X}$ and $\mathbf{Y}$ have profile-adjacent initial states and be coupled by the profile coupling. Let $T > 0$ be an arrival time where $X_{T-} \prec Y_{T-}$ at level l and where the X_{T-}-imbalance is the memory queue. Then*

$$\{T \text{ is helpful}\} \subseteq \{X_T\text{-imbalance is not the memory queue}\} \cup \{X_T \prec Y_T \text{ at level } l+1\}.$$

Moreover, we have equality if $d = 1$.

Proof. First note that since T is an arrival time, we do indeed have $X_T \prec Y_T$, by Lemma 3.5 (1). Let K denote the length of the queue joined by the $\mathbf{X}$-customer. Since the memory queue is the X_{T-}-imbalance, and its length is the level l, we have $K \leq l$. There are three cases to consider.

1. $K \leq l-1$. In this case, no queue of length K is the X_{T-}-imbalance, so the $\mathbf{X}$- and $\mathbf{Y}$-candidates lists both contain an equal number of shortest queues. It follows that the memory queues in X_T and Y_T have the same length, and thus are not the X_T- and Y_T-imbalances, respectively.

2. $K = l$ with the $\mathbf{X}$-customer selecting at least two taboo queues. In this case, the $\mathbf{X}$- and $\mathbf{Y}$-candidates lists both contain at least two shortest queues. Again, it follows that the memory queues in X_T and Y_T have the same length, and thus are not the X_T- and Y_T-imbalances, respectively.

3. $K = l$ with the **X**-customer selecting no taboo queues. In this case, the **X**-candidates list contains a unique shortest queue. By Lemma 3.5 (1), we have $X_T \prec Y_T$ at level $l+1$.

It remains to show that if $d = 1$, then we have equality. Thus, we suppose that T is unhelpful, so that the **X**-customer selects a taboo queue. In this case, the **X**-candidates list contains two queues of length l, whilst the **Y**-candidates list contains queues of length l and $l+1$. It follows that the memory queues in X_T and Y_T have different lengths (l and $l+1$, respectively), and thus are the X_T- and Y_T-imbalances, respectively. By Lemma 3.5 (1), we also have $X_T \prec Y_T$ at level l. □

The following lemma says that if $d \geq 2$, n is sufficiently large and **S** is bad, then the next event time has probability bounded away from 0 of being a helpful arrival.

Lemma 3.11. *Let $d \geq 2$. Then there exists $n^* \geq 1$ such that the following holds. Let $n \geq n^*$, let* **X** *and* **Y** *have profile-adjacent initial states, and let* **X** *and* **Y** *be coupled by the profile coupling. Let $T > 0$ be an event time where $X_{T-} \prec Y_{T-}$ and where the X_{T-}-imbalance is the memory queue. Then*

$$\mathbb{P}\left(T \text{ is a helpful arrival} \mid \mathcal{F}_{T-}\right) \geq \frac{\lambda}{\lambda+1}\frac{1}{4}.$$

Proof. If T is an unhelpful arrival, then the **X**-customer selects exactly one taboo queue. Hence there exists a choice $1 \leq R \leq d$ such that choice R is a taboo queue, choices $1, \ldots, R-1$ are non-taboo queues, and choices $R+1, \ldots, d$ are either non-taboo queues or the same as choice R. Let $M \geq 1$ denote the number of non-taboo queues, then

$$\begin{aligned}\mathbb{P}\left(T \text{ is an unhelpful arrival} \mid \mathcal{F}_{T-}\right) &= \frac{\lambda}{\lambda+1}\sum_{r=1}^{d}\left(\frac{M}{n}\right)^{r-1}\frac{n-M}{n}\left(\frac{M+1}{n}\right)^{d-r} \\ &= \frac{\lambda}{\lambda+1}\frac{n-M}{n^d}\left[(M+1)^d - M^d\right],\end{aligned}$$

since $\sum_{r=1}^{d} x^{r-1}y^{d-r} = \frac{y^d - x^d}{y-x}$ for all distinct $x, y \in \mathbb{R}$. Expanding the binomial term and using Lemma 1.5 (1) (with $x = M$ and $y = n$) gives

$$\begin{aligned}\mathbb{P}\left(T \text{ is an unhelpful arrival} \mid \mathcal{F}_{T-}\right) &\leq \frac{\lambda}{\lambda+1}\frac{n-M}{n^d}\left(dM^{d-1} + 2^d M^{d-2}\right) \\ &\leq \frac{\lambda}{\lambda+1}\left[\frac{d\left(n-M\right)M^{d-1}}{n^d} + \frac{2^d}{n}\right] \\ &\leq \frac{\lambda}{\lambda+1}\left[\frac{1}{2} + \frac{2^d}{n}\right] \leq \frac{\lambda}{\lambda+1}\frac{3}{4},\end{aligned}$$

if n^* is sufficiently large. Hence the result follows. □

Now we will show that in a profile coupling, under reasonable initial conditions, the two lengths processes rapidly become profile-equivalent.

Lemma 3.12. *Let $c > \frac{\lambda}{1-\lambda}$. Then there exists $0 < \beta = \beta(c) < 1$ such that the following holds. Let $n \geq 1$, let* **X** *and* **Y** *have initial states $x, y \in \mathcal{Q}_n$, respectively, where $x \prec y$ and*

$\|x\|_1 \le cn$, *and let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be coupled by the profile coupling. Then*

$$\mathbb{E}\left[d_p\left(X_t, Y_t\right)\right] = \mathbb{E}\left[\mathbf{1}_{X_t \not\equiv Y_t}\right] \le e^{-\beta t} + 2e^{-\beta n}$$

for all $t \ge \frac{1}{\beta}\|x\|_\infty$.

Proof. Since the left-hand side is bounded by 1 and $\beta > 0$ may be arbitrarily small, it suffices to show the result for all sufficiently large n.

Let $\mathbf{W} = (W_t)_{t \ge 0}$ denote the level walk, and note that

$$T_{\mathrm{co}} = \inf\left\{t > 0 : W_t = 0\right\}.$$

For $0 \le t < T_{\mathrm{co}}$, say $\mathbf{X}$ is *good* at time t if the X_t-imbalance is not the memory queue, and *bad* otherwise. If U_t denotes the X_t-imbalance, then let

$$D_t := \begin{cases} \{U_t \ne \Xi_t\}, & \text{if } 0 \le t < T_{\mathrm{co}}, \\ \Omega, & \text{if } t \ge T_{\mathrm{co}}. \end{cases} \tag{3.2}$$

Thus, for $0 \le t < T_{\mathrm{co}}$, D_t denotes the event that $\mathbf{X}$ is good at time t. Define the *change times* $J_0 := 0$ and

$$J_i := \inf\left\{t > J_{i-1} : \mathbf{1}_{D_t} \ne \mathbf{1}_{D_{t-}} \text{ or } W_t \ne W_{t-}\right\},$$

for all $i \ge 1$. That is, let J_i be the first time after J_{i-1} when either $\mathbf{X}$ starts/stops being good or when $\mathbf{W}$ changes value. The filtration $(\mathcal{G}_i)_{i=0}^{\infty}$ we will be using for Lemma 2.8 will be based on these change times: for $i \ge 0$, set $\mathcal{G}_i := \mathcal{F}_{J_{i+1}-}$ to be the σ-field generated by all events before J_{i+1}.

Now, for $t \ge 0$, let

$$C_t := \{\|X_r\|_1 \le 2cn \text{ for all } 0 \le r < t\}, \quad m := \left\lceil \tfrac{1}{4}t \right\rceil.$$

Then

$$\begin{aligned} \mathbb{P}\left(X_t \not\equiv Y_t\right) \le{}& \mathbb{P}\left(\{X_t \not\equiv Y_t\} \cap \{J_m \le t\} \cap C_t\right) \\ &+ \mathbb{P}\left(\{X_t \not\equiv Y_t\} \cap \{J_m > t\}\right) + \mathbb{P}\left(\overline{C_t}\right), \end{aligned} \tag{3.3}$$

for all $t \ge 0$. The first term will be where we apply Lemma 2.8, but let us bound the last two terms first.

We claim that on $\{X_t \not\equiv Y_t\}$, change times occur at rate at least 1 over $[0, t]$. To see this claim, consider a time $0 \le r < t$. As we have not yet become profile-equivalent, $\mathbf{W}$ is non-zero, and a sufficient condition for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is the $\mathbf{X}$-imbalance. Such events occur at rate $n \cdot \frac{1}{n} = 1$. Hence the number of change times $N_t := \max\left\{i \ge 0 : J_i \le t\right\}$ in $[0, t]$ stochastically dominates a $\mathrm{Po}\left(t\right)$ random variable on the event $\{X_t \not\equiv Y_t\}$. By Lemma 1.4 (with $\varepsilon = \frac{1}{2}$), we have

$$\mathbb{P}\left(\{X_t \not\equiv Y_t\} \cap \{J_m > t\}\right) \le \mathbb{P}\left(N_t < m\right) \le \mathbb{P}\left(\mathrm{Po}\left(t\right) \le \tfrac{1}{2}t\right) \le 2e^{-\frac{1}{12}t}, \tag{3.4}$$

for all $t \geq 2$. To see the second inequality, note that $m = \left\lceil \frac{1}{4}t \right\rceil \leq \frac{1}{2}t$.

By Lemma 2.5, there exists $\eta_1 = \eta_1(c) > 0$ such that

$$\mathbb{P}\left(\overline{C_t}\right) \leq 2e^{-\eta_1 n}, \tag{3.5}$$

for all $0 \leq t \leq e^{\eta_1 n}$.

Hence, by (3.3)-(3.5), we have

$$\mathbb{P}\left(X_t \not\equiv Y_t\right) \leq \mathbb{P}\left(\{X_t \not\equiv Y_t\} \cap \{J_m \leq t\} \cap C_t\right) + 2e^{-\frac{1}{12}t} + 2e^{-\eta_1 n}, \tag{3.6}$$

for all $2 \leq t \leq e^{\eta_1 n}$. Having bounded the last two terms in (3.3) to obtain (3.6), we now turn our attention to the first term, which is where we will apply Lemma 2.8. We have already defined the filtration $(\mathcal{G}_i)_{i=0}^{\infty}$ for this lemma by setting each $\mathcal{G}_i := \mathcal{F}_{J_{i+1}-}$. The background events are

$$A_i := C_{J_{i+1}},$$

for $i \geq 0$. For $i \geq 0$, let

$$B_i := \{\mathbf{1}_{D_r} = 1 \text{ for all } J_i \leq r < J_{i+1}\}$$

denote the event that $\mathbf{X}$ is good at all times $J_i \leq r < J_{i+1}$. Note that A_i and B_i are both $\mathcal{G}_i$-measurable, since they depend only on the history of the process until but excluding J_{i+1}. The random walk is $\mathbf{S} = \mathbf{W_J}$, that is,

$$S_i := W_{J_i},$$

where $i \geq 0$. Note that each increment $Z_i := S_i - S_{i-1} = W_{J_i} - W_{J_{i-1}}$ is $\mathcal{G}_i$-measurable and $\{-1, 0, 1\}$-valued. Let the initial value be $S_0 = s \geq 1$. We will say that $\mathbf{S} = \mathbf{W_J}$ is *good* at step i if $\mathbf{X}$ is good at time J_i, and that $\mathbf{S}$ is *bad* otherwise. Thus, $\mathbf{S}$ is good at i if and only if D_{J_i} holds, and because $\mathbf{1_D}$ is constant between change times, it follows that $\mathbf{S}$ is good at step i if and only if B_i holds.

Having defined the sequences of events and the random walk, we may now write (3.6) as

$$\begin{aligned}
\mathbb{P}\left(X_t \not\equiv Y_t\right) &\leq \mathbb{P}\left(\{X_t \not\equiv Y_t\} \cap \{J_m \leq t\} \cap C_t\right) + 2e^{-\frac{1}{12}t} + 2e^{-\eta_1 n} \\
&\leq \mathbb{P}\left(\bigcap_{i=1}^{m} \left(A_{i-1} \cap \{S_i > 0\}\right)\right) + 2e^{-\frac{1}{12}t} + 2e^{-\eta_1 n} \\
&\leq \mathbb{P}\left(\bigcap_{i=1}^{m} \left(A_{i-1} \cap \left(\overline{B_i} \cup \{S_i > 0\}\right)\right)\right) + 2e^{-\frac{1}{12}t} + 2e^{-\eta_1 n},
\end{aligned} \tag{3.7}$$

for all $2 \leq t \leq e^{\eta_1 n}$.

Next we define some constants. Let

$$\delta := \min\left(\frac{\lambda}{\lambda+1}\frac{1}{4}, \frac{\lambda}{\lambda d+1}, \frac{1-\lambda}{\lambda d+1}\right).$$

Define $0 < \varepsilon, \omega < 1$ as follows. If $d = 1$, then let $\varepsilon := 1$, else let ε be sufficiently small so

that

$$d\varepsilon^{d-1} \leq \frac{1}{\lambda d+1}.$$

Let ω be sufficiently small so that

$$\frac{1-\omega^d}{1+\omega^d} \geq 1-\tfrac{1}{2}\delta.$$

Finally, let

$$\kappa = \kappa(c) := \left\lceil \frac{2c}{\min(\varepsilon,\omega)} \right\rceil + 1.$$

Now we will show that the hypotheses of Lemma 2.8 hold with the filtration $(\mathcal{G}_i)_{i=0}^{\infty}$, the sequences of events $(A_i)_{i=0}^{\infty}$ and $(B_i)_{i=0}^{\infty}$, the random walk $\mathbf{S} = (S_i)_{i=0}^{\infty}$, and the constants δ and κ, all as defined above. There are now four conditions to verify: (2.18)-(2.21).

1. For condition (2.18), we are looking at

$$\mathbb{P}\left(B_{i+1} \cup \{Z_{i+1}=1\} \mid \mathcal{G}_i\right), \quad \text{on } A_i \cap \overline{B_i}.$$

   Now

$$A_i \cap \overline{B_i} \subseteq F_{i,1} := \left\{U_{J_{i+1}-} = \Xi_{J_{i+1}-}\right\} \subseteq \left\{W_{J_{i+1}-} > 0\right\},$$

   where the last inclusion holds by the definition in (3.2). We will work on the event $F_{i,1}$, which says that immediately before J_{i+1}, the level walk is non-zero and the $\mathbf{X}$-imbalance is the memory queue. Since $B_{i+1} \cup \{Z_{i+1}=1\}$ denotes the event that the $(i+1)^{\text{st}}$ change time is one where $\mathbf{S}$ becomes good or increases, we may write

$$\mathbb{P}\left(B_{i+1} \cup \{Z_{i+1}=1\} \mid \mathcal{G}_i\right) \geq \frac{p_1}{q_1}, \quad \text{on } A_i \cap \overline{B_i}, \tag{3.8}$$

   where p_1 is a lower bound on the rate of events where $\mathbf{X}$ becomes good or $\mathbf{W}$ increases, and q_1 is an upper bound on the rate of events where $\mathbf{X}$ becomes good or $\mathbf{W}$ changes value (i.e., change times). There are now two cases to consider.

   (a) Case 1: $d \geq 2$. We may take the lower bound $p_1 := \frac{1}{4}\lambda n$, if n is sufficiently large. To see this, note that a sufficient condition for $\mathbf{X}$ to become good or for $\mathbf{W}$ to increase is if we have a helpful arrival, by Lemma 3.10. Since $d \geq 2$, helpful arrivals occur at rate at least $(\lambda+1)\,n \cdot \frac{\lambda}{\lambda+1}\frac{1}{4} = \frac{1}{4}\lambda n$, if n is sufficiently large; this holds by Lemma 3.11.

   We may take the upper bound $q_1 = (\lambda+1)\,n$, the rate of all events.

   Then (3.8) gives

$$\mathbb{P}\left(B_{i+1} \cup \{Z_{i+1}=1\} \mid \mathcal{G}_i\right) \geq \frac{\frac{1}{4}\lambda n}{(\lambda+1)\,n} \geq \delta, \quad \text{on } A_i \cap \overline{B_i},$$

   and (2.18) holds, if n is sufficiently large.

   (b) Case 2: $d = 1$. Recall that, in this case, an arrival is helpful if and only if the $\mathbf{X}$-customer selects a non-taboo queue; let $M_i \geq 1$ denote the number of non-taboo queues immediately before J_{i+1}.

We may take the lower bound $p_1 := \lambda M_i$. To see this, note that a sufficient condition for $\mathbf{X}$ to become good is if we have a helpful arrival, by Lemma 3.10. Such events occur at rate $\lambda n \cdot \frac{M_i}{n} = \lambda M_i$.

We may take the upper bound $q_1 := \lambda M_i + 1$. To see this, note that a necessary for $\mathbf{X}$ to become good or for $\mathbf{W}$ to increase is if we have a helpful arrival, by the equality in Lemma 3.10. Such events occur at rate $\lambda n \cdot \frac{M_i}{n} = \lambda M_i$. A necessary condition for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is the $\mathbf{X}$-imbalance. Such events occur at rate $n \cdot \frac{1}{n} = 1$.

Then (3.8) gives

$$\mathbb{P}\left(B_{i+1} \cup \{Z_{i+1} = 1\} \mid \mathcal{G}_i\right) \geq \frac{\lambda M_i}{\lambda M_i + 1} \geq \frac{\lambda \cdot 1}{\lambda \cdot 1 + 1} \geq \delta, \quad \text{on } A_i \cap \overline{B_i},$$

and (2.18) holds.

2. For condition (2.19), we are looking at

$$\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = -1\} \mid \mathcal{G}_i\right), \quad \text{on } A_i \cap B_i \cap \{S_i > 0\}.$$

Now

$$A_i \cap B_i \cap \{S_i > 0\} \subseteq F_{i,2} := \left\{W_{J_{i+1}-} > 0\right\} \cap \left\{U_{J_{i+1}-} \neq \Xi_{J_{i+1}-}\right\}.$$

We will work on the event $F_{i,2}$, which says that immediately before J_{i+1}, the level walk is non-zero and the $\mathbf{X}$-imbalance is not the memory queue. Since $B_{i+1} \cap \{Z_{i+1} = -1\}$ denotes the event that the $(i+1)^{\text{st}}$ change time is one where $\mathbf{S}$ remains good and decreases, we may write

$$\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = -1\} \mid \mathcal{G}_i\right) \geq \frac{p_2}{q_2}, \quad \text{on } A_i \cap B_i \cap \{S_i > 0\}, \tag{3.9}$$

where p_2 is a lower bound on the rate of events where $\mathbf{X}$ remains good and $\mathbf{W}$ decreases, and q_2 is an upper bound on the rate of events where $\mathbf{X}$ becomes bad or $\mathbf{W}$ changes value (i.e., change times).

We may take the lower bound $p_2 := 1$. To see this, note that a sufficient condition for $\mathbf{X}$ to remain good and for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is the $\mathbf{X}$-imbalance. Such events occur at rate $n \cdot \frac{1}{n} = 1$.

We may take the upper bound $q_2 := \lambda d + 1$. To see this, note that a necessary condition for $\mathbf{X}$ to become bad or for $\mathbf{W}$ to increase is if we have an arrival where the $\mathbf{X}$-customer selects the $\mathbf{X}$-imbalance at least once (since it is not the memory queue). Such events occur at rate at most $\lambda n \cdot \frac{d}{n} = \lambda d$. A necessary condition for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is the $\mathbf{X}$-imbalance. Such events occur at rate $n \cdot \frac{1}{n} = 1$.

Then (3.9) gives

$$\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = -1\} \mid \mathcal{G}_i\right) \geq \frac{1}{\lambda d + 1} \geq \delta, \quad \text{on } A_i \cap B_i \cap \{S_i > 0\}, \tag{3.10}$$

and (2.19) holds.

3. For condition (2.20), we are looking at

$$\mathbb{P}\left(Z_{i+1}=1 \mid \mathcal{G}_i\right), \quad \text{on } A_i \cap B_i \cap \{S_i > \kappa\}.$$

Now

$$A_i \cap B_i \cap \{S_i > \kappa\} \subseteq F_{i,3} := \left\{W_{J_{i+1}-} > \kappa\right\} \cap \left\{U_{J_{i+1}-} \neq \Xi_{J_{i+1}-}\right\} \cap C_{J_{i+1}}.$$

We will work on the event $F_{i,3}$, which says that immediately before J_{i+1}, the level walk is greater than κ, the **X**-imbalance is not the memory queue, and the number of customers is at most $2cn$. Hence the **X**-imbalance has length at least κ (see Definition 3.8), whence the proportion of queues at least as long as the **X**-imbalance is at most

$$u_\kappa\left(X_{J_{i+1}-}\right) \leq \frac{\left\|X_{J_{i+1}-}\right\|_1}{n\kappa} \leq \frac{2c}{\kappa} \leq \min\left(\varepsilon, \omega\right) \leq \varepsilon.$$

Since $\{Z_{i+1}=1\}$ denotes the event that the $(i+1)^{\text{st}}$ change time is one where **S** increases, we may write

$$\mathbb{P}\left(Z_{i+1}=1 \mid \mathcal{G}_i\right) \leq \frac{p_3}{q_3}, \quad \text{on } A_i \cap B_i \cap \{S_i > \kappa\}, \tag{3.11}$$

where p_3 is an upper bound on the rate of events where **W** increases, and q_3 is a lower bound on the rate of events where **X** becomes bad or **W** changes value (i.e., change times). Note that if **W** increases at J_{i+1}, then immediately before J_{i+1}, the **X**-imbalance cannot be longer than the memory queue. That is, we have

$$X_{J_{i+1}-}\left(U_{J_{i+1}-}\right) \leq X_{J_{i+1}-}\left(\Xi_{J_{i+1}-}\right). \tag{3.12}$$

There are now two cases to consider.

(a) Case 1: $d \geq 2$. We may take the upper bound $p_3 := \lambda d \varepsilon^{d-1}$. To see this, note that a necessary condition for **W** to increase is if we have an arrival where the **X**-customer selects only queues as long as the **X**-imbalance, and he/she selects the **X**-imbalance at least once (since the **X**-imbalance is not the memory queue). Such events occur at rate at most $\lambda n \cdot \frac{d}{n}\varepsilon^{d-1} = \lambda d \varepsilon^{d-1}$.

We may take the lower bound $q_3 := 1$. To see this, note that a sufficient condition for **W** to decrease is if we have a potential departure where the **X**-selection is the **X**-imbalance. Such events occur at rate $n \cdot \frac{1}{n} = 1$.

Then (3.11) gives

$$\mathbb{P}\left(Z_{i+1}=1 \mid \mathcal{G}_i\right) \leq \frac{\lambda d \varepsilon^{d-1}}{1} \leq \frac{\lambda}{\lambda d + 1}, \quad \text{on } A_i \cap B_i \cap \{S_i > \kappa\}.$$

The last inequality holds since $d \geq 2$. By (3.10), we have

$$\begin{aligned}\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = -1\} \mid \mathcal{G}_i\right) &\geq \frac{1}{\lambda d + 1} \\ &= \frac{\lambda}{\lambda d + 1} + \frac{1-\lambda}{\lambda d + 1} \\ &\geq \mathbb{P}\left(Z_{i+1} = 1 \mid \mathcal{F}_i\right) + \delta, \quad \text{on } A_i \cap B_i \cap \{S_i > \kappa\},\end{aligned}$$

and (2.20) holds.

(b) Case 2: $d = 1$. We may take the upper bound $p_3 := \lambda d$. To see this, note that a necessary condition for $\mathbf{W}$ to increase is if we have an arrival time where the $\mathbf{X}$-customer selects the $\mathbf{X}$-imbalance (since the $\mathbf{X}$-imbalance is not the memory queue). Such events occurs at rate $\lambda n \cdot \frac{1}{n} = \lambda$.
We may take the lower bound $q_3 := \lambda d + 1$. To see this, note that a sufficient condition for $\mathbf{W}$ to increase is if we have an arrival where the $\mathbf{X}$-customer selects the $\mathbf{X}$-imbalance (since the $\mathbf{X}$-imbalance is neither the memory queue nor longer than it, by (3.12)). Such events occur at rate $\lambda n \cdot \frac{1}{n} = \lambda d$. A sufficient condition for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is the $\mathbf{X}$-imbalance. Such events occur at rate $n \cdot \frac{1}{n} = 1$.
Then (3.11) gives

$$\mathbb{P}\left(Z_{i+1} = 1 \mid \mathcal{G}_i\right) \leq \frac{\lambda}{\lambda d + 1}, \quad \text{on } A_i \cap B_i \cap \{S_i > \kappa\},$$

and (2.20) holds by the same calculation as in case 1.

4. For condition (2.21), we are looking at

$$\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = 0\} \mid \mathcal{G}_i\right), \quad \text{on } A_i \cap \overline{B_i} \cap \{S_i > \kappa\}.$$

Now

$$A_i \cap \overline{B_i} \cap \{S_i > \kappa\} \subseteq F_{i,4} := \left\{W_{J_{i+1}-} > \kappa\right\} \cap \left\{U_{J_{i+1}-} = \Xi_{J_{i+1}-}\right\} \cap C_{J_{i+1}}.$$

We will work on the event $F_{i,4}$, which says that immediately before J_{i+1}, the level walk is greater than κ, the $\mathbf{X}$-imbalance is the memory queue, and the number of customers is at most $2cn$. Hence the $\mathbf{X}$-imbalance has length at least κ (see Definition 3.8), and the proportion of queues of length at least $\kappa - 1$ is at most

$$u_{\kappa-1}\left(X_{J_{i+1}-}\right) \leq \frac{\left\|X_{J_{i+1}-}\right\|_1}{n\left(\kappa - 1\right)} \leq \frac{2c}{\kappa - 1} \leq \min\left(\varepsilon, \omega\right) \leq \omega.$$

Since $B_{i+1} \cap \{Z_{i+1} = 0\}$ denotes the event that the $(i+1)^{\text{st}}$ change time is one where $\mathbf{S}$ becomes good and does not change value, we may write

$$\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = 0\} \mid \mathcal{G}_i\right) \geq \frac{p_4}{q_4}, \quad \text{on } A_i \cap \overline{B_i} \cap \{S_i > \kappa\}, \tag{3.13}$$

where p_4 is a lower bound on the rate of events where $\mathbf{X}$ becomes good and $\mathbf{W}$ does not change value, and q_4 is an upper bound on the rate of events where $\mathbf{X}$ becomes

good or $\mathbf{W}$ changes value (i.e., change times).

We may take the lower bound $p_4 := \lambda n\left(1-\omega^d\right)$. To see this, note that a sufficient condition for $\mathbf{X}$ to become good and for $\mathbf{W}$ to not change value is if we have an arrival where the $\mathbf{X}$-customer selects a queue shorter than $\kappa-1$ (since the $\mathbf{X}$-imbalance has length at least κ, the $\mathbf{X}$- and $\mathbf{Y}$-candidates lists will both contain an equal number of shortest queues, and these are shorter than $\kappa-1$). Such events occur at rate at least $\lambda n\left(1-\omega^d\right)$.

We may take the upper bound $q_4 := \lambda n + 1$. To see this, note that a necessary condition for $\mathbf{X}$ to become good or for $\mathbf{W}$ to increase is if we have an arrival. Arrivals occur at rate λn. A necessary condition for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is the $\mathbf{X}$-imbalance. Such events occur at rate $n \cdot \frac{1}{n} = 1$.

Then (3.13) gives

$$\begin{aligned}\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1}=0\} \mid \mathcal{F}_i\right) &\geq \frac{\lambda n\left(1-\omega^d\right)}{\lambda n+1} \\ &\geq \frac{1-\omega^d}{1+\omega^d} \geq 1-\tfrac{1}{2}\delta, \quad \text{on } A_i \cap \overline{B_i} \cap \{S_i > \kappa\},\end{aligned}$$

if n is sufficiently large, and (2.21) holds.

Since we have shown that the hypotheses of Lemma 2.8 hold if n is sufficiently large, there exists a constant $\eta_2 = \eta_2(c) > 0$ such that (3.7) becomes

$$\begin{aligned}\mathbb{P}\left(X_t \not\equiv Y_t\right) &\leq \mathbb{P}\left(\bigcap_{i=1}^{m}\left(A_{i-1} \cap \left(\overline{B_i} \cup \{S_i > 0\}\right)\right)\right) + 2e^{-\frac{1}{12}t} + 2e^{-\eta_1 n} \\ &\leq 2e^{-\eta_2 m} + \mathbf{1}_{s>\eta_2 m} + 2e^{-\frac{1}{12}t} + 2e^{-\eta_1 n},\end{aligned}$$

for all $2 \leq t \leq e^{\eta_1 n}$, and if n is sufficiently large. We will also assume, without loss of generality, that $0 < \eta_2 < 1$.

Let $\eta_3 = \eta_3(c) := \frac{1}{2}\min\left(\eta_1, \eta_2, \frac{1}{12}\right)$, then

$$\mathbb{P}\left(X_t \not\equiv Y_t\right) \leq 4e^{-2\eta_3 t} + 2e^{-2\eta_3 n} \leq e^{-\eta_3 t} + e^{-\eta_3 n},$$

for all $\frac{4}{\eta_3}s \leq t \leq e^{\eta_3 n}$, and if n is sufficiently large. To see the first inequality, note that $t \geq \frac{4}{\eta_3}s \geq 2$ (since $s \geq 1$) and that $\eta_2 m \geq \eta_3 \frac{t}{4} \geq s$. To see the second inequality, note that $4 \leq e^{\eta_3 t}$ (since $t \geq \frac{4}{\eta_3}s \geq \frac{\ln 4}{\eta_3}$). We can remove the upper bound on t as follows. If $t > e^{\eta_3 n}$, then

$$\mathbb{P}\left(X_t \not\equiv Y_t\right) \leq \mathbb{P}\left(X_n \not\equiv Y_n\right) \leq 2e^{-\eta_3 n},$$

if n is sufficiently large so that $e^{\eta_3 n} > n$. Let $\beta = \beta(c) := \frac{1}{4}\eta_3$, then

$$\mathbb{P}\left(X_t \not\equiv Y_t\right) \leq e^{-\beta t} + 2e^{-\beta n}$$

for all $t \geq \frac{1}{\beta}\|x\|_\infty$, and if n is sufficiently large. To see this, note that $\frac{1}{\beta}\|x\|_\infty \geq \frac{4}{\eta_3}s$. $\square$

We then have the following result.

Lemma 3.13. *Let $c > \frac{\lambda}{1-\lambda}$, then let $0 < \beta = \beta(c) < 1$ denote the constant given by Lemma 3.12. Let $n \geq 1$, let $\mathbf{X}$ and $\mathbf{Y}$ have initial states $x, y \in \mathcal{Q}_n$, respectively, where $\max(\|x\|_1, \|y\|_1) \leq cn$ and $\max(\|x\|_\infty, \|y\|_\infty) \leq \beta t$, and let $\mathbf{X}$ and $\mathbf{Y}$ be coupled by a profile coupling. Then*

$$\mathbb{E}\left[d_p(X_t, Y_t)\right] \leq 2cn\left(e^{-\beta t} + 2e^{-\beta n}\right)$$

for all $t \geq 0$.

Proof. Let m and $\mathbf{Z}^i$ be as in Definition 3.6. By Lemma 3.3, we have

$$m = d_p(x, y) \leq \|x\|_1 + \|y\|_1 \leq 2cn,$$

and

$$\|z_i\|_1 \leq \max(\|x\|_1, \|y\|_1) \leq cn, \quad \|z_i\|_\infty \leq \max(\|x\|_\infty, \|y\|_\infty) \leq \beta t,$$

for all $0 \leq i \leq m$. Hence

$$\begin{aligned}\mathbb{E}\left[d_p(X_t, Y_t)\right] &\leq \sum_{i=1}^{m} \mathbb{E}\left[d_p\left(Z_t^{i-1}, Z_t^i\right)\right] \\ &\leq \sum_{i=1}^{m}\left(e^{-\beta t} + 2e^{-\beta n}\right) \leq 2cn\left(e^{-\beta t} + 2e^{-\beta n}\right),\end{aligned}$$

and we are done. □

The following is the main result of this section.

Theorem 3.14. *Let $c > \frac{\lambda}{1-\lambda}$. Then there exists $\eta = \eta(c) > 0$ such that the following holds. Let $n \geq 1$, let $\mathbf{X}$ have an arbitrary initial distribution, let $\mathbf{Y}$ be in equilibrium, and let $\mathbf{X}$ and $\mathbf{Y}$ be coupled by a profile coupling. Then*

$$\mathbb{P}(X_t \not\equiv Y_t) \leq ne^{-\eta t} + 2e^{-\eta n} + \mathbb{P}(\|X_0\|_1 > cn) + \mathbb{P}(\|X_0\|_\infty > \eta t)$$

for all $t \geq 0$.

Remark. This proof is essentially the same as the proof of Theorem 1.1 in [10], the analogous result for the standard supermarket model.

Proof. First we will define some constants. Let $0 < \beta = \beta(c) < 1$ denote the constant given by Lemma 3.13 (with the same c). Let $\eta_1 = \eta_1(c) > 0$ and $\eta_2 > 0$ denote the constants given by Lemma 2.5 (with the same c). Let

$$\begin{aligned}\eta_3 = \eta_3(c) &:= \tfrac{1}{2}\min(\eta_1, \eta_2\beta, \beta), \\ t^* = t^*(c) &:= \frac{\ln(2c+1)}{\eta_3}.\end{aligned}$$

Let $n^* \geq 1$ be sufficiently large so that

$$4cn + 1 \leq 2e^{\eta_3 n}, \tag{3.14}$$

for all $n \geq n^*$. Finally, let

$$\eta = \eta(c) := \min\left(\eta_3, \frac{\ln 2}{n^*}\right),$$

so that $e^{\eta n^*} \leq 2$.

Note that if $t \leq t^*$ and $n \geq n^*$, then $ne^{-\eta t} \geq ne^{-\eta_3 t} \geq ne^{-\eta_3 t^*} \geq 1$. Similarly, if $n \leq n^*$, then $2e^{-\eta n} \geq 2e^{-\eta n^*} \geq 1$. Hence, we will assume that $t \geq t^*$ and $n \geq n^*$, since there is nothing to prove otherwise.

Let A, B_t, C and D_t denote the events that $\|X_0\|_1 \leq cn$, $\|X_0\|_\infty \leq \beta t$, $\|Y_0\|_1 \leq cn$ and $\|Y_0\|_\infty \leq \beta t$, respectively. Then

$$\begin{aligned}\mathbb{P}\left(X_t \not\equiv Y_t\right) &\leq \mathbb{E}\left[\mathbf{1}_{X_t \not\equiv Y_t}\mathbf{1}_A\mathbf{1}_{B_t}\mathbf{1}_C\mathbf{1}_{D_t}\right] + \mathbb{P}\left(\overline{A} \cup \overline{B_t} \cup \overline{C} \cup \overline{D_t}\right) \\ &\leq \mathbb{E}\left[d_p\left(X_t, Y_t\right)\mathbf{1}_A\mathbf{1}_{B_t}\mathbf{1}_C\mathbf{1}_{D_t}\right] + \mathbb{P}\left(\overline{A}\right) + \mathbb{P}\left(\overline{B_t}\right) + \mathbb{P}\left(\overline{C}\right) + \mathbb{P}\left(\overline{D_t}\right). \quad (3.15)\end{aligned}$$

The constant $0 < \beta = \beta(c) < 1$ given by Lemma 3.13 satisfies

$$\mathbb{E}\left[d_p\left(X_t, Y_t\right)\mathbf{1}_A\mathbf{1}_{B_t}\mathbf{1}_C\mathbf{1}_{D_t}\right] \leq 2cn\left(e^{-\beta t} + 2e^{-\beta n}\right), \quad (3.16)$$

and the constants $\eta_1 = \eta_1(c) > 0$ and $\eta_2 > 0$ given by Lemma 2.5 satisfy

$$\mathbb{P}\left(\overline{C}\right) \leq e^{-\eta_1 n}, \quad \mathbb{P}\left(\overline{D_t}\right) \leq ne^{-\eta_2 \beta t}. \quad (3.17)$$

Hence, by (3.15)-(3.17), we have

$$\begin{aligned}\mathbb{P}\left(X_t \not\equiv Y_t\right) &\leq 2cn\left(e^{-\beta t} + 2e^{-\beta n}\right) + \mathbb{P}\left(\overline{A}\right) + \mathbb{P}\left(\|X_0\|_\infty > \beta t\right) + e^{-\eta_1 n} + ne^{-\eta_2 \beta t} \\ &\leq (2c+1)\,ne^{-2\eta_3 t} + (4cn+1)\,e^{-2\eta_3 n} + \mathbb{P}\left(\|X_0\|_1 > cn\right) + \mathbb{P}\left(\|X_0\|_\infty > 2\eta_3 t\right).\end{aligned}$$

Now $2c + 1 \leq e^{\eta_3 t}$ (since $t \geq t^*$) and $4cn + 1 \leq 2e^{\eta_3 n}$ (by (3.14)), so

$$\mathbb{P}\left(X_t \not\equiv Y_t\right) \leq ne^{-\eta_3 t} + 2e^{-\eta_3 n} + \mathbb{P}\left(\|X_0\|_1 > cn\right) + \mathbb{P}\left(\|X_0\|_\infty > \eta_3 t\right).$$

Hence the result follows if $\eta \leq \eta_3$. □

Before we close this chapter, we present a related result concerning the expected profile-distance.

Lemma 3.15. *Let $c > \frac{\lambda}{1-\lambda}$. Then there exists $\eta = \eta(c) > 0$ such that the following holds. Let $n \geq 1$, let $\mathbf{X}$ have any initial distribution where $\mathbb{E}\left[\|X_0\|_1\right] < \infty$, let $\mathbf{Y}$ be in equilibrium, let X_0 and Y_0 be independent, and let $\mathbf{X}$ and $\mathbf{Y}$ be coupled by a profile coupling. Then*

$$\begin{aligned}\mathbb{E}\left[d_p\left(X_t, Y_t\right)\right] &\leq 2cne^{-\eta t} + 6cne^{-\eta n} \\ &\quad + 2\mathbb{E}\left[\|X_0\|_1 \mathbf{1}_{\|X_0\|_1 > cn}\right] + 2cn\mathbb{P}\left(\max\left(\|X_0\|_\infty, \|Y_0\|_\infty\right) > \eta t\right)\end{aligned}$$

for all $t \geq 0$.

Proof. Let $0 < \beta = \beta(c) < 1$ denote the constant given by Lemma 3.13 (with the same c). Let A, B_t, C and D_t denote the events that $\|X_0\|_1 \leq cn$, $\|X_0\|_\infty \leq \beta t$, $\|Y_0\|_1 \leq cn$ and $\|Y_0\|_\infty \leq \beta t$, respectively. Then

$$\mathbb{E}\left[d_p\left(X_t, Y_t\right)\mathbf{1}_A\mathbf{1}_{B_t}\mathbf{1}_C\mathbf{1}_{D_t}\right] \leq 2cn\left(e^{-\beta t} + 2e^{-\beta n}\right). \quad (3.18)$$

By Lemma 3.7 and (3.1), we have

$$\begin{aligned}\mathbb{E}\left[d_p\left(X_t,Y_t\right)\mathbf{1}_{\overline{A}\cup\overline{B_t}\cup\overline{C}\cup\overline{D_t}}\right] &\leq \mathbb{E}\left[d_p\left(X_0,Y_0\right)\mathbf{1}_{\overline{A}\cup\overline{B_t}\cup\overline{C}\cup\overline{D_t}}\right]\\ &\leq \mathbb{E}\left[\left(\left\|X_0\right\|_1+\left\|Y_0\right\|_1\right)\left(\mathbf{1}_{\overline{A}}+\mathbf{1}_A\mathbf{1}_{\overline{C}}+\mathbf{1}_A\mathbf{1}_C\mathbf{1}_{\overline{B_t}\cup\overline{D_t}}\right)\right].\end{aligned}$$

By Lemma 2.5 (1), we have $\mathbb{E}\left[\left\|Y_0\right\|_1\right]\leq\frac{\lambda n}{1-\lambda}\leq cn$, so

$$\begin{aligned}\mathbb{E}\left[d_p\left(X_t,Y_t\right)\mathbf{1}_{\overline{A}\cup\overline{B_t}\cup\overline{C}\cup\overline{D_t}}\right] &\leq \mathbb{E}\left[\left\|X_0\right\|_1\mathbf{1}_{\overline{A}}\right]+\mathbb{E}\left[\left\|Y_0\right\|_1\mathbf{1}_{\overline{A}}\right]\\ &\quad+2cn\mathbb{P}\left(\overline{C}\right)+2cn\mathbb{P}\left(\overline{B_t}\cup\overline{D_t}\right).\end{aligned}\tag{3.19}$$

Let $\eta_1=\eta_1(c)>0$ denote the constant given by Lemma 2.5 (with the same c). Then Lemma 2.5 (1) and the independence of X_0 and Y_0 give the inequalities

$$\mathbb{P}\left(\overline{C}\right)\leq e^{-\eta_1 n},\quad \mathbb{E}\left[\left\|Y_0\right\|_1\mathbf{1}_{\overline{A}}\right]\leq cn\mathbb{P}\left(\overline{A}\right)\leq\mathbb{E}\left[\left\|X_0\right\|_1\mathbf{1}_{\overline{A}}\right].\tag{3.20}$$

By (3.18)-(3.20), we have

$$\begin{aligned}\mathbb{E}\left[d_p\left(X_t,Y_t\right)\right] &\leq \mathbb{E}\left[d_p\left(X_t,Y_t\right)\mathbf{1}_A\mathbf{1}_{B_t}\mathbf{1}_C\mathbf{1}_{D_t}\right]+\mathbb{E}\left[d_p\left(X_t,Y_t\right)\mathbf{1}_{\overline{A}\cup\overline{B_t}\cup\overline{C}\cup\overline{D_t}}\right]\\ &\leq 2cn\left(e^{-\beta t}+2e^{-\beta n}\right)+2\mathbb{E}\left[\left\|X_0\right\|_1\mathbf{1}_{\overline{A}}\right]+2cne^{-\eta_1 n}+2cn\mathbb{P}\left(\overline{B_t}\cup\overline{D_t}\right),\end{aligned}$$

and the result follows if $\eta:=\min\left(\beta,\eta_1\right)$. □

Chapter 4

Concentration of measure

In this chapter, we will show some concentration of measure results for lengths processes. This chapter is based on Chapter 4 of [10] by Luczak and McDiarmid.

4.1 General concentration results

We will say that $f : \mathcal{Q}_n \to \mathbb{R}$ is *Lipschitz* if

$$|f(x) - f(y)| \leq d_p(x, y) \tag{4.1}$$

for all $x, y \in \mathcal{Q}_n$. Note that it suffices to check that (4.1) holds for all profile-equivalent and profile-adjacent pairs. To see this, suppose we are given $x, y \in \mathcal{Q}_n$. Let $x = z_0 \sim z_1 \sim \ldots \sim z_m = y$ be a shortest profile-path between x and y. If $m = 0$, so that x and y are profile-equivalent, then (4.1) is already assumed to hold. Else if $m \geq 1$, then

$$|f(x) - f(y)| \leq \sum_{i=1}^{m} |f(z_{i-1}) - f(z_i)| \leq \sum_{i=1}^{m} d_p(z_{i-1}, z_i) = m = d_p(x, y).$$

First we will need the following result by McDiarmid [15] which concerns the concentration of functions of random variables which satisfy the *bounded differences inequality*.

Theorem 4.1 ([15], Theorem 3.1)**.** *Let* $\mathbf{W} = (W_1, \ldots, W_n)$ *be a vector of independent random variables where* $W_i : \Omega_i \to \mathbb{R}$ *for all* $1 \leq i \leq n$*. Let* $f : \prod_{i=1}^{n} \Omega_i \to \mathbb{R}$*, and suppose that for all* $1 \leq i \leq n$*, there exists* $c_i > 0$ *such that*

$$\left|f(\mathbf{w}) - f\left(\mathbf{w}'\right)\right| \leq c_i$$

for all $\mathbf{w}, \mathbf{w}' \in \prod_{i=1}^{n} \Omega_i$ *differing only in the* i^{th} *coordinate. Then*

$$\mathbb{P}\left(|f(\mathbf{W}) - \mathbb{E}[f(\mathbf{W})]| \geq w\right) \leq 2\exp\left(-\frac{2w^2}{\sum_{i=1}^{n} c_i^2}\right)$$

for all $w \geq 0$*.*

Now we will show a general concentration of measure result for Lipschitz functions of lengths processes.

Lemma 4.2. *There exists $\eta > 0$ such that the following holds. Let $n \geq 1$, let $\mathbf{X}$ have any initial distribution, and let $f : \mathcal{Q}_n \to \mathbb{R}$ be Lipschitz. Then*

$$\mathbb{P}\left(|f(X_t) - \mathbb{E}[f(X_t)]| \geq y\right) \leq nt \exp\left(-\frac{\eta y^2}{nt + y}\right)$$

for all $t, y > 0$.

Remark. This proof is essentially the same as the proof of Lemma 4.3 in [10], the analogous result for the standard supermarket model. The main difference is that here we deal with arrival and potential departure times together as event times, instead of dealing with them separately.

Proof. Since we may consider the translation $f(X_t) - f(X_0)$ instead of $f(X_t)$, we will assume that $f(X_0) = 0$. For $t \geq 0$, let $N_t := \max\{i \geq 0 : T_i \leq t\}$ denote the number of event times in $[0, t]$. Then

$$|f(X_t)| = |f(X_t) - 0| \leq d_p(X_t, X_0) \leq N_t. \tag{4.2}$$

Now we will define some constants. Let

$$\lambda' := \lambda + 1, \quad \beta := 4e\lambda',$$

then let

$$\rho := \max\left(\sqrt{96e^2\lambda'}, \sqrt{7}\beta\right), \quad \eta := \min\left(\frac{1}{\rho^2}, \frac{1}{48e^2\lambda'}, \frac{\ln 2}{2}\right).$$

Note that if $y \leq \rho\sqrt{nt \ln nt}$, then $nt \exp\left(-\frac{\eta y^2}{nt+y}\right) \geq nt \exp\left(-\frac{y^2}{\rho^2 nt}\right) \geq 1$. Hence, we will assume that $y \geq \rho\sqrt{nt \ln nt}$, since there is nothing to prove otherwise. There are now two cases to consider.

Case 1 $\rho\sqrt{nt \ln nt} \leq y \leq \beta nt$.

First note that the bounds on y and the fact that $\rho \geq \sqrt{7}\beta$ imply that

$$nt \geq \frac{\rho\sqrt{nt \ln nt}}{\beta} \geq \sqrt{7nt \ln nt},$$

from which we deduce that $nt \geq 21$. For such values of nt, we have

$$\frac{4e\lambda' nt}{2^{\lfloor 2e\lambda' nt \rfloor}} \leq \frac{1}{nt}, \quad \frac{29}{nt} \leq \tfrac{1}{4}\sqrt{nt \ln nt} \leq \tfrac{1}{4}y. \tag{4.3}$$

In the latter inequality, we have used the fact that $\rho \geq 1$.

Let $\mathcal{I} = \mathcal{I}(n, t, y)$ denote the set of integers k such that $|k - \lambda' nt| \leq \frac{y}{4e} = \frac{y}{\beta nt}\lambda' nt$. Since $N_t \sim \mathrm{Po}(\lambda' nt)$, Lemma 1.4 (with $\varepsilon = \frac{y}{\beta nt} \leq 1$) gives

$$\mathbb{P}(N_t \notin \mathcal{I}) = \mathbb{P}\left(\left|N_t - \lambda' nt\right| > \frac{y}{\beta nt}\lambda' nt\right) \leq 2e^{-\frac{1}{3}\varepsilon^2\lambda' nt} = 2\exp\left(-\frac{y^2}{48e^2\lambda' nt}\right). \tag{4.4}$$

Applying the lower bound on y and the fact that $\rho \geq \sqrt{96e^2\lambda'}$ gives

$$\mathbb{P}\left(N_t \notin \mathcal{I}\right) \leq 2\exp\left(-\frac{\rho^2 \ln nt}{48e^2\lambda'}\right) \leq \frac{2}{(nt)^2}. \tag{4.5}$$

Now, for any $\mathbb{Z}_+$-valued random variable W and any real $k \geq 1$, we have

$$\mathbb{E}\left[W\mathbf{1}_{W>k}\right] \leq \mathbb{E}\left[W\mathbf{1}_{W>\lfloor k\rfloor}\right] = \sum_{i=\lfloor k\rfloor+1}^{\infty} i\mathbb{P}\left(W=i\right) \leq 2k \sum_{i=\lfloor k\rfloor+1}^{\infty} \mathbb{P}\left(W \geq i\right).$$

Applying this inequality with $W := N_t \sim \mathrm{Po}\left(\lambda' nt\right)$ and $k := 2e\lambda' nt$, along with Lemma 1.4 (noting that $i \geq \lfloor k\rfloor + 1 \geq 2e\lambda' nt$) and (4.3), we have

$$\mathbb{E}\left[N_t\mathbf{1}_{N_t>2e\lambda' nt}\right] \leq 2k \sum_{i=\lfloor k\rfloor+1}^{\infty} \frac{1}{2^i} = 2k \cdot \frac{1}{2^{\lfloor k\rfloor}} \leq \frac{1}{nt}.$$

Hence, (4.2) and (4.5) give

$$\begin{aligned}
\mathbb{E}\left[\left|f\left(X_t\right)\mathbf{1}_{N_t\notin\mathcal{I}}\right|\right] &\leq \mathbb{E}\left[N_t\mathbf{1}_{N_t\notin\mathcal{I}}\right] \\
&= \mathbb{E}\left[N_t\left(\mathbf{1}_{N_t<\lambda' nt-\frac{y}{4e}} + \mathbf{1}_{\lambda' nt+\frac{y}{4e}<N_t\leq 2e\lambda' nt}\right)\right] + \mathbb{E}\left[N_t\mathbf{1}_{N_t>2e\lambda' nt}\right] \\
&\leq 2e\lambda' nt\mathbb{P}\left(N_t \notin \mathcal{I}\right) + \frac{1}{nt} \\
&\leq 2e\lambda' nt\frac{2}{(nt)^2} + \frac{1}{nt} \leq \frac{25}{nt}.
\end{aligned} \tag{4.6}$$

For $t > 0$ and $k \geq 0$, let

$$\mu_{t,k} := \mathbb{E}\left[f\left(X_t\right) \mid N_t = k\right],$$

so that (4.2) gives

$$\min_{k\in\mathcal{I}} \mu_{t,k} \leq \min_{k\in\mathcal{I}} \mathbb{E}\left[N_t \mid N_t = k\right] = \min\mathcal{I} \leq \lambda' nt. \tag{4.7}$$

Now write

$$\mu_t := \mathbb{E}\left[f\left(X_t\right)\right] = \sum_{k\in\mathcal{I}} \mu_{t,k}\mathbb{P}\left(N_t = k\right) + \mathbb{E}\left[f\left(X_t\right)\mathbf{1}_{N_t\notin\mathcal{I}}\right].$$

We may bound μ_t above using (4.6), so that

$$\begin{aligned}
\mu_t &\leq \max_{k\in\mathcal{I}} \mu_{t,k}\mathbb{P}\left(N_t \in \mathcal{I}\right) + \mathbb{E}\left[\left|f\left(X_t\right)\right|\mathbf{1}_{N_t\notin\mathcal{I}}\right] \\
&\leq \max_{k\in\mathcal{I}} \mu_{t,k} + \frac{25}{nt},
\end{aligned} \tag{4.8}$$

and below using (4.5)-(4.7), so that

$$\begin{aligned}\mu_t &\geq \min_{k\in\mathcal{I}} \mu_{t,k}\mathbb{P}\left(N_t \in \mathcal{I}\right) - \mathbb{E}\left[\left|f\left(X_t\right)\right| \mathbf{1}_{N_t\notin\mathcal{I}}\right] \\ &\geq \min_{k\in\mathcal{I}} \mu_{t,k} - \min_{k\in\mathcal{I}} \mu_{t,k}\mathbb{P}\left(N_t \notin \mathcal{I}\right) - \mathbb{E}\left[\left|f\left(X_t\right)\right| \mathbf{1}_{N_t\notin\mathcal{I}}\right] \\ &\geq \min_{k\in\mathcal{I}} \mu_{t,k} - \lambda' nt\frac{2}{(nt)^2} - \frac{25}{nt} \geq \min_{k\in\mathcal{I}} \mu_{t,k} - \frac{29}{nt}. \end{aligned} \tag{4.9}$$

We will also require the following result, which holds by Lemma 3.7: for all $k \geq 0$, we have

$$|\mu_{t,k} - \mu_{t,k+1}| \leq 1.$$

Then since $\mathcal{I}$ is an interval of length at most $\frac{y}{4e} \leq \frac{1}{4}y$, the bounds (4.8) and (4.9), along with (4.3), give

$$|\mu_t - \mu_{t,k}| \leq \tfrac{1}{4}y + \frac{29}{nt} \leq \tfrac{1}{2}y, \tag{4.10}$$

for all $k \in \mathcal{I}$.

For $t > 0$ and $k \geq 0$, let $\mathbb{P}_{t,k}$ denote the probability conditional on $N_t = k$. Then (4.10) and (4.4) give

$$\begin{aligned}\mathbb{P}\left(\left|f\left(X_t\right) - \mu_t\right| \geq y\right) &\leq \sum_{k\in\mathcal{I}} \mathbb{P}_{t,k}\left(\left|f\left(X_t\right) - \mu_t\right| \geq y\right)\mathbb{P}\left(N_t = k\right) + \mathbb{P}\left(N_t \notin \mathcal{I}\right) \\ &\leq \sum_{k\in\mathcal{I}} \mathbb{P}_{t,k}\left(\left|f\left(X_t\right) - \mu_{t,k}\right| \geq \tfrac{1}{2}y\right)\mathbb{P}\left(N_t = k\right) \\ &\qquad + 2\exp\left(-\frac{y^2}{48e^2\lambda' nt}\right). \end{aligned} \tag{4.11}$$

Thus it remains to show that $\mathbb{P}_{t,k}\left(\left|f\left(X_t\right) - \mu_{t,k}\right| \geq \frac{1}{2}y\right)$ is small, for $k \in \mathcal{I}$.

We will use Theorem 4.1 to do this. Recall the definition of $\mathbf{C}^{\mathrm{a}} = \left(C_i^{\mathrm{a}}\right)_{i=1}^{\infty}$ and $\mathbf{S}^{\mathrm{d}} = \left(S_i^{\mathrm{d}}\right)_{i=1}^{\infty}$ in Section 2.1. Conditional on $N_t = k$, X_t depends only on the random variables $C_1^{\mathrm{a}}, \ldots, C_k^{\mathrm{a}}, S_1^{\mathrm{d}}, \ldots, S_k^{\mathrm{d}}$, and none others (in fact, only on exactly k of the random variables, since there are only k event times). Hence, $f\left(X_t\right)$ also only depends on $C_1^{\mathrm{a}}, \ldots, C_k^{\mathrm{a}}, S_1^{\mathrm{d}}, \ldots, S_k^{\mathrm{d}}$. As required by Theorem 4.1, these $2k$ random variables are independent of each other. Next we must verify the bounded differences inequality. Let $\mathbf{x}$ and $\mathbf{y}$ be realisations of lengths processes, with the same initial state and differing only in their choices at one arrival time t_i. Then they are identical until time t_i, which is when the $\mathbf{x}$- and $\mathbf{y}$-customers join possibly different queues. By Lemma 3.7 (and by taking a profile coupling), we have

$$\left|f\left(x_t\right) - f\left(y_t\right)\right| \leq d_p\left(x_t, y_t\right) \leq d_p\left(x_{t_i}, y_{t_i}\right) \leq 2.$$

We may argue similarly if $\mathbf{x}$ and $\mathbf{y}$ differ only in their selection at one potential departure time. Hence, by Theorem 4.1 (with each $c_i = 2$), we have

$$\mathbb{P}_{t,k}\left(\left|f\left(X_t\right) - \mu_{t,k}\right| \geq \tfrac{1}{2}y\right) \leq 2\exp\left(-\frac{2\left(\frac{1}{2}y\right)^2}{\sum_{i=1}^{2k} 2^2}\right) = 2\exp\left(-\frac{y^2}{16k}\right)$$

for all $k \geq 0$. Substituting this into (4.11) then gives

$$\begin{aligned}\mathbb{P}\left(|f\left(X_t\right)-\mu_t| \geq y\right) &\leq 2\sum_{k\in\mathcal{I}} \exp\left(-\frac{y^2}{16k}\right)\mathbb{P}\left(N_t=k\right)+2\exp\left(-\frac{y^2}{48e^2\lambda' nt}\right)\\ &\leq 2\exp\left(-\frac{y^2}{16\left(\lambda' nt+\frac{y}{4e}\right)}\right)+2\exp\left(-\frac{y^2}{48e^2\lambda'\left(nt+y\right)}\right)\\ &\leq nt\exp\left(-\frac{y^2}{48e^2\lambda'\left(nt+y\right)}\right).\end{aligned}$$

The result follows since $\eta \leq \frac{1}{48e^2\lambda'}$.

Case 2 $y \geq \beta nt$.

First note that

$$|f\left(X_t\right)-\mu_t| \leq |f\left(X_t\right)|+|\mu_t| \leq N_t+\mathbb{E}\left[N_t\right]=N_t+\lambda' nt \leq N_t+\tfrac{1}{2}y.$$

Let $\eta_1 := \frac{1}{2}\ln 2$. Since $\frac{1}{2}y \geq 2e\lambda' nt$, Lemma 1.4 gives

$$\mathbb{P}\left(|f\left(X_t\right)-\mu_t| \geq y\right) \leq \mathbb{P}\left(N_t \geq \tfrac{1}{2}y\right) \leq 2^{-\frac{1}{2}y}=e^{-\eta_1 y} \leq \exp\left(-\frac{\eta_1 y^2}{nt+y}\right).$$

The result follows since $\eta \leq \eta_1$. □

We will now show concentration of measure for Lipschitz functions of lengths process in equilibrium.

Lemma 4.3. *There exists $\eta > 0$ such that the following holds. Let $n \geq 1$, let X have the equilibrium distribution for the lengths process, and let $f : \mathcal{Q}_n \to \mathbb{R}$ be Lipschitz. Then*

$$\mathbb{P}\left(|f\left(X\right)-\mathbb{E}\left[f\left(X\right)\right]| \geq y\right) \leq n^2\exp\left(-\frac{\eta y}{\sqrt{n}}\right)$$

for all $y > 0$.

Remark. This proof is essentially the same as the proof of Lemma 4.1 in [10], the analogous result for the standard supermarket model.

Proof. Let $\mathbf{X}$ be in equilibrium, and let $\mathbf{0}_1 := \left(\left(0,\ldots,0\right),1\right)$ denote the empty state with memory queue 1. Since we may consider the translation $f\left(X_t\right)-f\left(\mathbf{0}_1\right)$ instead of $f\left(X_t\right)$, we will assume that $f\left(\mathbf{0}_1\right)=0$.

Now we will define some constants. Let $\eta_1=\eta_1\left(c\right)>0$ and $\eta_2>0$ denote the constants given by Lemma 2.5 with $c := \frac{2\lambda}{1-\lambda}$. Let $\eta_3=\eta_3\left(c\right)>0$ and $\eta_4=\eta_4\left(c\right)>0$ denote the constants given by Theorem 3.14 and Lemma 3.15 with $c := \frac{2\lambda}{1-\lambda}$, respectively. Let $\eta_5=\eta_5\left(c\right) := \min\left(\eta_2,\eta_2\eta_4\right)>0$. Let $\eta_6>0$ denote the constant given by Lemma 4.2. Let

$$\beta := \max\left(\frac{8\lambda}{1-\lambda},\frac{1}{2}\right), \quad \eta_7=\eta_7\left(c\right) := \tfrac{1}{2}\min\left(\eta_3,\frac{\eta_3}{\beta},\eta_5,\frac{\eta_5}{\beta},\frac{\eta_6}{6}\right).$$

Let $n^* \geq 1$ be sufficiently large so that if $n \geq 1$ satisfies $\frac{n}{\ln n} \geq n^*$, then

$$\frac{10c}{n} \leq \frac{3}{2\eta_7}\sqrt{n}\ln n, \quad \beta+3 \leq n^3. \tag{4.12}$$

Let

$$\rho = \rho(c) := \max\left(\frac{3}{\eta_7}, \beta n^*\right), \quad \eta = \eta(c) := \min\left(\frac{1}{\rho}, \eta_1, \eta_7\right).$$

Note that if $y \leq \rho\sqrt{n}\ln n$, then $n^2 \exp\left(-\frac{\eta y}{\sqrt{n}}\right) \geq n\exp\left(-\frac{y}{\rho\sqrt{n}}\right) \geq 1$. Hence, we will assume that $y \geq \rho\sqrt{n}\ln n$, since there is nothing to prove otherwise. There are now two cases to consider.

Case 1 $\rho\sqrt{n}\ln n \leq y \leq \beta n^{3/2}$.

Let $t := \frac{y}{\sqrt{n}}$. Then the bounds on y imply that

$$\max\left(\frac{3}{\eta_7}, \beta n^*\right)\ln n = \rho \ln n \leq t = \frac{y}{\sqrt{n}} \leq \beta n. \tag{4.13}$$

This implies that $\frac{n}{\ln n} \geq n^*$, whence (4.12) holds, and that

$$n^3 \leq e^{\eta_7 t}. \tag{4.14}$$

Let $\mathbf{Y}$ be started from $\mathbf{0}_1$, and let $\mathbf{X}$ and $\mathbf{Y}$ be coupled by a profile coupling. First note that

$$|f(X_t) - f(Y_t)| \leq d_p(X_t, Y_t),$$

whence the constant $\eta_4 > 0$ (given by Lemma 3.15) satisfies

$$|\mathbb{E}[f(X_t) - f(Y_t)]| \leq \mathbb{E}[d_p(X_t, Y_t)] \leq 2cn\left(e^{-\eta_4 t} + 3e^{-\eta_4 n} + \mathbb{P}\left(\|X_0\|_\infty > \eta_4 t\right)\right).$$

Hence, the constant $\eta_2 > 0$ (given by Lemma 2.5) satisfies

$$|\mathbb{E}[f(X_t) - f(Y_t)]| \leq 2cn\left(e^{-\eta_4 t} + 3e^{-\eta_4 n} + ne^{-\eta_2\eta_4 t}\right) \leq 2cn^2\left(2e^{-\eta_5 t} + 3e^{-\eta_5 n}\right).$$

Then, by (4.13), (4.14) and then (4.12), we have

$$\begin{aligned}
|\mathbb{E}[f(X_t) - f(Y_t)]| &\leq 2cn^2\left(2e^{-\eta_5 t} + 3e^{-\eta_5 t/\beta}\right) \\
&\leq 10cn^2 e^{-\eta_7 t} \\
&\leq \frac{10c}{n} \leq \frac{3}{2\eta_7}\sqrt{n}\ln n \leq \tfrac{1}{2}\rho\sqrt{n}\ln n \leq \tfrac{1}{2}y.
\end{aligned}$$

By the triangle inequality, we have

$$\begin{aligned}
|f(X_t) - \mathbb{E}[f(X_t)]| &\leq |f(X_t) - f(Y_t)| + |f(Y_t) - \mathbb{E}[f(Y_t)]| + |\mathbb{E}[f(Y_t)] - \mathbb{E}[f(X_t)]| \\
&\leq d_p(X_t, Y_t) + |f(Y_t) - \mathbb{E}[f(Y_t)]| + \tfrac{1}{2}y.
\end{aligned}$$

It follows that the constants $\eta_3, \eta_6 > 0$ (given by Theorem 3.14 and Lemma 4.2, respectively) satisfy

$$\begin{aligned}
\mathbb{P}\left(|f(X_t) - \mathbb{E}[f(X_t)]| \geq y\right) &\leq \mathbb{P}\left(d_p(X_t, Y_t) > 0\right) + \mathbb{P}\left(|f(Y_t) - \mathbb{E}[f(Y_t)]| \geq \tfrac{1}{2}y\right) \\
&\leq ne^{-\eta_3 t} + 2e^{-\eta_3 n} + nt\exp\left(-\frac{\eta_6\left(\frac{1}{2}y\right)^2}{nt + \frac{1}{2}y}\right).
\end{aligned}$$

By (4.13) and the fact that $y = t\sqrt{n}$, we have

$$\begin{aligned}\mathbb{P}\left(\left|f\left(X_t\right) - \mathbb{E}\left[f\left(X_t\right)\right]\right| \geq y\right) &\leq ne^{-\eta_3 t} + 2e^{-\eta_3 t/\beta} + \beta n^2 \exp\left(-\frac{\eta_6 nt}{2\left(2n+\sqrt{n}\right)}\right) \\ &\leq n^2\left(e^{-\eta_3 t} + 2e^{-\eta_3 t/\beta} + \beta e^{-\frac{1}{6}\eta_6 t}\right) \\ &\leq \left(\beta+3\right) n^2 e^{-2\eta_7 t} \leq n^2 e^{-\eta_7 t}.\end{aligned}$$

The last inequality holds by (4.12) and (4.14). The result follows since $\eta \leq \eta_5$.

Case 2 $y > \beta n^{3/2}$.

First note that, by Lemma 3.3, we have

$$\left|f\left(X\right)\right| = \left|f\left(X\right) - 0\right| \leq d_p\left(X, \mathbf{0}_1\right) \leq \left\|X\right\|_1,$$

whence Lemma 2.5 gives

$$\left|\mathbb{E}\left[f\left(X\right)\right]\right| \leq \mathbb{E}\left[\left\|X\right\|_1\right] \leq \frac{\lambda n}{1-\lambda} < \frac{4\lambda n}{1-\lambda} \leq \tfrac{1}{2}\beta n^{3/2} < \tfrac{1}{2}y, \tag{4.15}$$

and thus

$$\left|f\left(X\right) - \mathbb{E}\left[f\left(X\right)\right]\right| < \left\|X\right\|_1 + \tfrac{1}{2}y.$$

Since $y > \beta n^{3/2} \geq \frac{1}{2}$, we have $y > \frac{1}{2}\lceil y \rceil$, and thus

$$\mathbb{P}\left(\left|f\left(X\right) - \mathbb{E}\left[f\left(X\right)\right]\right| \geq y\right) \leq \mathbb{P}\left(\left\|X\right\|_1 > \tfrac{1}{2}y\right) \leq \mathbb{P}\left(\left\|X\right\|_1 > \tfrac{1}{4}\lceil y \rceil\right).$$

Using (4.15), we see that $\frac{1}{4}\lceil y \rceil > \frac{2\lambda n}{1-\lambda}$, so the constant $\eta_1 = \eta_1\left(c\right) > 0$ (given by Lemma 2.5) satisfies

$$\mathbb{P}\left(\left|f\left(X\right) - \mathbb{E}\left[f\left(X\right)\right]\right| \geq y\right) \leq e^{-\eta_1 \lceil y \rceil} \leq n^2 \exp\left(-\frac{\eta_1 y}{\sqrt{n}}\right).$$

The result follows since $\eta \leq \eta_1$. □

4.2 Concentration of the tail functions

In this section, we will apply our concentration of measure results to the functions $l_i\left(\cdot\right)$, which give the number of queues of length at least i; these were defined in (2.1). However, we will express our results in terms of the *tail functions*

$$u_i\left(x\right) := \tfrac{1}{n} l_i\left(x\right),$$

which give the proportion of queues in $x \in \mathcal{Q}_n$ of length at least $i \geq 1$.

Note that the $l_i\left(\cdot\right)$ are Lipschitz, for all $i \geq 1$, since (4.1) holds for all $x', y' \in \mathcal{Q}_n$ such that $x' \equiv y'$ or $x' \sim y'$. The first lemma bounds the equilibrium deviation of the tail functions from their means, over long periods of time.

Lemma 4.4. *Let $z > 0$. Then there exists $\eta = \eta\left(z\right) > 0$ such that the following holds. Let*

$n \geq 1$, *and let* X *have the equilibrium distribution for the lengths process. Then*

$$\mathbb{P}\left(|u_i(X) - \mathbb{E}[u_i(X)]| \geq z \text{ for some } 0 \leq t \leq e^{\eta\sqrt{n}}\right) \leq 2e^{-\eta\sqrt{n}}$$

for all $i \geq 1$.

Proof. Since the left-hand side is bounded by 1 and $\eta > 0$ may be arbitrarily small, it suffices to show the result for all sufficiently large n.

Let $\mathbf{X}$ be in equilibrium. For $i \geq 1$, $t \geq 0$ and $h > 0$, let

$$E_{i,t,h} := \{|u_i(X_t) - \mathbb{E}[u_i(X_t)]| \geq h\}.$$

Then $\overline{E_{i,t,z/2}}$ holds with high probability at each individual time, since Lemma 4.3 (with $l_i(\cdot)$ and $y := \frac{1}{2}zn$) gives $\eta_1 > 0$ such that

$$\begin{aligned}\mathbb{P}\left(E_{i,t,z/2}\right) &= \mathbb{P}\left(|l_i(X_t) - \mathbb{E}[l_i(X_t)]| \geq \tfrac{1}{2}zn\right) \\ &\leq n^2 e^{-\frac{1}{2}\eta_1 z\sqrt{n}} \leq e^{-\frac{1}{4}\eta_1 z\sqrt{n}},\end{aligned} \tag{4.16}$$

for all $i \geq 1$, $t \geq 0$ and $z > 0$, if n is sufficiently large.

Now we will extend this to the interval $\left[0, e^{\eta\sqrt{n}}\right]$, where

$$\eta = \eta(z) := \tfrac{1}{3}\min\left(\tfrac{1}{4}\eta_1, \tfrac{1}{12}\right) z.$$

Consider covering this with sub-intervals of length $\delta = \delta(z) := \frac{z}{4(\lambda+1)}$; clearly $m = m(z) := \left\lceil \frac{e^{\eta\sqrt{n}}}{\delta} \right\rceil$ such sub-intervals will cover $\left[0, e^{\eta\sqrt{n}}\right]$. For $k \geq 0$, let $t_k := k\delta$, then

$$\mathbb{P}\left(\bigcup_{0 \leq t \leq e^{\eta\sqrt{n}}} E_{i,t,z}\right) \leq \sum_{k=0}^{m} \mathbb{P}\left(E_{i,t_k,z/2}\right) + m\mathbb{P}\left(\mathrm{Po}\left(\tfrac{1}{4}zn\right) \geq \tfrac{1}{2}zn\right).$$

To see the last term in this inequality, suppose that $\overline{E_{i,t_k,z/2}}$ holds for all end-points t_k. Then there exists a sub-interval $\mathcal{I}_l := (t_{l-1}, t_l)$ containing t. Since $\overline{E_{i,t_{l-1},z/2}}$ and $E_{i,t,z}$ hold, we deduce that over $\mathcal{I}_l$ the proportion of queues of length at least i changes by at least $\frac{1}{2}z$, and thus over $\mathcal{I}_l$, we have at least $\frac{1}{2}zn$ events. However, the number of events over $\mathcal{I}_l$, an interval of length δ, is Poisson with mean $(\lambda+1)\delta n = \frac{1}{4}zn$. By (4.16) and Lemma 1.4 (with $\varepsilon = 1$), we have

$$\mathbb{P}\left(\bigcup_{0 \leq t \leq e^{\eta\sqrt{n}}} E_{i,t,z}\right) \leq (m+1)\, e^{-\frac{1}{4}\eta_1 z\sqrt{n}} + m\left(2e^{-\frac{1}{12}zn}\right).$$

Straightforward manipulation gives

$$\mathbb{P}\left(\bigcup_{0 \leq t \leq e^{\eta\sqrt{n}}} E_{i,t,z}\right) \leq 2\left(\frac{e^{\eta\sqrt{n}}}{\delta} + 2\right) e^{-3\eta\sqrt{n}} \leq \frac{2 \cdot 3e^{\eta\sqrt{n}}}{\min(\delta, 1)} e^{-3\eta\sqrt{n}} \leq 2e^{-\eta\sqrt{n}},$$

if n is sufficiently large. □

The second lemma uniformly bounds the deviation of the equilibrium tail functions

from their means.

Lemma 4.5. *Let $z > 0$. Then there exists $\eta = \eta(z) > 0$ such that the following holds. Let $n \geq 1$, and let X have the equilibrium distribution for the lengths process. Then*

$$\mathbb{P}\left(\sup_{i\geq 1} |u_i(X) - \mathbb{E}[u_i(X)]| \geq \frac{z\ln^2 n}{\sqrt{n}}\right) \leq 2e^{-\eta \ln^2 n}.$$

Proof. Since the left-hand side is bounded by 1 and $\eta > 0$ may be arbitrarily small, it suffices to show the result for all sufficiently large n.

Let $c := \frac{2\lambda}{1-\lambda}$. First consider the case where the supremum of $|u_i(X) - \mu_i|$ is attained at some $i \geq cn$. By Lemma 2.5, there exists $\eta_1 = \eta_1(c) > 0$ such that

$$\mathbb{P}\left(\sup_{i\geq cn} |u_i(X) - \mathbb{E}[u_i(X)]| \geq \frac{z\ln^2 n}{\sqrt{n}}\right) \leq \mathbb{P}\left(l_{\lceil cn\rceil}(X) \geq 1\right) \leq \mathbb{P}\left(\|X\|_1 \geq cn\right) = e^{-\eta_1 n}.$$

Next consider the case where the supremum is attained at some $i \leq cn$. By Lemma 4.3 (with $l_i(\cdot)$ and $y := z\sqrt{n}\ln^2 n$), there exists $\eta_2 > 0$ such that

$$\mathbb{P}\left(\sup_{i\leq cn} |u_i(X) - \mathbb{E}[u_i(X)]| \geq \frac{z\ln^2 n}{\sqrt{n}}\right) \leq n^2 e^{-z\eta_2 \ln^2 n} \leq e^{-\frac{1}{2}\eta_2 z \ln^2 n},$$

if n is sufficiently large. Hence the result follows by taking $\eta = \eta(z) := \min\left(\eta_1, \frac{1}{2}\eta_2 z\right)$. □

The third lemma uniformly bounds the deviation of powers of the equilibrium tail functions from the same powers of their means.

Lemma 4.6. *Let $r \geq 2$ be an integer. Then there exists $c = c(r) > 0$ such that the following holds. Let $n \geq 1$, and let X have the equilibrium distribution for the lengths process. Then*

$$\sup_{i\geq 1} |\mathbb{E}[u_i(X)^r] - \mathbb{E}[u_i(X)]^r| \leq \frac{c\ln^2 n}{n}.$$

Proof. Since the left-hand side is bounded by 1 and $c > 0$ may be arbitrarily large, it suffices to show the result for all sufficiently large n.

For brevity, let $U_i := u_i(X)$ and $\mu_i := \mathbb{E}[U_i]$. Let $\eta > 0$ denote the constant given by Lemma 4.3, then let $c_1 = c_1(r) := \frac{r+2}{\eta}$. By Lemma 4.3 with $y := \frac{c_1 \ln n}{\sqrt{n}}$, we have

$$\begin{aligned}\mathbb{P}(|U_i - \mu_i| \geq y) &= \mathbb{P}\left(|l_i(X) - \mathbb{E}[l_i(X)]| \geq c_1\sqrt{n}\ln n\right)\\ &\leq n^2 e^{-\eta c_1 \ln n} = n^{-r},\end{aligned}$$

and thus

$$\begin{aligned}\mathbb{E}[|U_i - \mu_i|^s] &\leq \mathbb{E}\left[|U_i - \mu_i|^s \mathbf{1}_{|U_i-\mu_i|\leq y}\right] + \mathbb{E}\left[|U_i - \mu_i|^s \mathbf{1}_{|U_i-\mu_i|\geq y}\right]\\ &\leq y^s + \mathbb{P}(|U_i - \mu_i| \geq y)\\ &\leq \left(\frac{c_1 \ln n}{\sqrt{n}}\right)^s + \frac{1}{n^r} \leq 2c_1^r \left(\frac{\ln n}{\sqrt{n}}\right)^s,\end{aligned}$$

for all $1 \leq s \leq r$, if n is sufficiently large. This gives

$$\begin{aligned}
0 \leq \mathbb{E}\left[U_i^r\right] - \mu_i^r &\leq \mathbb{E}\left[\sum_{s=0}^{r} \binom{r}{s} (U_i - \mu_i)^s \mu_i^{r-s}\right] - \mu_i^r \\
&= \sum_{s=2}^{r} \binom{r}{s} \mathbb{E}\left[(U_i - \mu_i)^s\right] \mu_i^{r-s} \\
&\leq 2^r \sum_{s=2}^{r} \mathbb{E}\left[|U_i - \mu_i|^s\right] \\
&\leq 2^{r+1} c_1^r \sum_{s=2}^{r} \left(\frac{\ln n}{\sqrt{n}}\right)^s \leq (r-1)\, 2^{r+1} c_1^r \frac{\ln^2 n}{n}.
\end{aligned}$$

The result follows by taking $c = c(r) := (r-1)\, 2^{r+1} c_1^r$. □

Chapter 5

Tail functions and memory queue length

In this chapter, we will analyse the equilibrium behaviour of the tail functions $u_i(\cdot)$ and the indicators $\mathbf{1}_{v(\cdot)\geq i}$, where

$$v(x) := x(\xi)$$

is the length of the memory queue in x. This chapter is based on Chapter 5 of [10] by Luczak and McDiarmid.

5.1 Balance equations

In this section, we will determine the balance equations for the tail functions and the indicators $\mathbf{1}_{v(\cdot)\geq i}$.

Lemma 5.1. *Let $n \geq 1$, let $\mathbf{X}$ be in equilibrium, and let $\mathcal{G}$ denote the generator operator of $\mathbf{X}$. For $t \geq 0$ and $i \geq 0$, let*

$$U_{t,i} := u_i(X_t), \quad V_t := v(X_t), \quad P_{t,i} = \sum_{s=1}^{d} U_{t,i}^{s-1}\left(U_{t,i-1} - U_{t,i}\right)\left(U_{t,i} + \tfrac{1}{n}\right)^{d-s}.$$

Then

$$\mathcal{G}U_{t,i} = \lambda\left(U_{t,i-1}^d \mathbf{1}_{V_t\geq i-1} - U_{t,i}^d \mathbf{1}_{V_t\geq i}\right) - \left(U_{t,i} - U_{t,i+1}\right), \tag{5.1}$$

$$\begin{aligned}\mathcal{G}\mathbf{1}_{V_t\geq i} = \lambda n\Big[\left(U_{t,i} + \tfrac{1}{n}\right)^d \mathbf{1}_{V_t\geq i-1} - \left(1 + \left(U_{t,i} + \tfrac{1}{n}\right)^d - U_{t,i}^d - P_{t,i}\right)\mathbf{1}_{V_t\geq i}\Big] \\ - \left[\mathbf{1}_{V_t\geq i} - \mathbf{1}_{V_t\geq i+1}\right],\end{aligned} \tag{5.2}$$

for all $t \geq 0$ and $i \geq 1$.

Proof. To show (5.1), we will show that

$$\mathcal{G}U_{t,i} = \lambda n\left[\tfrac{1}{n}\left(U_{t,i-1}^d - U_{t,i}^d\right)\mathbf{1}_{V_t\geq i} + \tfrac{1}{n}U_{t,i-1}^d \mathbf{1}_{V_t=i-1}\right] + n\left[-\tfrac{1}{n}\left(U_{t,i} - U_{t,i+1}\right)\right]. \tag{5.3}$$

Here, the terms in the two square brackets in (5.3) correspond to changes in $U_{t,i}$ at arrival and potential departure times, respectively. The factors λn and n correspond to the fact that arrivals and potential departures occur at rate λn and n, respectively.

At an arrival time, $U_{t,i}$ can only change by $+\frac{1}{n}$. This occurs if and only if the customer joins a queue of length $i-1$, which is if and only if the shortest queue in the candidates list has length $i-1$. If $V_t > i-1$ immediately before the arrival, then $U_{t,i}$ changes by $+\frac{1}{n}$ if and only if the customer selects a shortest queue of length $i-1$; this gives the term $\frac{1}{n}\left(U^d_{t,i-1} - U^d_{t,i}\right)\mathbf{1}_{V_t \geq i}$. Else if $V_t = i-1$ immediately before the arrival, then $U_{t,i}$ changes by $+\frac{1}{n}$ if and only if the customer selects only queues of length at least $i-1$; this gives the term $\frac{1}{n}U^d_{t,i-1}\mathbf{1}_{V_t=i-1}$.

At a potential departure time, $U_{t,i}$ can only change by $-\frac{1}{n}$. This occurs if and only if the selection has length i; this gives the term $-\frac{1}{n}\left(U_{t,i} - U_{t,i+1}\right)$. Thus, (5.3) holds.

Now we point out that the term

$$P_{t,i} = \sum_{s=1}^{d} U^{s-1}_{t,i}\left(U_{t,i-1} - U_{t,i}\right)\left(U_{t,i} + \tfrac{1}{n}\right)^{d-s}$$

corresponds to the probability of an arriving customer selecting a unique shortest queue of length $i-1$. For if such a queue exists, then there exists $1 \leq s \leq d$ such that choice s is a queue of length $i-1$, choices $1, \ldots, s-1$ are queues of length at least i, and choices $s+1, \ldots, d$ are queues of length at least i or the same as choice s.

To show (5.2), we will show that

$$\mathcal{G}\mathbf{1}_{V_t \geq i} = \lambda n\left[\left(U_{t,i} + \tfrac{1}{n}\right)^d \mathbf{1}_{V_t=i-1} - \left(1 - U^d_{t,i} - P_{t,i}\right)\mathbf{1}_{V_t \geq i}\right] + n\left[-\tfrac{1}{n}\mathbf{1}_{V_t=i}\right]. \tag{5.4}$$

Again, the terms in the two square brackets in (5.4) correspond to changes in $\mathbf{1}_{V_t \geq i}$ at arrival and potential departure times, respectively, and the factors λn and n correspond to the fact that arrivals and potential departures occur at rate λn and n, respectively.

At an arrival time, $\mathbf{1}_{V_t \geq i}$ can change by $+1$ or -1. Now $\mathbf{1}_{V_t \geq i}$ changes by $+1$ if and only if $V_t = i-1$ immediately before the arrival, and if the customer selects only queues of length at least i or the memory queue; this gives the term $\left(U_{t,i} + \frac{1}{n}\right)^d \mathbf{1}_{V_t=i-1}$. On the other hand, $\mathbf{1}_{V_t \geq i}$ changes by -1 if and only if $V_t \geq i$ immediately before the arrival, and if the customer selects some queue shorter than i, but he/she does not end up selecting a unique shortest queue of length $i-1$ (for the memory queue would then have length i at time t); this gives the term $-\left(1 - U^d_{t,i} - P_{t,i}\right)\mathbf{1}_{V_t \geq i}$.

At a potential departure time, $\mathbf{1}_{V_t \geq i}$ can only change by -1. This occurs if and only if $V_t = i$ immediately before the potential departure, and if the selection is the memory queue; this gives the term $-\frac{1}{n}\mathbf{1}_{V_t=i}$. Thus, (5.4) holds. □

5.2 Approximate recurrence relations

In this section, we will show that the equilibrium means of the tail and indicator functions closely follow two families of recurrence relations. As mentioned in Section 1.3, these relations also appear in [21, 13].

For the rest of this chapter, we will let X have the equilibrium distribution for the lengths process, for some $n \geq 1$. In this case, for $i \geq 0$, let

$$U_i := u_i(X), \quad V := v(X),$$

and

$$\mu_i := \mathbb{E}\left[U_i\right], \quad \nu_i := \mathbb{E}\left[\mathbf{1}_{V\geq i}\right].$$

Thus the μ_i and ν_i all depend on n.

We begin our analysis by taking expectations of the balance equations.

Lemma 5.2. *There exists $c_1 > 0$ such that the following holds. Let $n \geq 1$, and let X have the equilibrium distribution for the lengths process. Then*

$$\mu_i = \lambda\mathbb{E}\left[U_{i-1}^d \mathbf{1}_{V\geq i-1}\right] \leq \lambda^i \tag{5.5}$$

for all $i \geq 1$, and

$$\sup_{i\geq 1}\left|\mathbb{E}\left[U_i^d \mathbf{1}_{V\geq i-1}\right] - \mathbb{E}\left[\left(1 - d\left(U_{i-1} - U_i\right)U_i^{d-1}\right)\mathbf{1}_{V\geq i}\right]\right| \leq \frac{c_1}{n}. \tag{5.6}$$

Proof. Let $\mathbf{X}$ be in equilibrium, then let $\mathcal{G}$ denote the generator operator of $\mathbf{X}$. It is known (e.g., see [5], Chapters 1 and 4) that if $f : \mathcal{Q}_n \to \mathbb{R}$ is bounded, then

$$\mathbb{E}\left[\mathcal{G}f\left(X_t\right)\right] = \frac{d\mathbb{E}\left[f\left(X_t\right)\right]}{dt}.$$

This is 0 since $\mathbf{X}$ is in equilibrium. We will apply this to the bounded functions U_i and $\mathbf{1}_{V\geq i}$. Thus, taking expectations in (5.1) and (5.2), and then rearranging, gives

$$\mu_i - \mu_{i+1} = \lambda\left(\mathbb{E}\left[U_{i-1}^d \mathbf{1}_{V\geq i-1}\right] - \mathbb{E}\left[U_i^d \mathbf{1}_{V\geq i}\right]\right), \tag{5.7}$$

$$\tfrac{1}{n}\left(\nu_i - \nu_{i+1}\right) = \lambda\mathbb{E}\left[\left(U_i + \tfrac{1}{n}\right)^d \mathbf{1}_{V\geq i-1}\right] - \lambda\mathbb{E}\left[\left(1 + \left(U_i + \tfrac{1}{n}\right)^d - U_i^d - P_i\right)\mathbf{1}_{V\geq i}\right], \tag{5.8}$$

for all $i \geq 1$, where

$$P_i = \sum_{s=1}^{d} U_i^{s-1}\left(U_{i-1} - U_i\right)\left(U_i + \tfrac{1}{n}\right)^{d-s}.$$

Now we will show (5.5). By Lemma 2.5, we have $\sum_{k=1}^{\infty}\mu_k = \frac{1}{n}\mathbb{E}\left[\|X\|_1\right] < \infty$, and thus $\lim_{k\to\infty}\mu_k = 0$. Hence, for each $i \geq 1$, we may sum (5.7) over $\{i, i+1, \dots\}$ to obtain

$$\mu_i = \lambda\mathbb{E}\left[U_{i-1}^d \mathbf{1}_{V\geq i-1}\right] \leq \lambda\mathbb{E}\left[U_{i-1}\right] = \lambda\mu_{i-1}.$$

The inequality in (5.5) easily follows by induction.

Next we will show (5.6). First let

$$Q_i := \left(U_i + \tfrac{1}{n}\right)^d - U_i^d = \sum_{k=1}^{d}\binom{d}{k}\frac{U_i^{d-k}}{n^k}, \quad R_i := Q_i - P_i + d\left(U_{i-1} - U_i\right)U_i^{d-1},$$

so we may decompose the terms in (5.6) as follows:

$$\begin{aligned} U_i^d \mathbf{1}_{V\geq i-1} &= \left(U_i + \tfrac{1}{n}\right)^d \mathbf{1}_{V\geq i-1} - Q_i \mathbf{1}_{V\geq i-1}, \\ \left(1 - d\left(U_{i-1} - U_i\right)U_i^{d-1}\right)\mathbf{1}_{V\geq i} &= \left(1 + Q_i - P_i\right)\mathbf{1}_{V\geq i} - R_i \mathbf{1}_{V\geq i}. \end{aligned}$$

Taking a difference of these two decompositions, and then using (5.8) and the fact that

$Q_i \le \frac{2^d}{n}$, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[U_i^d \mathbf{1}_{V\ge i-1}\right] - \mathbb{E}\left[\left(1 - d\left(U_{i-1}-U_i\right)U_i^{d-1}\right)\mathbf{1}_{V\ge i}\right]\right| \\
&\le \mathbb{E}\left[\left|\left(U_i+\tfrac{1}{n}\right)^d \mathbf{1}_{V\ge i-1} - Q_i\mathbf{1}_{V\ge i-1} - \left(1+Q_i-P_i\right)\mathbf{1}_{V\ge i} + R_i\mathbf{1}_{V\ge i}\right|\right] \\
&\le \mathbb{E}\left[\left|\left(U_i+\tfrac{1}{n}\right)^d \mathbf{1}_{V\ge i-1} - \left(1+Q_i-P_i\right)\mathbf{1}_{V\ge i}\right|\right] + \mathbb{E}\left[\left|R_i\mathbf{1}_{V\ge i} - Q_i\mathbf{1}_{V\ge i-1}\right|\right] \\
&\le \frac{\nu_i - \nu_{i+1}}{\lambda n} + \frac{2^d}{n} + \mathbb{E}\left[R_i\right].
\end{aligned}$$

Thus it suffices to show that R_i is also of order $O\left(\frac{1}{n}\right)$. Let us write

$$P_i = n\left(U_{i-1}-U_i\right)Q_i,$$

since $\sum_{s=1}^d x^{s-1}y^{d-s} = \frac{y^d-x^d}{y-x}$ for all distinct $x, y \in \mathbb{R}$. Then

$$\begin{aligned}
R_i &= Q_i - \left[n\left(U_{i-1}-U_i\right)Q_i\right] + d\left(U_{i-1}-U_i\right)U_i^{d-1} \\
&= Q_i - \left(U_{i-1}-U_i\right)\left[nQ_i - dU_i^{d-1}\right] \\
&= Q_i - \left(U_{i-1}-U_i\right)\sum_{k=2}^{d}\binom{d}{k}\frac{U_i^{d-k}}{n^{k-1}} \le \frac{2^d}{n}.
\end{aligned}$$

Hence the result follows by taking $c_1 := \frac{1}{\lambda} + 2^{d+1}$. □

Equations (5.5) and (5.6) contain terms of the form $\mathbb{E}\left[U_{i_1}U_{i_2}^{d-1}\mathbf{1}_{V\ge j}\right]$. Such terms should be strongly concentrated around

$$\mathbb{E}\left[\mu_{i_1}\mu_{i_2}^{d-1}\mathbf{1}_{V\ge j}\right] = \mu_{i_1}\mu_{i_2}^{d-1}\mathbb{E}\left[\mathbf{1}_{V\ge j}\right] = \mu_{i_1}\mu_{i_2}^{d-1}\nu_j,$$

since the U_i are strongly concentrated around their means μ_i. This is expressed in the following lemma.

Lemma 5.3. *There exists $c_2 > 0$ such that the following holds. Let $n \ge 1$, and let X have the equilibrium distribution for the lengths process. Then*

$$\sup_{i_1,\ldots,i_d,j\ge 0}\left|\mathbb{E}\left[U_{i_1}\ldots U_{i_d}\mathbf{1}_{V\ge j}\right] - \mu_{i_1}\ldots\mu_{i_d}\nu_j\right| \le \frac{c_2\ln^2 n}{\sqrt{n}}.$$

Proof. Since the left-hand side is bounded by 1 and $c_2 > 0$ may be arbitrarily large, it suffices to show the result for all sufficiently large n.

Let

$$A := \left\{\sup_{i\ge 1}\left|U_i - \mu_i\right| \le \frac{\ln^2 n}{\sqrt{n}}\right\}.$$

By Lemma 4.5 with $z = 1$, there exists $\eta > 0$ such that

$$\mathbb{E}\left[\mathbf{1}_{\overline{A}}\right] = \mathbb{P}\left(\overline{A}\right) \le 2e^{-\eta\ln^2 n} \le \frac{\ln^2 n}{\sqrt{n}},$$

if n is sufficiently large.

Noting that the $\mu_{i_k} \leq 1$ and $\nu_j \leq 1$, we have the easy upper bound

$$\begin{aligned}\mathbb{E}\left[U_{i_1} \dots U_{i_d} \mathbf{1}_{V \geq j} \mathbf{1}_A\right] &\leq \mathbb{E}\left[\prod_{k=1}^{d} \left(\mu_{i_k} + \frac{\ln^2 n}{\sqrt{n}}\right) \mathbf{1}_{V \geq j}\right] \\ &= \prod_{k=1}^{d} \left(\mu_{i_k} + \frac{\ln^2 n}{\sqrt{n}}\right) \nu_j \\ &\leq \mu_{i_1} \dots \mu_{i_d} \nu_j + \binom{d}{1} \frac{\ln^2 n}{\sqrt{n}} + \left[\binom{d}{2} \frac{\ln^4 n}{n} + \dots + \frac{\ln^{2d} n}{n^{d/2}}\right],\end{aligned}$$

However, the sum of the terms in the square brackets is $O\left(\frac{\ln^2 n}{\sqrt{n}}\right)$, so

$$\mathbb{E}\left[U_{i_1} \dots U_{i_d} \mathbf{1}_{V \geq j} \mathbf{1}_A\right] \leq \mu_{i_1} \dots \mu_{i_d} \nu_j + \frac{(d+1) \ln^2 n}{\sqrt{n}},$$

if n is sufficiently large.

For a lower bound, we will make use of the inequalities

$$U_i \geq \max\left(\mu_i - \frac{\ln^2 n}{\sqrt{n}}, 0\right) \geq 0, \quad \mathbf{1}_{V \geq j} \mathbf{1}_A = \mathbf{1}_{V \geq j} \left(1 - \mathbf{1}_{\overline{A}}\right) \geq \mathbf{1}_{V \geq j} - \mathbf{1}_{\overline{A}}.$$

Thus, by the same reasoning above,

$$\begin{aligned}\mathbb{E}\left[U_{i_1} \dots U_{i_d} \mathbf{1}_{V \geq j} \mathbf{1}_A\right] &\geq \mathbb{E}\left[\prod_{k=1}^{d} \max\left(\mu_{i_k} - \frac{\ln^2 n}{\sqrt{n}}, 0\right) \left(\mathbf{1}_{V \geq j} - \mathbf{1}_{\overline{A}}\right)\right] \\ &= \prod_{k=1}^{d} \max\left(\mu_{i_k} - \frac{\ln^2 n}{\sqrt{n}}, 0\right) \left(\nu_j - \mathbb{P}\left(\overline{A}\right)\right) \\ &\geq \prod_{k=1}^{d} \max\left(\mu_{i_k} - \frac{\ln^2 n}{\sqrt{n}}, 0\right) \nu_j - \mathbb{P}\left(\overline{A}\right) \\ &\geq \prod_{k=1}^{d} \left(\mu_{i_k} - \min\left(\frac{\ln^2 n}{\sqrt{n}}, \mu_{i_k}\right)\right) \nu_j - \frac{\ln^2 n}{\sqrt{n}} \\ &\geq \mu_{i_1} \dots \mu_{i_d} \nu_j - \frac{(d+2) \ln^2 n}{\sqrt{n}},\end{aligned}$$

if n is sufficiently large.

Combining our upper and lower bounds gives

$$\begin{aligned}\left|\mathbb{E}\left[U_{i_1} \dots U_{i_d} \mathbf{1}_{V \geq j}\right] - \mu_{i_1} \dots \mu_{i_d} \nu_j\right| &\leq \left|\mathbb{E}\left[U_{i_1} \dots U_{i_d} \mathbf{1}_{V \geq j}\right] - \mathbb{E}\left[U_{i_1} \dots U_{i_d} \mathbf{1}_{V \geq j} \mathbf{1}_A\right]\right| \\ &\quad + \left|\mathbb{E}\left[U_{i_1} \dots U_{i_d} \mathbf{1}_{V \geq j} \mathbf{1}_A\right] - \mu_{i_1} \dots \mu_{i_d} \nu_j\right| \\ &\leq \mathbb{E}\left[\mathbf{1}_{\overline{A}}\right] + \frac{(d+2) \ln^2 n}{\sqrt{n}} \leq \frac{(d+3) \ln^2 n}{\sqrt{n}},\end{aligned}$$

if n is sufficiently large. The result follows by taking $c_2 := d + 3$. □

Lemma 5.2 and the concentration of measure results of Lemma 5.3 then imply the following uniform bounds.

Lemma 5.4. *There exists $c_3 > 0$ such that the following holds. Let $n \geq 1$, and let X have the equilibrium distribution for the lengths process. Then*

$$\sup_{i \geq 1} \left| \mu_i - \lambda \mu_{i-1}^d \nu_{i-1} \right| \leq \frac{c_3 \ln^2 n}{\sqrt{n}},$$

and

$$\sup_{i \geq 1} \left| \mu_i^d \nu_{i-1} - \left(1 - d\left(\mu_{i-1} - \mu_i\right) \mu_i^{d-1}\right) \nu_i \right| \leq \frac{c_3 \ln^2 n}{\sqrt{n}}.$$

Proof. Let $c_1, c_2 > 0$ denote the constants given by Lemma 5.2 and Lemma 5.3, respectively. Then

$$\left| \mu_i - \lambda \mu_{i-1}^d \nu_{i-1} \right| = \lambda \left| \mathbb{E}\left[U_{i-1}^d \mathbf{1}_{V \geq i-1} \right] - \mu_{i-1}^d \nu_{i-1} \right| \leq \lambda \frac{c_2 \ln^2 n}{\sqrt{n}},$$

and

$$\begin{aligned}
&\left| \mu_i^d \nu_{i-1} - \left(1 - d\left(\mu_{i-1} - \mu_i\right) \mu_i^{d-1}\right) \nu_i \right| \\
&\leq \left| \mu_i^d \nu_{i-1} - \mathbb{E}\left[U_i^d \mathbf{1}_{V \geq i-1} \right] \right| \\
&\quad + \left| \mathbb{E}\left[U_i^d \mathbf{1}_{V \geq i-1} \right] - \mathbb{E}\left[\left(1 - d\left(U_{i-1} - U_i\right) U_i^{d-1}\right) \mathbf{1}_{V \geq i} \right] \right| \\
&\quad + \left| \mathbb{E}\left[\left(1 - d\left(U_{i-1} - U_i\right) U_i^{d-1}\right) \mathbf{1}_{V \geq i} \right] - \left(1 - d\left(\mu_{i-1} - \mu_i\right) \mu_i^{d-1}\right) \nu_i \right| \\
&\leq \frac{c_2 \ln^2 n}{\sqrt{n}} + \frac{c_1}{n} + d\left(\left| \mathbb{E}\left[U_{i-1} U_i^{d-1} \mathbf{1}_{V \geq i} \right] - \mu_{i-1} \mu_i^{d-1} \nu_i \right| + \left| \mathbb{E}\left[U_i^d \mathbf{1}_{V \geq i} \right] - \mu_i^d \nu_i \right| \right) \\
&\leq \frac{c_2 \ln^2 n}{\sqrt{n}} + \frac{c_1}{n} + d\left(\frac{2 c_2 \ln^2 n}{\sqrt{n}} \right) \leq \frac{\left(c_1 + c_2\left(2d + 1\right)\right) \ln^2 n}{\sqrt{n}}.
\end{aligned}$$

Hence the result follows by taking $c_3 := c_1 + c_2\left(2d + 1\right)$. □

These uniform bounds suggest that the μ_i and ν_i satisfy

$$\mu_i \approx \lambda \mu_{i-1}^d \nu_{i-1}, \tag{5.9}$$

$$\mu_i^d \nu_{i-1} \approx \left(1 - d\left(\mu_{i-1} - \mu_i\right) \mu_i^{d-1}\right) \nu_i, \tag{5.10}$$

for all $i \geq 1$. We will analyse the two families of recurrence relations suggested by (5.9) and (5.10) in the next section.

5.3 Solutions to the recurrence relations

In this section, we will analyse the two families of recurrence relations suggested in the previous section. The calculations are in-depth for the sake of completeness, but are routine and easy.

Equations (5.9) and (5.10) suggest that the means μ_i and ν_i should be close to a_i and b_i, as defined in (1.1). We remind the reader that $a_0 = b_0 = 1$, and that

$$a_i := \lambda a_{i-1}^d b_{i-1}, \quad b_i := \frac{a_i^d b_{i-1}}{1 - d\left(a_{i-1} - a_i\right) a_i^{d-1}},$$

for all $i \geq 1$. For brevity, we will let

$$p_i := d\,(a_{i-1} - a_i)\,a_i^{d-1}.$$

It is clear that $(\mu_i)_{i=0}^{\infty}$ and $(\nu_i)_{i=0}^{\infty}$ are decreasing sequences in $(0,1]$. This suggests that $(a_i)_{i=0}^{\infty}$ and $(b_i)_{i=0}^{\infty}$ should also be decreasing sequences in $(0,1]$. However, since $b_0 = b_1 = 1$ if $d = 1$, the claim for the latter sequence should only be for the indices $i \geq 1$.

Lemma 5.5. *We have*

$$a_{i+1} < a_i \leq 1, \quad b_{i+2} < b_{i+1} \leq 1,$$

for all $i \geq 0$.

Proof. First we claim that if $0 < a_i \leq 1$, then

$$b_{i+1} = \frac{a_{i+1}^d b_i}{1 - p_{i+1}} \leq b_i, \tag{5.11}$$

with strict inequality if $0 < a_i < 1$. To see this inequality, use Lemma 1.5 (2) (with $x = a_{i+1}$ and $y = a_i$) to obtain $p_{i+1} \leq a_i^d - a_{i+1}^d \leq 1 - a_{i+1}^d$, with the last inequality being strict if $0 < a_i < 1$. Also note that if $b_i \leq 1$, then

$$a_{i+1} = \lambda a_i^d b_i < a_i. \tag{5.12}$$

Now let us use induction to show that the result holds. For the base step of $i = 0$, the first inequality easily holds since $a_0 = 1$ and $a_1 = \lambda$. For the second inequality, there are two cases to consider. If $d = 1$, then we simply calculate that

$$b_1 = 1, \quad a_2 = \lambda^2, \quad b_2 = \frac{\lambda^2}{1 - (\lambda - \lambda^2)}.$$

Else if $d \geq 2$, then (5.11) (with $i = 0$) first gives $b_1 \leq b_0 = 1$. Then, by (5.12) (with $i = 1$), we have $a_2 < a_1$. Finally, by (5.11) again (with $i = 1$), we have $b_2 < b_1$.

For the inductive step, suppose that

$$a_i < a_{i-1} \leq 1, \quad b_{i+1} < b_i \leq 1,$$

for some $i \geq 1$. By (5.12) and the hypotheses $a_i, b_i \leq 1$, we have $a_{i+1} < a_i \leq 1$. Then, by (5.12) and the hypothesis $b_{i+1} \leq 1$, we have $b_{i+2} < b_{i+1} \leq 1$. □

In particular, $(a_i)_{i=0}^{\infty}$ is a decreasing sequence in $(0,1]$ with $a_1 < 1$.

Lemma 5.6. *Let* $(r_i)_{i=0}^{\infty}$ *be a decreasing sequence in* $(0,1]$ *with* $r_1 < 1$. *Then there exists* $\kappa > 1$ *such that*

$$\prod_{i=1}^{\infty} \frac{1}{1 - d\,(r_{i-1} - r_i)\,r_i^{d-1}} \leq \kappa.$$

Proof. By Lemma 1.5 (2), we have

$$q_i := d\,(r_{i-1} - r_i)\,r_i^{d-1} \leq r_{i-1}^d - r_i^d \leq \begin{cases} r_0^d - r_1^d \leq 1 - r_1^d, & \text{if } i = 1, \\ r_{i-1}^d < r_1^d, & \text{if } i \geq 2. \end{cases}$$

Let $\rho := \max\left(r_1^d, 1 - r_1^d\right)$, and note that $0 < \rho < 1$. Then

$$\prod_{i=1}^{\infty} \frac{1}{1-q_i} = \prod_{i=1}^{\infty} \left(1 + \frac{q_i}{1-q_i}\right) \leq \prod_{i=1}^{\infty} \left(1 + \frac{q_i}{1-\rho}\right) < \infty,$$

with the infinite product converging since $\sum_{i=1}^{\infty} q_i \leq \sum_{i=1}^{\infty} \left(r_{i-1}^d - r_i^d\right) = r_0^d < \infty$. □

Now we will show that the a_i and b_i are asymptotically doubly exponential. We begin with some heuristic calculations: we have

$$a_i = \lambda a_{i-1}^d \prod_{j=1}^{i-1} \frac{a_j^d}{1-p_j} \approx \lambda a_{i-1}^d \prod_{j=1}^{i-1} a_j^d,$$

for all large i, since $p_j \approx 0$ for large j. If we suppose that $a_i \approx \omega^{f_i}$ for some $0 < \omega < 1$, then

$$\omega^{f_i} \approx \lambda \omega^{df_{i-1}} \prod_{j=1}^{i-1} \omega^{df_j} = \omega^{\ln\lambda/\ln\omega} \omega^{df_{i-1}} \prod_{j=1}^{i-1} \omega^{df_j}.$$

Treating this as an equality, we have

$$f_i = \frac{\ln\lambda}{\ln\omega} + d\left(\sum_{j=1}^{i-2} f_j + 2f_{i-1}\right),$$

which satisfies the recurrence relation

$$\begin{aligned} f_{i+2} - (2d+1)\, f_{i+1} + df_i &= \frac{\ln\lambda}{\ln\omega} + d\left(\sum_{j=1}^{i-1} f_j + f_i + 2f_{i+1}\right) - (2d+1)\, f_{i+1} + df_i \\ &= \frac{\ln\lambda}{\ln\omega} + d\left(\sum_{j=1}^{i-1} f_j + 2f_i\right) - f_{i+1} = 0. \end{aligned}$$

This has solutions

$$f_i = c_1\alpha^i + c_2\bar{\alpha}^i,$$

where $c_1, c_2 > 0$, where α is as defined in (1.3) and $\bar{\alpha} := d + \frac{1}{2} - \sqrt{d^2 + \frac{1}{4}}$. Since $\alpha > \bar{\alpha}$, this suggests that the asymptotic behaviour of the solution is $f_i \approx c_1\alpha^i$, as $i \to \infty$. We will need the following result by Luczak and Norris [13].

Lemma 5.7 ([13], Proposition 2.5)**.** *There exists $c > 1$ such that*

$$\frac{a_i^\alpha}{c} \leq a_{i+1} \leq c a_i^\alpha$$

for all $i \geq 0$.

Now we will show that $(a_i)_{i=0}^{\infty}$ and $(b_i)_{i=0}^{\infty}$ are asymptotically doubly exponential.

Lemma 5.8. *There exists $0 < \sigma < \tau < 1$ such that*

$$\sigma^{\alpha^i} \leq a_i \leq \tau^{\alpha^i} \tag{5.13}$$

for all $i \geq 1$, *and*

$$\sigma^{d\alpha^{i+1}/(\alpha-1)} \leq b_i \leq \tau^{d\alpha^{i+1}/(\alpha-1)} \tag{5.14}$$

for all $i \geq 2$.

Proof. Let $c > 1$ denote the constant given by Lemma 5.7. Then, by induction, we have

$$\frac{a_{i-j}^{\alpha^j}}{c^{1+\alpha+\alpha^2+\cdots+\alpha^{j-1}}} \leq a_i \leq c^{1+\alpha+\alpha^2+\cdots+\alpha^{j-1}} a_{i-j}^{\alpha^j}$$

for all $i \geq 1$ and $0 \leq j \leq i$. Since $\alpha > 2d \geq 2$, we have

$$1 + \alpha + \alpha^2 + \cdots + \alpha^{j-1} = \frac{\alpha^j - 1}{\alpha - 1} < \alpha^j$$

for all $j \geq 1$, and thus

$$\left(\frac{a_{i-j}}{c}\right)^{\alpha^j} \leq a_i \leq (ca_{i-j})^{\alpha^j} \tag{5.15}$$

for all $i \geq 1$ and $1 \leq j \leq i$.

By Lemma 5.5, we have $0 < a_i, b_j < 1$ for all $i \geq 1$ and $j \geq 2$. Hence, there exist $0 < \rho_i, \beta_j < 1$ such that

$$\rho_i^{\alpha^i} = a_i, \quad \beta_j^{d\alpha^{j+1}/(\alpha-1)} = b_j,$$

for all $i \geq 1$ and $j \geq 2$. The same lemma also implies that there exists $m \geq 1$ such that

$$\omega := ca_m < \tfrac{1}{2}, \quad p_m \leq \tfrac{1}{2}.$$

Let

$$\begin{aligned}
\sigma &:= \min\left(\tfrac{1}{c}, \rho_1, \rho_2, \ldots, \rho_m, \beta_2, \ldots, \beta_m\right), \\
\tau &:= \max\left((2\omega)^{1/\alpha^m}, \rho_1, \rho_2, \ldots, \rho_m, \beta_2, \ldots, \beta_m\right).
\end{aligned}$$

Note that $0 < \sigma < \tau < 1$, where we have used the fact that $2\omega < 1$. Also note that

$$\omega^{\alpha^{i-m}} < (2\omega)^{\alpha^{i-m}} \leq \tau^{\alpha^i}. \tag{5.16}$$

We will directly show that (5.13) holds for all $i \geq 1$. For the lower bound, use (5.15) (with $j = i \geq 1$) and the fact that $\sigma \leq \frac{1}{c}$ to obtain

$$a_i \geq \left(\tfrac{1}{c}\right)^{\alpha^i} \geq \sigma^{\alpha^i}.$$

For the upper bound, if $1 \leq i \leq m$, then we simply have $a_i = \rho_i^{\alpha^i} \leq \tau^{\alpha^i}$. Else if $i > m$, then use (5.15) (with $j = i - m \geq 1$) and (5.16) to obtain

$$a_i \leq (ca_m)^{\alpha^{i-m}} = \omega^{\alpha^{i-m}} < \tau^{\alpha^i}. \tag{5.17}$$

Thus (5.13) holds for all $i \geq 1$.

Now let us use induction to show that (5.14) holds for all $i \geq 2$. For the base steps of

$2 \leq i \leq m$, we have $\sigma \leq \beta_i \leq \tau$ and thus

$$\sigma^{d\alpha^{i+1}/(\alpha-1)} \leq b_i = \beta_i^{d\alpha^{i+1}/(\alpha-1)} \leq \tau^{d\alpha^{i+1}/(\alpha-1)}.$$

For the inductive step, suppose that

$$\sigma^{d\alpha^{i}/(\alpha-1)} \leq b_{i-1} \leq \tau^{d\alpha^{i}/(\alpha-1)}$$

for some $i > m$. Then

$$a_i^d b_{i-1} \leq b_i = \frac{a_i^d b_{i-1}}{1-p_i} \leq \frac{a_i^d b_{i-1}}{1-p_{m+1}} \leq 2a_i^d b_{i-1}.$$

Using the lower bound, (5.13) and the inductive hypothesis, we have

$$b_i \geq \sigma^{d\alpha^i}\sigma^{d\alpha^i/(\alpha-1)} = \sigma^{d\alpha^{i+1}/(\alpha-1)}.$$

Using the upper bound, (5.16), (5.17) and the inductive hypothesis, we have

$$b_i \leq 2\omega^{d\alpha^{i-m}}\tau^{d\alpha^i/(\alpha-1)} \leq (2\omega)^{d\alpha^{i-m}}\tau^{d\alpha^i/(\alpha-1)} \leq \tau^{d\alpha^i}\tau^{d\alpha^i/(\alpha-1)} = \tau^{d\alpha^{i+1}/(\alpha-1)}.$$

This completes the inductive step. □

5.4 Long-term behaviour

In this section, we will show that the μ_i and ν_i are uniformly close to a_i and b_i, respectively, for long periods of time.

Lemma 5.9. *There exists $c > 0$ such that the following holds. Let $n \geq 1$, and let X have the equilibrium distribution for the lengths process. Then*

$$\sup_{i\geq 1}|\mu_i - a_i| \leq \frac{c\ln^2 n}{\sqrt{n}}, \quad \sup_{i\geq 1}|\nu_i - b_i| \leq \frac{c\ln^2 n}{\sqrt{n}}.$$

Remark. This proof closely follows the argument in Chapter 5 of [10], where the analogous result for the standard supermarket model is to be found. The main difference is that here we are seeking a pair of bounds instead of just one, so each part of the original argument is adapted into a pair of arguments here.

Proof. In this proof, we will do two pairs of inductions to bound $|\mu_i - a_i|$ and $|\nu_i - b_i|$. The first will establish bounds which depend on i, and then the second will establish bounds which are independent of i.

Let $0 < \tau < 1$ denote the constant given by Lemma 5.8, then let $\omega := \max(\lambda, \tau)$. By Lemma 5.2 and Lemma 5.8, we have

$$\mu_i \leq \lambda^i \leq \omega^i, \quad b_j \leq \tau^{d\alpha^{j+1}/(\alpha-1)} \leq \omega^j, \tag{5.18}$$

for all $i \geq 1$ and $j \geq 2$.

Recall that for $i \geq 1$, we defined $p_i := d\,(a_{i-1} - a_i)\,a_i^{d-1}$; also let $q_i := d\,(\mu_{i-1} - \mu_i)\,\mu_i^{d-1}$. By Lemma 5.2 and Lemma 5.5, $(\mu_i)_{i=0}^{\infty}$ and $(a_i)_{i=0}^{\infty}$ are both decreasing sequences in $(0, 1]$, with $\mu_1 = a_1 < 1$. Hence, by Lemma 5.6, there exists $\kappa > 1$ such that

$$\frac{1}{1-p_i} \leq \kappa, \quad \frac{1}{1-q_i} \leq \kappa,$$

for all $i \geq 1$. Let $c_3 > 0$ denote the constant given by Lemma 5.4, then let $h := c_3\kappa$ and $m := \kappa^2 d^2$.

We will also need the inequalities

$$\left|\mu_i^k - a_i^k\right| \leq d\,|\mu_i - a_i|\,, \tag{5.19}$$

$$|rs - tu| \leq s\,|r - t| + t\,|s - u|\,, \tag{5.20}$$

for all $1 \leq k \leq d$, $i \geq 1$ and $r, s, t, u \in \mathbb{R}$ with $s, t \geq 0$.

Before we set up the induction, we will need two preliminary results. Our first preliminary result is quick to derive. For $i \geq 1$, we may write

$$|\mu_i - a_i| = \left|\mu_i - \lambda a_{i-1}^d b_{i-1}\right| \leq \left|\mu_i - \lambda \mu_{i-1}^d \nu_{i-1}\right| + \lambda\left|\mu_{i-1}^d \nu_{i-1} - a_{i-1}^d b_{i-1}\right|.$$

Using Lemma 5.4 and (5.20) (with $r = a_{i-1}^d$, $s = b_{i-1}$, $t = \mu_{i-1}^d$ and $u = \nu_{i-1}$), we have

$$|\mu_i - a_i| \leq \frac{c_3 \ln^2 n}{\sqrt{n}} + b_{i-1}\left|\mu_{i-1}^d - a_{i-1}^d\right| + \mu_{i-1}^d\,|\nu_{i-1} - b_{i-1}|\,.$$

Using (5.19) on the second term and the fact that $\mu_{i-1} \leq 1$, we have

$$\begin{aligned} |\mu_i - a_i| &\leq \frac{h \ln^2 n}{\sqrt{n}} + d b_{i-1}\,|\mu_{i-1} - a_{i-1}| + \mu_{i-1}\,|\nu_{i-1} - b_{i-1}| \\ &\leq \frac{h \ln^2 n}{\sqrt{n}} + m\left(b_{i-1}\,|\mu_{i-1} - a_{i-1}| + \mu_{i-1}\,|\nu_{i-1} - b_{i-1}|\right), \end{aligned} \tag{5.21}$$

for all $i \geq 1$.

Our second preliminary result takes longer to derive. For $i \geq 1$, we may write

$$\begin{aligned} |\nu_i - b_i| &= \left|\nu_i - \frac{a_i^d b_{i-1}}{1-p_i}\right| \\ &\leq \left|\frac{(1-q_i)\,\nu_i - \mu_i^d \nu_{i-1}}{1-q_i}\right| + \left|\frac{\mu_i^d \nu_{i-1} - a_i^d b_{i-1}}{1-q_i}\right| + \left|\frac{a_i^d b_{i-1}}{1-q_i} - \frac{a_i^d b_{i-1}}{1-p_i}\right| \\ &\leq \kappa\left|\mu_i^d \nu_{i-1} - (1-q_i)\,\nu_i\right| + \kappa\left|\mu_i^d \nu_{i-1} - a_i^d b_{i-1}\right| + \kappa^2 a_i^d b_{i-1}\,|q_i - p_i|\,. \end{aligned}$$

Let us bound the first term using Lemma 5.4, and note that

$$\begin{aligned} |q_i - p_i| &= d\left|(\mu_{i-1} - \mu_i)\,\mu_i^{d-1} - (a_{i-1} - a_i)\,a_i^{d-1}\right| \\ &\leq d\left(\left|\mu_i^{d-1}\mu_{i-1} - a_i^{d-1}a_{i-1}\right| + \left|\mu_i^d - a_i^d\right|\right). \end{aligned}$$

Hence

$$|\nu_i - b_i| \le \kappa \frac{c_3 \ln^2 n}{\sqrt{n}} + \kappa \left|\mu_i^d \nu_{i-1} - a_i^d b_{i-1}\right| + \kappa^2 d a_i^d b_{i-1} \left(\left|\mu_i^{d-1}\mu_{i-1} - a_i^{d-1} a_{i-1}\right| + \left|\mu_i^d - a_i^d\right|\right).$$

Using (5.20) on the second term (with $r = a_i^d$, $s = b_{i-1}$, $t = \mu_i^d$ and $u = \nu_{i-1}$) and on the first term in the brackets (with $r = \mu_i^{d-1}$, $s = \mu_{i-1}$, $t = a_i^{d-1}$ and $u = a_{i-1}$), we have

$$|\nu_i - b_i| \le \frac{h \ln^2 n}{\sqrt{n}} + \kappa \left(b_{i-1}\left|\mu_i^d - a_i^d\right| + \mu_i^d \left|\nu_{i-1} - b_{i-1}\right|\right) + \kappa^2 d a_i^d b_{i-1} \left(\mu_{i-1}\left|\mu_i^{d-1} - a_i^{d-1}\right| + a_i^{d-1}\left|\mu_{i-1} - a_{i-1}\right| + \left|\mu_i^d - a_i^d\right|\right).$$

Using (5.19) on the three terms of the form $\left|\mu_i^k - a_i^k\right|$, we have

$$|\nu_i - b_i| \le \frac{h \ln^2 n}{\sqrt{n}} + \kappa \left(d b_{i-1}\left|\mu_i - a_i\right| + \mu_i^d \left|\nu_{i-1} - b_{i-1}\right|\right) + \kappa^2 d a_i^d b_{i-1} \left(d\mu_{i-1}\left|\mu_i - a_i\right| + a_i^{d-1}\left|\mu_{i-1} - a_{i-1}\right| + d\left|\mu_i - a_i\right|\right).$$

Finally, we use the fact that $\mu_i \le \mu_{i-1} \le 1$ and that $a_i \le 1$, so

$$|\nu_i - b_i| \le \frac{h \ln^2 n}{\sqrt{n}} + m\left(b_{i-1}\left|\mu_{i-1} - a_{i-1}\right| + \mu_{i-1}\left|\nu_{i-1} - b_{i-1}\right| + 3 b_{i-1}\left|\mu_i - a_i\right|\right), \quad (5.22)$$

for all $i \ge 1$.

The first pair of inductions will show that

$$|\mu_i - a_i| \le h \sum_{r=0}^{2i-3} (5m)^r \frac{\ln^2 n}{\sqrt{n}}, \quad (5.23)$$

$$|\nu_i - b_i| \le h \sum_{r=0}^{2i-2} (5m)^r \frac{\ln^2 n}{\sqrt{n}}, \quad (5.24)$$

for all $i \ge 1$.

The base steps will be $1 \le i \le 2$. First, Lemma 5.2 gives $\mu_1 = \lambda$, and thus $|\mu_1 - a_1| = 0$. By (5.22) (with $i = 1$), we have

$$|\nu_1 - b_1| \le \frac{h \ln^2 n}{\sqrt{n}} + m\left(|\mu_0 - a_0| + |\nu_0 - b_0| + 3|\mu_1 - a_1|\right) = \frac{h \ln^2 n}{\sqrt{n}}.$$

By (5.21) (with $i = 2$), we have

$$|\mu_2 - a_2| \le \frac{h \ln^2 n}{\sqrt{n}} + m\left(|\mu_1 - a_1| + |\nu_1 - b_1|\right) \le h(1+m)\frac{\ln^2 n}{\sqrt{n}}.$$

Finally, by (5.22) (with $i = 2$), we have

$$|\nu_2 - b_2| \le \frac{h \ln^2 n}{\sqrt{n}} + m\left(|\mu_1 - a_1| + |\nu_1 - b_1| + 3|\mu_2 - a_2|\right) \le h\left(1 + 4m + 3m^2\right)\frac{\ln^2 n}{\sqrt{n}}.$$

Thus, (5.23) and (5.24) hold for $1 \le i \le 2$.

For the inductive step, suppose that

$$|\mu_{i-1} - a_{i-1}| \leq h \sum_{r=0}^{2i-5} (5m)^r \frac{\ln^2 n}{\sqrt{n}}, \quad |\nu_{i-1} - b_{i-1}| \leq h \sum_{r=0}^{2i-4} (5m)^r \frac{\ln^2 n}{\sqrt{n}},$$

for some $i \geq 3$. Since $i - 1 \geq 2$, we may use (5.18) to bound $\mu_{i-1} \leq \omega^{i-1}$ and $b_{i-1} \leq \omega^{i-1}$. Hence, the preliminary results (5.21) and (5.22) give

$$|\mu_i - a_i| \leq \frac{h \ln^2 n}{\sqrt{n}} + m\omega^{i-1} \left(|\mu_{i-1} - a_{i-1}| + |\nu_{i-1} - b_{i-1}| \right), \tag{5.25}$$

$$|\nu_i - b_i| \leq \frac{h \ln^2 n}{\sqrt{n}} + m\omega^{i-1} \left(|\mu_{i-1} - a_{i-1}| + |\nu_{i-1} - b_{i-1}| + 3 |\mu_i - a_i| \right), \tag{5.26}$$

for all $i \geq 3$.

Substituting the inductive hypotheses into (5.25) gives

$$\begin{aligned} |\mu_i - a_i| &\leq \frac{h \ln^2 n}{\sqrt{n}} + m \left(h \sum_{r=0}^{2i-5} (5m)^r \frac{\ln^2 n}{\sqrt{n}} + h \sum_{r=0}^{2i-4} (5m)^r \frac{\ln^2 n}{\sqrt{n}} \right) \\ &= h \left(1 + 2m \sum_{r=0}^{2i-5} (5m)^r + m (5m)^{2i-4} \right) \frac{\ln^2 n}{\sqrt{n}} \\ &\leq h \left(1 + 5m \sum_{r=0}^{2i-5} (5m)^r + 5m (5m)^{2i-4} \right) \frac{\ln^2 n}{\sqrt{n}} = h \sum_{r=0}^{2i-3} (5m)^r \frac{\ln^2 n}{\sqrt{n}}. \end{aligned}$$

Substituting the inductive hypotheses and this result into (5.26) then gives

$$\begin{aligned} |\nu_i - b_i| &\leq \frac{h \ln^2 n}{\sqrt{n}} + m \left(h \sum_{r=0}^{2i-5} (5m)^r \frac{\ln^2 n}{\sqrt{n}} + h \sum_{r=0}^{2i-4} (5m)^r \frac{\ln^2 n}{\sqrt{n}} + 3h \sum_{r=0}^{2i-3} (5m)^r \frac{\ln^2 n}{\sqrt{n}} \right) \\ &= h \left(1 + 5m \sum_{r=0}^{2i-5} (5m)^r + 4m (5m)^{2i-4} + 3m (5m)^{2i-3} \right) \frac{\ln^2 n}{\sqrt{n}} \\ &\leq h \left(1 + 5m \sum_{r=0}^{2i-5} (5m)^r + 5m (5m)^{2i-4} + 5m (5m)^{2i-3} \right) \frac{\ln^2 n}{\sqrt{n}} = h \sum_{r=0}^{2i-2} (5m)^r \frac{\ln^2 n}{\sqrt{n}}. \end{aligned}$$

Thus, (5.23) and (5.24) hold for all $i \geq 1$.

Now let $j \geq 1$ be sufficiently large so that $5m\omega^{j-1} \leq \frac{4}{5}$, then let

$$c := h \sum_{r=0}^{2j-2} (5m)^r .$$

The second pair of inductions will show that

$$|\mu_i - a_i|, |\nu_i - b_i| \leq \frac{c \ln^2 n}{\sqrt{n}}, \tag{5.27}$$

for all $i \geq 1$.

The base steps will be $0 \leq i \leq j$, and these trivially hold, since (5.23) and (5.24) give

$$|\mu_i - a_i| \leq h \sum_{r=0}^{2i-3} (5m)^r \frac{\ln^2 n}{\sqrt{n}} \leq \frac{c \ln^2 n}{\sqrt{n}}, \quad |\nu_i - b_i| \leq h \sum_{r=0}^{2i-2} (5m)^r \frac{\ln^2 n}{\sqrt{n}} \leq \frac{c \ln^2 n}{\sqrt{n}},$$

for all $i \leq j$.

For the inductive step, suppose that

$$|\mu_{i-1} - a_{i-1}|, |\nu_{i-1} - b_{i-1}| \leq \frac{c \ln^2 n}{\sqrt{n}},$$

for some $i > j$. Then

$$h + 5cm\omega^{i-1} \leq \tfrac{1}{5}c + 5cm\omega^{j-1} \leq \tfrac{1}{5}c + \tfrac{4}{5}c = c,$$

since $c \geq 5h$.

Substituting the inductive hypotheses into (5.25) gives

$$\begin{aligned} |\mu_i - a_i| &\leq \frac{h \ln^2 n}{\sqrt{n}} + m\omega^{i-1} \left(\frac{c \ln^2 n}{\sqrt{n}} + \frac{c \ln^2 n}{\sqrt{n}} \right) \\ &= \left(h + 2cm\omega^{i-1}\right) \frac{\ln^2 n}{\sqrt{n}} \leq \frac{c \ln^2 n}{\sqrt{n}}. \end{aligned}$$

Substituting the inductive hypotheses and this result into (5.26) then gives

$$\begin{aligned} |\nu_i - b_i| &\leq \frac{h \ln^2 n}{\sqrt{n}} + m\omega^{i-1} \left(\frac{c \ln^2 n}{\sqrt{n}} + \frac{c \ln^2 n}{\sqrt{n}} + 3\frac{c \ln^2 n}{\sqrt{n}} \right) \\ &= \left(h + 5cm\omega^{i-1}\right) \frac{\ln^2 n}{\sqrt{n}} \leq \frac{c \ln^2 n}{\sqrt{n}}. \end{aligned}$$

Thus, (5.27) holds for all $i \geq 1$. □

Finally, we will uniformly bound the equilibrium deviation of the tail functions from the a_i, over long periods of time.

Lemma 5.10. *Let $c > 0$ denote the constant given by Lemma 5.9. For all $z > c$ and $r > 0$, there exists $\eta = \eta(z, r) > 0$ such that the following holds. Let $n \geq 1$, and let $\mathbf{X}$ be in equilibrium. Then*

$$\mathbb{P}\left(\sup_{i \geq 1} |u_i(X_t) - a_i| \geq \frac{z \ln^2 n}{\sqrt{n}} \text{ for some } 0 \leq t \leq n^r \right) \leq 2e^{-\eta \ln^2 n}.$$

Proof. Since the left-hand side is bounded by 1 and $\eta > 0$ may be arbitrarily small, it suffices to show the result for all sufficiently large n.

For $t \geq 0$ and $h > 0$, let

$$E_{t,h} := \left\{ \sup_{i \geq 1} |u_i(X_t) - \mu_i| \geq \frac{h \ln^2 n}{\sqrt{n}} \right\}.$$

Let $y = y(z) := z - c > 0$. Then $\overline{E_{t,y/2}}$ holds with high probability at each individual

time, since Lemma 4.5 (with $z := \frac{1}{2}y$) gives $\eta_1 = \eta_1(z) > 0$ such that

$$\mathbb{P}\left(E_{t,y/2}\right) \leq 2e^{-\eta_1 \ln^2 n} \tag{5.28}$$

for all $t \geq 0$.

Now we will extend this to the interval $[0, n^r]$. Consider covering this with sub-intervals of length

$$\delta = \delta(z) := \frac{y \ln^2 n}{4(\lambda+1)\sqrt{n}}.$$

Clearly $m = m(z, r) := \left\lceil \frac{n^r}{\delta} \right\rceil$ such sub-intervals will cover $[0, n^r]$. For $k \geq 0$, let $t_k := k\delta$, then

$$\mathbb{P}\left(\bigcup_{0 \leq t \leq n^r} E_{t,y}\right) \leq \sum_{k=0}^{m} \mathbb{P}\left(E_{t_k,y/2}\right) + m\mathbb{P}\left(\mathrm{Po}\left(\tfrac{1}{4}y\sqrt{n}\ln^2 n\right) \geq \tfrac{1}{2}y\sqrt{n}\ln^2 n\right).$$

To see the last term in this inequality, suppose that $\overline{E_{t_k,y/2}}$ holds for all end-points t_k. Then there exists a sub-interval $\mathcal{I}_l := (t_{l-1}, t_l)$ containing t. Since $\overline{E_{t_{k-1},y/2}}$ and $E_{t,y}$ hold, we deduce that over $\mathcal{I}_l$ the proportion of queues of length at least i changes by at least $\frac{y\ln^2 n}{2\sqrt{n}}$, for some $i \geq 1$, and thus over $\mathcal{I}_l$, we have at least $\frac{1}{2}y\sqrt{n}\ln^2 n$ events. However, the number of events over $\mathcal{I}_l$, an interval of length δ, is Poisson with mean $(\lambda+1)\delta n = \frac{1}{4}y\sqrt{n}\ln^2 n$. By (5.28) and Lemma 1.4 (with $\varepsilon = 1$), we have

$$\mathbb{P}\left(\bigcup_{0 \leq t \leq n^r} E_{t,y}\right) \leq (m+1)\left(2e^{-\eta_1 \ln^2 n}\right) + m\left(2e^{-\frac{1}{12}y\sqrt{n}\ln^2 n}\right).$$

Let $\eta = \eta(z) := \frac{1}{2}\min\left(\eta_1, \frac{1}{12}y\right)$, then straightforward manipulation gives

$$\mathbb{P}\left(\bigcup_{0 \leq t \leq n^r} E_{t,y}\right) \leq 3\left(\frac{n^r}{\delta} + 2\right)e^{-2\eta\ln^2 n} \leq \frac{3 \cdot 3n^r}{\min(\delta, 1)}e^{-2\eta\ln^2 n} \leq 2e^{-\eta\ln^2 n},$$

if n is sufficiently large. Note that how large n must be for the last inequality to hold will also depend on r. □

Chapter 6

Rapid mixing — part two

In this chapter, we will complete our proof of rapid mixing of the lengths process. This will require defining *swap-adjacency* for a pair of profile-equivalent lengths vectors, and then the *swap coupling* of two lengths processes with swap-adjacent initial states. Like Chapter 3, this chapter is based on Chapter 2 of [10] by Luczak and McDiarmid.

6.1 Swap-adjacency and distance

In this section, we will define *swap-adjacency* and then the *swap-distance* between a pair of profile-equivalent lengths vectors. We begin with the concepts of being *lengths-swapped* and *memory-aligned*.

Definition 6.1. We will say that profile-equivalent $x, y \in \mathcal{Q}_n$ are *lengths-swapped* at k and l (the *swapped queues*), where $k \neq l$, if

1. queue i in x and queue i in y have the same length for all $i \neq k, l$, and
2. queue k (resp., l) in x and queue l (resp., k) in y have the same length.

If x and y are lengths-swapped at k and l, then we will say they are *memory-aligned* if the memory queues in x and in y are

1. the same non-swapped queue (that is, queue $i \neq k, l$ in both lengths vectors), or
2. different swapped queues (that is, queue k in one lengths vector and queue l in the other).

We will say that a pair of queues are *indistinguishable* if the two queues have the same length and neither is the memory queue, and *distinguishable* otherwise. Informally, we will say that two lengths vectors are *swap-adjacent* if we take a pair of identical lengths vectors, and then swap a pair of distinguishable queues.

Definition 6.2. We will say that profile-equivalent $x, y \in \mathcal{Q}_n$ are *swap-adjacent* at k and l, and write $x \frown y$, if

1. x and y are lengths-swapped at k and l,
2. x and y are memory-aligned, and

3. k and l are distinguishable.

For $x, y \in \mathcal{Q}_n$ such that $x \equiv y$, define a *swap-path* of *length* m between x and y to be a sequence

$$x = z_0 \frown z_1 \frown \ldots \frown z_m = y.$$

Note that, in any such path, we have $z_i \equiv x \equiv y$ for all $0 \leq i \leq m$. The following lemma says that swap-adjacency induces a connected structure on each class of profile-equivalent states, that is, each equivalence class in the quotient space $\mathcal{Q}_n/\equiv$.

Lemma 6.3. *Let $x, y \in \mathcal{Q}_n$ satisfy $x \equiv y$. Then there exists a swap-path $x = z_0 \frown z_1 \frown \ldots \frown z_m = y$ of length at most* $\min(2\|x\|_1 + 1, n - 1)$.

Proof. Since x and y each have at most $\|x\|_1 + 1 = \|y\|_1 + 1$ non-empty, non-memory queues, they differ by a permutation on at most $\min(2(\|x\|_1 + 1), n)$ indices. Since any permutation on k indices is a product of at most $k - 1$ transpositions, by successively transposing pairs of queues in x, we obtain a swap-path of length at most $\min(2(\|x\|_1 + 1), n) - 1$ from x to y. □

For $x, y \in \mathcal{Q}_n$ such that $x \equiv y$, let the *swap-distance* $d_s(x, y)$ denote the length of the shortest swap-path between x and y. Else, set $d_s(x, y) = \infty$. Then Lemma 6.3 gives

$$d_s(x, y) \leq \min(2\|x\|_1 + 1, n - 1) \tag{6.1}$$

for all $x, y \in \mathcal{Q}_n$ such that $x \equiv y$. Note that $d_s(x, y) = 0$ if and only if $x = y$, and that $d_s(x, y) = 1$ if and only if $x \frown y$.

6.2 The swap coupling

In this section, we will define the *swap coupling* of two lengths processes with swap-adjacent initial states. We will then show that under this coupling, at each event time, the two processes either remain swap-adjacent or coalesce.

Definition 6.4. The *swap coupling* is the following coupling of lengths processes $\mathbf{X}$ and $\mathbf{Y}$ with swap-adjacent initial states. Let $\mathbf{X}$ and $\mathbf{Y}$ share the same arrival and potential departure times. For an event time T such that $X_{T-} \frown Y_{T-}$ at k and l, pair the queues in X_{T-} and Y_{T-} as follows: pair the opposite swapped queues together (that is, pair queue k in X_{T-} to queue l in Y_{T-}, and vice versa), and then pair the remaining non-swapped queues by index.

1. If T is an arrival time, let the $\mathbf{X}$-choices $C = (C(1), \ldots, C(d))$ be an ordered list of d queues chosen uniformly at random with replacement, then define the $\mathbf{Y}$-choices $C' = (C'(1), \ldots, C'(d))$ by setting $C'(i)$ to be the queue paired with $C(i)$, for all $1 \leq i \leq d$.

2. If T is a potential departure time, let the $\mathbf{X}$-selection be a queue in X_{T-} selected uniformly at random, then set the $\mathbf{Y}$-selection to be the queue in Y_{T-} paired with the $\mathbf{X}$-selection.

Remark. It is easy to see that for an arrival time, the **Y**-choices is an ordered list of d queues chosen uniformly at random with replacement, and that for a potential departure time, the **Y**-selection is a queue in Y_{T-} selected uniformly at random. Thus, **Y** does have the distribution of a lengths process.

Now we will show that under this coupling, at each event time, the two processes either remain swap-adjacent or coalesce.

Lemma 6.5. *Let* **X** *and* **Y** *have initial states* $x, y \in \mathcal{Q}_n$*, respectively, where* $x \frown y$ *at* k *and* l*, and let* **X** *and* **Y** *be coupled by the swap coupling. Let* $T > 0$ *denote the first event time. Then either* $X_T \frown Y_T$ *at* k *and* l*, or* $X_T = Y_T$*.*

Proof. First note that it suffices to show that X_T and Y_T are lengths-swapped at k and l and are memory-aligned. To see this, note that in this case, it follows that $X_T \frown Y_T$ (resp., $X_T = Y_T$) if and only if k and l are distinguishable (resp., indistinguishable). There are now two cases to consider.

Case 1 T is an arrival time.

Since $x \frown y$, for all $1 \leq i \leq d+1$, the i^{th} queues in the **X**- and in the **Y**-candidates lists are either the same non-swapped queue or different swapped queues. Thus, the first shortest queues in the **X**- and **Y**-candidates lists (i.e., the **X**- and **Y**-choices with the corresponding memory queues appended) occur in the same position. This implies that the **X**- and **Y**-customers either join the same non-swapped queue or different swapped queues. In either case, X_T and Y_T are lengths swapped at k and l.

With the queue lengths updated, the first shortest queues in the **X**- and **Y**-candidates lists still occurs in the same position. This implies that the memory queues in X_T and in Y_T are either the same non-swapped queue or different swapped queues. In either case, X_T are Y_T are memory-aligned.

Case 2 T is a potential departure time.

Note that the **X**- and **Y**-selections are either the same non-swapped queue or different swapped queues. □

We remark that coalescence can occur in many different ways. For example, if the swapped queues have the same length m_1 and one is the memory queue, then an arrival where the customer joins a non-swapped queue and a non-swapped queue becomes the memory queue will give coalescence. Alternatively, if the swapped queues have lengths m_2 and $m_2 + 1$, respectively, and neither is the memory queue, then a departure from the longer swapped queue will also give coalescence. As we shall see in the next section, we will only be interested in the special cases where $m_1 = 0$ and $m_2 = 0$. That is, we will only be interested in the case when both swapped queues are empty and neither swapped queue is the memory queue.

We now extend the swap coupling of lengths processes with swap-adjacent initial states to lengths processes with arbitrary profile-equivalent initial states.

Definition 6.6. Let **X** and **Y** have initial states $x, y \in \mathcal{Q}_n$, respectively, where $x \equiv y$. Let $x = z_0 \frown z_1 \frown \ldots \frown z_m = y$ be a shortest swap-path of length $m = d_s(x, y)$ between

x and y. For all $0 \leq i \leq m$, let $\mathbf{Z}^i$ be a lengths process with initial state z_i, and let $\mathbf{Z}^{j-1}$ and $\mathbf{Z}^j$ be coupled by the swap coupling, for all $1 \leq j \leq m$ (using the fact that $z_i \equiv x \equiv y$ for all $0 \leq i \leq m$). This determines a coupling of $\mathbf{X}$ and $\mathbf{Y}$, which we will also call a *swap coupling*.

We then have the following result.

Lemma 6.7. *Let $\mathbf{X}$ and $\mathbf{Y}$ have initial states $x, y \in \mathcal{Q}_n$, respectively, where $x \equiv y$, and let $\mathbf{X}$ and $\mathbf{Y}$ be coupled by a swap coupling. Then $d_s(X_t, Y_t)$ is non-increasing over time.*

Proof. Let m and the $\mathbf{Z}^i$ be as in Definition 6.6. Then

$$d_s(X_t, Y_t) \leq \sum_{i=1}^{m} d_s\left(Z_t^{i-1}, Z_t^i\right).$$

Each summand takes the value 1 before the first event time, and by Lemma 6.5, a value in $\{0, 1\}$ at the first event time. Hence $d_s(X_t, Y_t)$ is non-increasing across the first event, and by induction, is non-increasing over all time. □

6.3 Rapid coalescence

In this section, we will show that in a swap coupling, under reasonable initial conditions, the two lengths processes in fact rapidly coalesce.

As in Section 3.3, we will begin by outlining our strategy for this section. Our strategy is to examine the maximum length of a swapped queue in the two lengths processes in a swap coupling, so we make the following definition.

Definition 6.8. Let $\mathbf{X}$ and $\mathbf{Y}$ have initial states $x, y \in \mathcal{Q}_n$, respectively, where $x \frown y$ at k and l. The *swap walk* is the random walk $\mathbf{W} = (W_t)_{t \geq 0}$ on $\mathbb{Z}_+$ defined by setting

$$W_t := \max\left(X_t(k), X_t(l)\right).$$

The *coalescence time* is

$$T_{\text{co}} := \inf\left\{t \geq 0 : X_t = Y_t\right\},$$

and let

$$T^* := \inf\left\{t \geq 0 : X_t(k) = X_t(l) = 0 \text{ and } \Xi_t \neq k, l\right\}.$$

Since $\mathbf{X}$ and $\mathbf{Y}$ have coalesced by time t if $X_t(k) = X_t(l) = 0$ and $\Xi_t \neq k, l$, we have

$$T_{\text{co}} \leq T^*.$$

We will show that T_{co} is small by showing that T^* is small, and to do this, we will show that with high probability there is soon a time when $\mathbf{W}$ is 0 and when neither swapped queue in $\mathbf{X}$ is the memory queue. We will analyse $\mathbf{W}$ at some times $(J_i)_{i=0}^{\infty}$ to be defined later (again, these are not the jump times as defined in Section 1.4), that is, we will analyse the random walk $\mathbf{W}_{\mathbf{J}} = (W_{J_i})_{i=0}^{\infty}$.

We will apply Lemma 2.8 to $\mathbf{S} = \mathbf{W_J}$ roughly the same way we applied it to the level walk in Section 3.3, but with two main differences. The first difference is that here we will keep track of when either swapped queue is the memory queue: we will say that $\mathbf{S}$ is *good* at step i if neither swapped queue in $\mathbf{X}$ is the memory queue at time J_i, and that $\mathbf{S}$ is *bad* otherwise. The other difference concerns the background events A_i, but we will discuss this later.

Now let us say a little about Lemma 2.8. For the first condition, (2.18), we must show that $\mathbf{S}$ will either become good or increase, with probability bounded away from 0, when it is bad. For the fourth condition, (2.21), we must show that $\mathbf{S}$ will become good without changing value, with probability close to 1, when it is bad and above κ. These requirements lead us to the following definition.

Definition 6.9. Let $\mathbf{X}$ and $\mathbf{Y}$ have swap-adjacent initial states and be coupled by the swap coupling. Let $T > 0$ be an arrival time where $X_{T-} \frown Y_{T-}$. If $v(X_{T-}) \geq 1$, then we will say that T is *aligning* if the $\mathbf{X}$-customer selects only non-swapped empty queues. Else if $v(X_{T-}) = 0$, then we will say that T is *aligning* if the $\mathbf{X}$-customer selects only non-swapped queues of length 1.

Now we will show that if $\mathbf{S}$ is good or bad, then given an aligning arrival time, it will stay/become good. Moreover, it will not change value if it was non-zero immediately before the aligning arrival.

Lemma 6.10. *Let $\mathbf{X}$ and $\mathbf{Y}$ have swap-adjacent initial states and be coupled by the swap coupling. Let $T > 0$ be an aligning arrival time where $X_{T-} \frown Y_{T-}$ at k and l. Then $\Xi_T \neq k, l$. Moreover, the maximum length of a swapped queue W_t does not increase at time T, if $W_{T-} \geq 1$.*

Proof. There are two cases to consider.

1. $v(X_{T-}) \geq 1$. In this case, the $\mathbf{X}$-candidates list contains d non-swapped, empty queues and the memory queue. If $d = 1$, then the first queue in the $\mathbf{X}$-candidates list receives the customer and is saved as the memory queue. Else if $d \geq 2$, then the first queue in the $\mathbf{X}$-candidates list receives the customer, and the second is saved as the memory.

2. $v(X_{T-}) = 0$. In this case, the $\mathbf{X}$-candidates list contains d non-swapped queues of length 1 and the empty memory queue. Thus, the empty memory queue receives the customer, and then the first queue in the $\mathbf{X}$-candidates list is saved as the memory queue.

In both cases, one of the d selections becomes the memory queue in X_T. Since each selection is a non-swapped queue, it follows that $\Xi_T \neq k, l$. Moreover, since each selection has length at most 1 at time T, it follows that the maximum length of a swapped queue W_t cannot increase at time T, if $W_{T-} \geq 1$. □

Now we come to the second main difference: here, the background events A_i will include the event that the proportion of queues in $\mathbf{X}$ of length at least k is close to

$$\mu_k := \mathbb{E}[u_k(X)]$$

for long periods of time, where X has the equilibrium distribution for the lengths process; these events will hold with high probability by the concentration of measure results in Section 4.2. We will only need this concentration for $k = 1, 2$ and for sufficiently large n, so on the background events, the proportion of empty queues and queues of length 1 in $\mathbf{X}$ will be bounded away from 0 for long periods of time.

Recall that the μ_k are close to the a_k (by Lemma 5.9), and that the a_k satisfy $0 < a_2 < a_1 < 1$ (by Lemma 5.5).

Definition 6.11. For $t \geq 0$, let

$$E_t := \bigcap_{i=1}^{2} \left\{ |u_i(X_u) - \mu_i| \leq \tfrac{1}{8} \min(1 - a_1, a_1 - a_2) \text{ for all } 0 \leq u < t \right\}.$$

If E_t holds, then we will say $\mathbf{X}$ has *concentrated proportions* over $[0, t)$.

The following lemma says that if $T > 0$ is an event time, E_T holds and n is sufficiently large, then T has probability bounded away from 0 of being an aligning arrival.

Lemma 6.12. *There exists $n^* \geq 1$ such that the following holds. Let $n \geq n^*$, let $\mathbf{X}$ have an arbitrary initial distribution, and let $T > 0$ be an event time. Then*

$$\mathbb{P}(T \text{ is an aligning arrival} \mid \mathcal{F}_{T-}) \geq \frac{\lambda}{\lambda + 1} \left[\tfrac{1}{2} \min(1 - a_1, a_1 - a_2) \right]^d \quad \text{on } E_T.$$

Proof. Let $\psi := \frac{1}{8} \min(1 - a_1, a_1 - a_2)$, and let $c > 0$ denote the constant given by Lemma 5.9. Then on the event E_T, we have

$$|u_i(X_t) - a_i| \leq |u_i(X_t) - \mu_i| + |\mu_i - a_i| \leq \psi + \frac{c \ln^2 n}{\sqrt{n}} \leq 2\psi,$$

for $i = 1, 2$ and all $0 \leq t < T$, and if n^* is sufficiently large. Hence the proportion of non-swapped empty queues immediately before T is at least

$$1 - u_1(X_{T-}) - \tfrac{2}{n} \geq 1 - (a_1 + 2\psi) - \tfrac{2}{n} \geq \tfrac{3}{4}(1 - a_1) - \tfrac{2}{n} \geq \tfrac{1}{2}(1 - a_1),$$

and the proportion of non-swapped queues of length 1 immediately before T is at least

$$u_1(X_{T-}) - u_2(X_{T-}) - \tfrac{2}{n} \geq (a_1 - 2\psi) - (a_2 + 2\psi) - \tfrac{2}{n} \geq \tfrac{3}{4}(a_1 - a_2) - \tfrac{2}{n} \geq \tfrac{1}{2}(a_1 - a_2),$$

if n^* is sufficiently large. The result follows, for if $v(X_{T-}) \geq 1$, then T is aligning if the customer selects only non-swapped empty queues, and if $v(X_{T-}) = 0$, then T is aligning if the customer selects only non-swapped queues of length 1. □

Now we will show that in a swap coupling, under reasonable initial conditions, the two lengths processes rapidly coalesce.

Lemma 6.13. *Let $c > \frac{\lambda}{1-\lambda}$. Then there exists $0 < \beta = \beta(c) < 1$ such that the following holds. Let $n \geq 1$, let $\mathbf{X}$ and $\mathbf{Y}$ have initial states $x, y \in \mathcal{Q}_n$ where $x \frown y$ and $\|x\|_1 \leq cn$, and let $\mathbf{X}$ and $\mathbf{Y}$ be coupled by the swap coupling. Then*

$$\mathbb{E}[d_s(X_t, Y_t)] = \mathbb{E}[\mathbf{1}_{X_t \neq Y_t}] \leq e^{-\beta t} + 2e^{-\beta n} + \mathbb{P}\left(\overline{E_t}\right)$$

for all $t \geq \frac{1}{\beta}(\|x\|_\infty + 1)$.

Proof. Since the left-hand side is bounded by 1 and $\beta > 0$ may be arbitrarily small, it suffices to show the result for all sufficiently large n.

Let $\mathbf{W} = (W_t)_{t \geq 0}$ denote the swap walk. Thus, if k and l denote the swapped queues, then

$$W_t = \max(X_t(k), X_t(l)),$$

For $t \geq 0$, say $\mathbf{X}$ is *good* at time t if neither swapped queue in X_t is the memory queue, and *bad* otherwise. For $t \geq 0$, let

$$D_t := \{\Xi_t \neq k, l\}$$

denote the event that $\mathbf{X}$ is good at time t. Define the *change times* $J_0 := 0$ and

$$J_i := \inf\left\{t > J_{i-1} : \mathbf{1}_{D_t} \neq \mathbf{1}_{D_{t-}} \text{ or } W_t \neq W_{t-}\right\},$$

for all $i \geq 1$. That is, let J_i be the first time after J_{i-1} when either $\mathbf{X}$ starts/stops being good or when $\mathbf{W}$ changes values. The filtration $(\mathcal{G}_i)_{i=0}^\infty$ we will be using for Lemma 2.8 will be based on these change times: for $i \geq 0$, set $\mathcal{G}_i := \mathcal{F}_{J_{i+1}-}$ to be the σ-field generated by all events before J_{i+1}.

Now, for $t \geq 0$, let

$$C_t := \{\|X_r\|_1 \leq 2cn \text{ for all } 0 \leq r < t\}, \quad m := \left\lceil \tfrac{1}{4}t \right\rceil.$$

Then

$$\begin{aligned}\mathbb{P}(X_t \neq Y_t) \leq{}& \mathbb{P}(\{X_t \neq Y_t\} \cap \{J_m \leq t\} \cap C_t \cap E_t) \\ &+ \mathbb{P}(\{X_t \neq Y_t\} \cap \{J_m > t\} \cap E_t) + \mathbb{P}\left(\overline{C_t}\right) + \mathbb{P}\left(\overline{E_t}\right), \end{aligned} \tag{6.2}$$

for all $t \geq 0$. The first term will be where we apply Lemma 2.8, but let us bound the two middle terms first.

We claim that on $\{X_t \neq Y_t\} \cap E_t$, change times occur at rate at least 1 over $[0, t]$, if n is sufficiently large. To see this claim, consider a time $0 \leq r < t$. There are now two cases to consider.

1. If D_r holds, then neither swapped queue is the memory (that is, $\Xi_r \neq k, l$). As we have not yet coalesced, there is a unique longer swapped queue, and a sufficient condition for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is the unique longer swapped queue. Such events occur at rate $n \cdot \frac{1}{n} = 1$.

2. If $\overline{D_r}$ holds (so $\mathbf{X}$ is bad immediately before r), then some swapped queue is the memory (that is, $\Xi_r = k$ or l). A sufficient condition for $\mathbf{X}$ to become good is if we have an aligning arrival, by Lemma 6.10. On E_t, aligning arrivals occur at rate at least $(\lambda + 1)\, n \cdot \frac{\lambda\gamma}{\lambda+1} \geq 1$ over $[0, t]$, if n is sufficiently large; this holds by Lemma 6.12.

Hence the number of change times $N_t := \max\{i \geq 0 : J_i \leq t\}$ in $[0, t]$ stochastically dominates a $\mathrm{Po}(t)$ random variable on the event $\{X_t \neq Y_t\} \cap E_t$, if n is sufficiently large. By

Lemma 1.4 (with $\varepsilon = \frac{1}{2}$), we have

$$\mathbb{P}\left(\{X_t \neq Y_t\} \cap \{J_m > t\} \cap E_t\right) \leq \mathbb{P}\left(N_t < m\right) \leq \mathbb{P}\left(\mathrm{Po}\left(t\right) \leq \tfrac{1}{2}t\right) \leq 2e^{-\frac{1}{12}t}, \tag{6.3}$$

for all $t \geq 2$, and if n is sufficiently large. To see the second inequality, note that $m = \lceil \frac{1}{4}t \rceil \leq \frac{1}{2}t$.

By Lemma 2.5, there exists $\eta_1 = \eta_1(c) > 0$ such that

$$\mathbb{P}\left(\overline{C_t}\right) \leq 2e^{-\eta_1 n}, \tag{6.4}$$

for all $0 \leq t \leq e^{\eta_1 n}$.

Hence, by (6.2)-(6.4), we have

$$\mathbb{P}\left(X_t \neq Y_t\right) \leq \mathbb{P}\left(\{X_t \neq Y_t\} \cap \{J_m \leq t\} \cap C_t \cap E_t\right) + 2e^{-\frac{1}{12}t} + 2e^{-\eta_1 n} + \mathbb{P}\left(\overline{E_t}\right), \tag{6.5}$$

for all $2 \leq t \leq e^{\eta_1 n}$, if n is sufficiently large. Having bounded the two middle terms in (6.2) to obtain (6.5), we now turn our attention to the first term, which is where we will apply Lemma 2.8. We already defined the filtration $(\mathcal{G}_i)_{i=0}^{\infty}$ for this lemma by setting each $\mathcal{G}_i$ to be $\mathcal{F}_{J_{i+1}-}$. The background events are

$$A_i := \{X_r \neq Y_r \text{ for all } J_i \leq r < J_{i+1}\} \cap C_{J_{i+1}} \cap E_{J_{i+1}},$$

for $i \geq 0$. For $i \geq 0$, let

$$B_i := \{\mathbf{1}_{D_r} = 1 \text{ for all } J_i \leq r < J_{i+1}\}$$

denote the event that $\mathbf{X}$ is good at all times $J_i \leq r < J_{i+1}$. Note that A_i and B_i are both $\mathcal{G}_i$-measurable, since they depend only on the history of the process until but excluding J_{i+1}. The random walk is $\mathbf{S} = \mathbf{W_J}$, that is,

$$S_i := W_{J_i},$$

where $i \geq 0$. Note that each increment $Z_i := S_i - S_{i-1} = W_{J_i} - W_{J_{i-1}}$ is $\mathcal{G}_i$-measurable and $\{-1, 0, 1\}$-valued. Let the initial value be $S_0 = s \geq 0$. We will say that $\mathbf{S} = \mathbf{W_J}$ is *good* at step i if $\mathbf{X}$ is good at time J_i, and that $\mathbf{S}$ is *bad* otherwise. Thus, $\mathbf{S}$ is good at i if and only if D_{J_i} holds, and because $\mathbf{1_D}$ is constant between change times, it follows that $\mathbf{S}$ is good at step i if and only if B_i holds.

Having defined sequences of events and the random walk, we may now write (6.5) as

$$\begin{aligned}\mathbb{P}\left(X_t \neq Y_t\right) &\leq \mathbb{P}\left(\{X_t \neq Y_t\} \cap \{J_m \leq t\} \cap C_t \cap E_t\right) + 2e^{-\frac{1}{12}t} + 2e^{-\eta_1 n} + \mathbb{P}\left(\overline{E_t}\right) \\ &\leq \mathbb{P}\left(\bigcap_{i=1}^{m} \left(A_{i-1} \cap \left(\overline{B_i} \cup \{S_i > 0\}\right)\right)\right) + 2e^{-\frac{1}{12}t} + 2e^{-\eta_1 n} + \mathbb{P}\left(\overline{E_t}\right), \end{aligned} \tag{6.6}$$

for all $2 \leq t \leq e^{\eta_1 n}$, and if n is sufficiently large.

Next we define some constants. Let

$$\delta := \min\left(\frac{\lambda\gamma}{\lambda+1}, \frac{1-\lambda}{2\lambda d+1}\right),$$

where γ is as defined above. Define $0 < \varepsilon, \omega < 1$ as follows. If $d = 1$, then let $\varepsilon := 1$, else let ε be sufficiently small so that

$$d\varepsilon^{d-1} \leq \frac{1}{2\lambda d + 1}.$$

Let ω be sufficiently small so that

$$\frac{2\omega^d}{\gamma} \leq \tfrac{1}{2}\delta.$$

Finally, let

$$\kappa = \kappa(c) := \left\lceil \frac{2c}{\min(\varepsilon, \omega)} \right\rceil.$$

Now we will show that the hypotheses of Lemma 2.8 hold with the filtration $(\mathcal{G}_i)_{i=0}^{\infty}$, the sequences of events $(A_i)_{i=0}^{\infty}$ and $(B_i)_{i=0}^{\infty}$, the random walk $\mathbf{S} = (S_i)_{i=0}^{\infty}$, and the constants δ and κ, all as defined above. There are now four conditions to verify: (2.18)-(2.21).

1. For condition (2.18), we are looking at

$$\mathbb{P}(B_{i+1} \cup \{Z_{i+1} = 1\} \mid \mathcal{G}_i), \quad \text{on } A_i \cap \overline{B_i}.$$

Now

$$A_i \cap \overline{B_i} \subseteq U_{i,1} := \left\{\Xi_{J_{i+1}-} = k \text{ or } l\right\} \cap E_{J_{i+1}}.$$

We will work on the event $U_{i,1}$, which says that immediately before J_{i+1}, some swapped queue is the memory queue, and $\mathbf{X}$ has concentrated proportions over $[0, J_{i+1})$. Since $B_{i+1} \cup \{Z_{i+1} = 1\}$ denotes the event that the $(i+1)^{\text{st}}$ change time is one where $\mathbf{S}$ becomes good or increases, we may write

$$\mathbb{P}(B_{i+1} \cup \{Z_{i+1} = 1\} \mid \mathcal{G}_i) \geq \frac{p_1}{q_1}, \quad \text{on } A_i \cap \overline{B_i}, \tag{6.7}$$

where p_1 is a lower bound on the rate of events where $\mathbf{X}$ becomes good or $\mathbf{W}$ increases, and q_1 is an upper bound on the rate of events where $\mathbf{X}$ becomes good or $\mathbf{W}$ changes value (i.e., change times).

We may take the lower bound $p_1 := \lambda\gamma n$, if n is sufficiently large. To see this, note that a sufficient condition for $\mathbf{X}$ to become good is if we have an aligning arrival, by Lemma 6.10. On $U_{i,1}$, aligning arrivals occur at rate at least $(\lambda+1)n \cdot \frac{\lambda\gamma}{\lambda+1} = \lambda\gamma n$ over $[0, J_{i+1}]$, if n is sufficiently large; this holds by Lemma 6.12.

We may take the upper bound $q_1 = (\lambda+1)n$, the rate of all events.

Then (6.7) gives

$$\mathbb{P}(B_{i+1} \cup \{Z_{i+1} = 1\} \mid \mathcal{G}_i) \geq \frac{\lambda\gamma n}{(\lambda+1)n} \geq \delta, \quad \text{on } A_i \cap \overline{B_i},$$

and (2.18) holds, if n is sufficiently large.

2. For condition (2.19), we are looking at

$$\mathbb{P}(B_{i+1} \cap \{Z_{i+1} = -1\} \mid \mathcal{G}_i), \quad A_i \cap B_i \cap \{S_i > 0\}.$$

Now

$$A_i \cap B_i \cap \{S_i > 0\} \subseteq U_{i,2} := \left\{X_{J_{i+1}-} \neq Y_{J_{i+1}-}\right\} \cap \left\{\Xi_{J_{i+1}-} \neq k, l\right\}.$$

We will work on the event $U_{i,2}$, which says that immediately before J_{i+1}, $\mathbf{X}$ and $\mathbf{Y}$ are not coalesced and neither swapped queue is the memory queue. Hence there is a unique longer swapped queue which we will assume, without loss of generality, is queue k. Since $B_{i+1} \cap \{Z_{i+1} = -1\}$ denotes the event that the $(i+1)^{\text{st}}$ change time is one where $\mathbf{S}$ remains good and decreases, we may write

$$\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = -1\} \mid \mathcal{G}_i\right) \geq \frac{p_2}{q_2}, \quad \text{on } A_i \cap B_i \cap \{S_i > 0\}, \tag{6.8}$$

where p_2 is a lower bound on the rate of events where $\mathbf{X}$ remains good and $\mathbf{W}$ decreases, and q_2 is an upper bound on the rate of events where $\mathbf{X}$ becomes bad or $\mathbf{W}$ changes value (i.e., change times).

We may take the lower bound $p_2 := 1$. To see this, note that a sufficient condition for $\mathbf{X}$ to remain good and for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is queue k. Such events occur at rate $n \cdot \frac{1}{n} = 1$.

We may take the upper bound $q_2 := 2\lambda d + 1$. To see this, note that a necessary condition for $\mathbf{X}$ to become bad or for $\mathbf{W}$ to increase is if we have an arrival where the $\mathbf{X}$-customer selects some swapped queue at least once (since neither is the memory queue). Such events occur at rate at most $\lambda n \cdot \frac{2d}{n} = 2\lambda d$. A necessary condition for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is queue k. Such events occur at rate $n \cdot \frac{1}{n} = 1$.

Then (6.8) gives

$$\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = -1\} \mid \mathcal{F}_i\right) \geq \frac{1}{2\lambda d + 1} \geq \delta, \quad \text{on } A_i \cap B_i \cap \{S_i > 0\}, \tag{6.9}$$

and (2.19) holds.

3. For condition (2.20), we are looking at

$$\mathbb{P}\left(Z_{i+1} = 1 \mid \mathcal{G}_i\right), \quad \text{on } A_i \cap B_i \cap \{S_i > \kappa\}.$$

Now

$$A_i \cap B_i \cap \{S_i > \kappa\} \subseteq U_{i,3} := \left\{X_{J_{i+1}-} \neq Y_{J_{i+1}-}\right\} \cap \left\{\Xi_{J_{i+1}-} \neq k, l\right\} \cap \left\{W_{J_{i+1}-} > \kappa\right\} \cap C_{J_{i+1}}.$$

We will work on the event $U_{i,3}$, which says that immediately before J_{i+1}, $\mathbf{X}$ and $\mathbf{Y}$ are not coalesced, neither swapped queue is the memory queue, the maximum length of a swapped queue is greater than κ, and the number of customers is at most $2cn$. Hence there is a unique longer swapped queue which we will assume, without loss of generality, is queue k (so $X_{J_{i+1}-}(k) > \kappa$), whence the proportion of queues at least

as long as queue k is less than

$$u_\kappa\left(X_{J_{i+1}-}\right) \leq \frac{\left\|X_{J_{i+1}-}\right\|_1}{n\kappa} \leq \frac{2c}{\kappa} \leq \min\left(\varepsilon, \omega\right) \leq \varepsilon.$$

Since $\{Z_{i+1} = 1\}$ denotes the event that the $(i+1)^{\text{st}}$ change time is one where $\mathbf{S}$ increases, we may write

$$\mathbb{P}\left(Z_{i+1} = 1 \mid \mathcal{G}_i\right) \leq \frac{p_3}{q_3}, \quad \text{on } A_i \cap B_i \cap \{S_i > \kappa\}, \tag{6.10}$$

where p_3 is an upper bound on the rate of events where $\mathbf{W}$ increases, and q_3 is a lower bound on the rate of events where $\mathbf{X}$ becomes bad or $\mathbf{W}$ changes value (i.e., change times). Note that if $\mathbf{W}$ increases at J_{i+1}, then immediately before J_{i+1}, queue k cannot be longer than the memory queue. That is, we have

$$X_{J_{i+1}-}(k) \leq X_{J_{i+1}-}\left(\Xi_{J_{i+1}-}\right) = v\left(X_{J_{i+1}-}\right). \tag{6.11}$$

There are now two cases to consider.

(a) Case 1: $d \geq 2$. We may take the upper bound $p_3 := \lambda d \varepsilon^{d-1}$. To see this, note that a necessary condition for $\mathbf{W}$ to increase is if we have an arrival where the $\mathbf{X}$-customer selects only queues as long as queue k, and he/she selects queue k at least once (since queue k is not the memory queue). Such events occur at rate at most $\lambda n \cdot \frac{d}{n}\varepsilon^{d-1} = \lambda d \varepsilon^{d-1}$.

We may take the lower bound $q_3 := 1$. To see this, note that a sufficient condition for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is queue k. Such events occur at rate $n \cdot \frac{1}{n} = 1$.

Then (6.10) gives

$$\mathbb{P}\left(Z_{i+1} = 1 \mid \mathcal{G}_i\right) \leq \frac{\lambda d \varepsilon^{d-1}}{1} \leq \frac{\lambda}{2\lambda d + 1}, \quad \text{on } A_i \cap B_i \cap \{S_i > \kappa\}.$$

The last inequality holds since $d \geq 2$. By (6.9), we have

$$\begin{aligned}
\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = -1\} \mid \mathcal{G}_i\right) &\geq \frac{1}{2\lambda d + 1} \\
&= \frac{\lambda}{2\lambda d + 1} + \frac{1 - \lambda}{2\lambda d + 1} \\
&\geq \mathbb{P}\left(Z_{i+1} = 1 \mid \mathcal{F}_i\right) + \delta, \quad \text{on } A_i \cap B_i \cap \{S_i > \kappa\},
\end{aligned}$$

and (2.20) holds.

(b) Case 2: $d = 1$. We may take the upper bound $p_3 := \lambda$. To see this, note that a necessary condition for $\mathbf{W}$ to increase is if we have an arrival time where the $\mathbf{X}$-customer selects queue k (since queue k is not the memory queue). Such events occurs at rate $\lambda n \cdot \frac{1}{n} = \lambda$.

We may take the lower bound $q_3 := 2\lambda d + 1$. To see this, note that a sufficient condition for $\mathbf{W}$ to increase is if we have an arrival where the $\mathbf{X}$-customer selects queue k (since queue k is neither the memory queue nor longer than it,

by (6.11)). Such events occur at rate $\lambda n \cdot \frac{1}{n} = \lambda d$. A sufficient condition for $\mathbf{X}$ to become bad is if we have an arrival where the $\mathbf{X}$-customer selects queue l (since queue l is shorter than queue k, and queue k is not longer that the memory queue, by (6.11)). Such events occur at rate $\lambda n \cdot \frac{1}{n} = \lambda d$. A sufficient condition for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is queue k. Such events occur at rate $n \cdot \frac{1}{n} = 1$.

Then (6.10) gives

$$\mathbb{P}\left(Z_{i+1} = 1 \mid \mathcal{G}_i\right) \leq \frac{\lambda}{2\lambda d + 1}, \quad \text{on } A_i \cap B_i \cap \{S_i > \kappa\},$$

and (2.20) holds by the same calculation as in case 1.

4. For condition (2.21), we will first look at

$$\mathbb{P}\left(Z_{i+1} \neq 0 \mid \mathcal{G}_i\right), \quad \text{on } A_i \cap \overline{B_i} \cap \{S_i > \kappa\}.$$

Now

$$A_i \cap \overline{B_i} \cap \{S_i > \kappa\} \subseteq U_{i,4} := \left\{\Xi_{J_{i+1}-} = k \text{ or } l\right\} \cap \left\{W_{J_{i+1}-} > \kappa\right\} \cap C_{J_{i+1}} \cap E_{J_{i+1}}.$$

We will work on the event $U_{i,4}$, which says that immediately before J_{i+1}, some swapped queue is the memory queue, the maximum length of a swapped queue is greater than κ, the number of customers is at most $2cn$, and $\mathbf{X}$ has concentrated proportions over $[0, J_{i+1})$. We will assume, without loss of generality, that queue k is not shorter than queue l (so $X_{J_{i+1}-}(k) > \kappa$), whence the proportion of queues at least as long as queue k is less than

$$u_\kappa\left(X_{J_{i+1}-}\right) \leq \frac{\left\|X_{J_{i+1}-}\right\|_1}{n\kappa} \leq \frac{2c}{\kappa} \leq \min\left(\varepsilon, \omega\right) \leq \omega.$$

Since $\{Z_{i+1} \neq 0\}$ denotes the event that the $(i+1)^{\text{st}}$ change time is one where $\mathbf{S}$ changes value, we may write

$$\mathbb{P}\left(Z_{i+1} \neq 0 \mid \mathcal{G}_i\right) \leq \frac{p_4}{q_4}, \quad \text{on } A_i \cap \overline{B_i} \cap \{S_i > \kappa\}, \tag{6.12}$$

where p_4 is an upper bound on the rate of events where $\mathbf{W}$ change values, and q_4 is a lower bound on the rate of events where $\mathbf{X}$ becomes good or $\mathbf{W}$ changes value (i.e., change times).

We may take the upper bound $p_4 := \lambda n \omega^d + 2$. To see this, note that a necessary condition for $\mathbf{W}$ to increase is if we have an arrival time where the $\mathbf{X}$-customer selects only queues as long as queue k. Such events occur at rate at most $\lambda n \omega^d$. A necessary condition for $\mathbf{W}$ to decrease is if we have a potential departure where the $\mathbf{X}$-selection is one of the two swapped queues. Such events occur at rate $n \cdot \frac{2}{n} = 2$.

We may take the lower bound $q_4 := \lambda \gamma n$, if n is sufficiently large. To see this, note that a sufficient condition for $\mathbf{X}$ to become good is if we have an aligning arrival time, by Lemma 6.10. On $U_{i,4}$, aligning arrivals occur at rate at least $(\lambda + 1)\, n \cdot \frac{\lambda\gamma}{\lambda+1} = \lambda\gamma n$

over $[0, J_{i+1}]$, if n is sufficiently large; this holds by Lemma 6.12.

Then (6.12) gives

$$\mathbb{P}\left(Z_{i+1} \neq 0 \mid \mathcal{G}_i\right) \leq \frac{\lambda n \omega^d + 2}{\lambda \gamma n} \leq \frac{2\omega^d}{\gamma} \leq \tfrac{1}{2}\delta, \quad \text{on } A_i \cap \overline{B_i} \cap \{S_i > \kappa\},$$

if n is sufficiently large.

At a change time, $\mathbf{S}$ must either become good and/or change value, so

$$\begin{aligned}\mathbb{P}\left(B_{i+1} \cap \{Z_{i+1} = 0\} \mid \mathcal{G}_i\right) &= \mathbb{P}\left(Z_{i+1} = 0 \mid \mathcal{G}_i\right) \\ &\geq 1 - \tfrac{1}{2}\delta, \quad \text{on } A_i \cap \overline{B_i} \cap \{S_i > \kappa\},\end{aligned}$$

if n is sufficiently large, and (2.21) holds.

Since we have shown that the hypotheses of Lemma 2.8 hold if n is sufficiently large, there exists a constant $\eta_2 = \eta_2(c) > 0$ such that (6.6) becomes

$$\begin{aligned}\mathbb{P}\left(X_t \neq Y_t\right) &\leq \mathbb{P}\left(\bigcap_{i=1}^{m} \left(A_{i-1} \cap \left(\overline{B_i} \cup \{S_i > 0\}\right)\right)\right) + 2e^{-\frac{1}{12}t} + 2e^{-\eta_1 n} + \mathbb{P}\left(\overline{E_t}\right) \\ &\leq 2e^{-\eta_2 m} + \mathbf{1}_{s > \eta_2 m} + 2e^{-\frac{1}{12}t} + 2e^{-\eta_1 n} + \mathbb{P}\left(\overline{E_t}\right),\end{aligned}$$

for all $2 \leq t \leq e^{\eta_1 n}$, and if n is sufficiently large. We will also assume, without loss of generality, that $0 < \eta_2 < 1$.

Let $\eta_3 = \eta_3(c) := \frac{1}{2}\min\left(\eta_1, \eta_2, \frac{1}{12}\right)$, then

$$\mathbb{P}\left(X_t \neq Y_t\right) \leq 4e^{-2\eta_3 t} + 2e^{-2\eta_3 n} + \mathbb{P}\left(\overline{E_t}\right) \leq e^{-\eta_3 t} + e^{-\eta_3 n} + \mathbb{P}\left(\overline{E_t}\right),$$

for all $\frac{4}{\eta_3}\max(s, 1) \leq t \leq e^{\eta_3 n}$, and if n is sufficiently large. To see the first inequality, note that $t \geq \frac{4}{\eta_3}\max(s, 1) \geq 2$ and that $\eta_2 m \geq \eta_3 \frac{t}{4} \geq \max(s, 1) \geq s$. To see the second inequality, note that $4 \leq e^{\eta_3 t}$ (since $t \geq \frac{4}{\eta_3} \geq \frac{\ln 4}{\eta_3}$). We can remove the upper bound on t as follows. If $t > e^{\eta_3 n}$, then

$$\mathbb{P}\left(X_t \neq Y_t\right) \leq \mathbb{P}\left(X_n \neq Y_n\right) + \mathbb{P}\left(\overline{E_t}\right) \leq 2e^{-\eta_3 n} + \mathbb{P}\left(\overline{E_t}\right)$$

if n is sufficiently large so that $e^{\eta_3 n} > n$. Let $\beta = \beta(c) := \frac{1}{4}\eta_3$, then

$$\mathbb{P}\left(X_t \neq Y_t\right) \leq e^{-\beta t} + 2e^{-\beta n} + \mathbb{P}\left(\overline{E_t}\right)$$

for all $t \geq \frac{1}{\beta}\left(\|x\|_\infty + 1\right)$, and if n is sufficiently large. To see this, note that $\frac{1}{\beta}\left(\|x\|_\infty + 1\right) \geq \frac{4}{\eta_3}\max(s, 1)$. □

We then have the following result.

Lemma 6.14. *Let $c > \frac{\lambda}{1-\lambda}$, then let $0 < \beta = \beta(c) < 1$ denote the constant given by Lemma 6.13. Let $n \geq 1$, let $\mathbf{X}$ and $\mathbf{Y}$ have initial states $x, y \in \mathcal{Q}_n$, respectively, where $x \equiv y$, $\|x\|_1 \leq cn$ and $\|x\|_1 \leq \beta t - 1$, and let $\mathbf{X}$ and $\mathbf{Y}$ be coupled by a swap coupling. Then*

$$\mathbb{E}\left[d_s\left(X_t, Y_t\right)\right] \leq n\left(e^{-\beta t} + 2e^{-\beta n} + \mathbb{P}\left(\overline{E_t}\right)\right)$$

for all $t \geq 0$.

Proof. Let m and $\mathbf{Z}^j$ be as in Definition 6.6. By Lemma 6.3, we have

$$m = d_s(x, y) \leq n.$$

We also have

$$\|z_i\|_1 = \|x\|_1 \leq cn, \quad \|z_i\|_\infty = \|x\|_\infty \leq \beta t - 1,$$

for all $0 \leq i \leq m$. Hence

$$\begin{aligned}
\mathbb{E}\left[d_s(X_t, Y_t)\right] &\leq \sum_{i=1}^{m} \mathbb{E}\left[d_s\left(Z_t^{i-1}, Z_t^i\right)\right] \\
&\leq \sum_{i=1}^{m} \left(e^{-\beta t} + 2e^{-\beta n} + \mathbb{P}\left(\overline{E_t}\right)\right) \\
&\leq n\left(e^{-\beta t} + 2e^{-\beta n} + \mathbb{P}\left(\overline{E_t}\right)\right),
\end{aligned}$$

and we are done. □

The following is the main result of this section, and will subsequently give Theorem 1.1.

Theorem 6.15. *Let* $c > \frac{\lambda}{1-\lambda}$. *Then there exists* $\eta = \eta(c) > 0$ *such that the following holds. Let* $n \geq 1$, *let* $\mathbf{X}$ *have an arbitrary initial distribution, and let* $\mathbf{Y}$ *be in equilibrium. Then there exists a coupling of* $\mathbf{X}$ *and* $\mathbf{Y}$ *such that*

$$\mathbb{P}(X_t \neq Y_t) \leq ne^{-\eta t} + 2e^{-\eta\sqrt{n}} + \mathbb{P}\left(\|X_0\|_1 > cn\right) + \mathbb{P}\left(\|X_0\|_\infty > \eta t\right)$$

for all $t \geq 0$.

Proof. First we will define some constants. Let $0 < \beta = \beta(c) < 1$ denote the constant given by Lemma 6.14 (with the same c). Let

$$\delta := \tfrac{1}{8}\min(1 - a_1, a_1 - a_2) > 0.$$

Let $\eta_1 > 0$ denote the constant given by Lemma 4.4 with $z = \delta$, so that

$$\mathbb{P}\left(|u_i(Y_r) - \mathbb{E}\left[u_i(Y_r)\right]| \geq \delta \text{ for some } 0 \leq r \leq e^{\eta_1\sqrt{n}}\right) \leq 2e^{-\eta_1\sqrt{n}} \tag{6.13}$$

for all $i \geq 1$. Let $\eta_2 = \eta_2(c) > 0$ and $\eta_3 > 0$ denote the constants given by Lemma 2.5 (with the same c), and let $\eta_4 = \eta_4(c) > 0$ denote the constant given by Theorem 3.14 (with the same c). Let

$$\begin{aligned}
\eta_5 = \eta_5(c) &:= \tfrac{1}{2}\min\left(\tfrac{1}{2}\beta, \eta_1\beta, \eta_1, \eta_2, \tfrac{1}{4}\eta_3\beta, \tfrac{1}{2}\eta_4\right), \\
t^* = t^*(c) &:= \max\left(\frac{5}{\beta}, \frac{\ln 3}{\eta_5}\right).
\end{aligned}$$

Let $n^* \geq 1$ be sufficiently large so that

$$\frac{1}{\beta} < e^{\eta_1\sqrt{n}}, \quad 7n + 3 \leq 2e^{\eta_5\sqrt{n}}, \tag{6.14}$$

for all $n \geq n^*$. Finally, let

$$\eta = \eta(c) := \min\left(\eta_5, \frac{\ln 2}{\sqrt{n^*}}\right),$$

so that $e^{\eta\sqrt{n^*}} \leq 2$.

Note that if $t \leq t^*$ and $n \geq n^*$, then $ne^{-\eta t} \geq ne^{-\eta_5 t} \geq ne^{-\eta_5 t^*} \geq 1$. Similarly, if $n \leq n^*$, then $2e^{-\eta\sqrt{n}} \geq 2e^{-\eta\sqrt{n^*}} \geq 1$. Hence, we will assume that $t \geq t^*$ and $n \geq n^*$, since there is nothing to prove otherwise.

The coupling $(\mathbf{X}, \mathbf{Y})$ will be defined as follows: run a profile coupling until the processes are profile-equivalent, and then run a swap coupling until the processes coalesce. In particular, we will be checking for profile-equivalence at time $\frac{1}{2}t$. If we have this, then we will check for coalescence at time

$$\tfrac{1}{2}t + h(t,n) \leq t,$$

where $h(t,n) := \min\left(\frac{1}{2}t, e^{\eta_1\sqrt{n}}\right)$.

Let F_t and G_t denote the events that $\left\|X_{t/2}\right\|_1 \leq cn$ and $\left\|X_{t/2}\right\|_\infty \leq \beta h(t,n) - 1$, respectively. Note that $\beta h(t,n) - 1 > 0$ by (6.14) and the fact that $t \geq t^* > \frac{2}{\beta}$. Then, by Lemma 6.7, we have

$$\begin{aligned}
\mathbb{P}(X_t \neq Y_t) &\leq \mathbb{E}\left[\mathbf{1}_{X_t \neq Y_t}\mathbf{1}_{X_{t/2} \equiv Y_{t/2}}\mathbf{1}_{F_t}\mathbf{1}_{G_t}\right] + \mathbb{P}\left(\left\{X_{t/2} \not\equiv Y_{t/2}\right\} \cup \overline{F_t} \cup \overline{G_t}\right) \\
&\leq \mathbb{E}\left[d_s(X_t, Y_t)\,\mathbf{1}_{X_{t/2} \equiv Y_{t/2}}\mathbf{1}_{F_t}\mathbf{1}_{G_t}\right] + \mathbb{P}\left(\left\{X_{t/2} \not\equiv Y_{t/2}\right\} \cup \overline{F_t} \cup \overline{G_t}\right) \\
&\leq \mathbb{E}\left[d_s\left(X_{t/2+h(t,n)}, Y_{t/2+h(t,n)}\right)\mathbf{1}_{X_{t/2} \equiv Y_{t/2}}\mathbf{1}_{F_t}\mathbf{1}_{G_t}\right] \\
&\qquad + \mathbb{P}\left(\left\{X_{t/2} \equiv Y_{t/2}\right\} \cap \left(\overline{F_t} \cup \overline{G_t}\right)\right) + \mathbb{P}\left(X_{t/2} \not\equiv Y_{t/2}\right). \qquad (6.15)
\end{aligned}$$

On the event that $X_{t/2} \equiv Y_{t/2}$, we have $X_r \equiv Y_r$ for all $r \geq \frac{1}{2}t$, so $d_s(X_t, Y_t) < \infty$. Moreover, by time $\frac{1}{2}t$, we are running a swap coupling of the lengths processes $\mathbf{Z} = (Z_r)_{r \geq 0}$ and $\mathbf{W} = (W_r)_{r \geq 0}$ defined by $Z_r := X_{t/2+r}$ and $W_r := Y_{t/2+r}$, whose initial states satisfy

$$Z_0 \equiv W_0, \quad \|Z_0\|_1 \leq cn, \quad \|Z_0\|_\infty \leq \beta h(t,n) - 1, \quad \text{on } \left\{X_{t/2} \equiv Y_{t/2}\right\} \cap F_t \cap G_t.$$

Hence, Lemma 6.14 implies that

$$\begin{aligned}
&\mathbb{E}\left[d_s\left(X_{t/2+h(t,n)}, Y_{t/2+h(t,n)}\right)\mathbf{1}_{X_{t/2} \equiv Y_{t/2}}\mathbf{1}_{F_t}\mathbf{1}_{G_t}\right] \\
&\leq \mathbb{E}\left[d_s\left(Z_{h(t,n)}, W_{h(t,n)}\right)\mathbf{1}_{Z_0 \equiv W_0}\mathbf{1}_{\|Z_0\|_1 \leq cn}\mathbf{1}_{\|Z_0\|_\infty \leq \beta h(t,n)-1}\right] \\
&\leq n\left(e^{-\beta h(t,n)} + 2e^{-\beta n} + \mathbb{P}\left(\overline{E'_{h(t,n)}}\right)\right), \qquad (6.16)
\end{aligned}$$

where

$$E'_{h(t,n)} := \bigcap_{i=1}^{2}\left\{|u_i(Z_r) - \mu_i| \leq \delta \text{ for all } 0 \leq r < h(t,n)\right\}$$

is an analogue of the event E_t, as defined in Definition 6.11. Note that $\mathbf{Z}$ and $\mathbf{W}$ are profile-equivalent for all time, so that $u_i(Z_r) = u_i(W_r)$ for all $r \geq 0$ and $i \geq 1$. Also note

that $\mathbf{W}$ is in equilibrium. Hence, (6.13) gives

$$\mathbb{P}\left(\overline{E'_{h(t,n)}}\right) \le \mathbb{P}\left(\bigcup_{i=1}^{2}\left\{|u_i(W_r)-\mu_i|>\delta \text{ for some } 0\le r<e^{\eta_1\sqrt{n}}\right\}\right) \le 4e^{-\eta_1\sqrt{n}}. \quad (6.17)$$

Regardless of whether $\frac{1}{2}t$ or $e^{\eta_1\sqrt{n}}$ is smaller, we have

$$e^{-\beta h(t,n)} \le e^{-\frac{1}{2}\beta t} + e^{-\exp\left(\eta_1\sqrt{n}\right)} \le e^{-2\eta_5 t} + e^{-2\eta_5\sqrt{n}},$$

whence (6.16) and (6.17) give

$$\begin{aligned}
&\mathbb{E}\left[d_s\left(X_{t/2+h(t,n)}, Y_{t/2+h(t,n)}\right)\mathbf{1}_{X_{t/2}\equiv Y_{t/2}}\mathbf{1}_{F_t}\mathbf{1}_{G_t}\right]\\
&\le n\left(e^{-\beta h(t,n)} + 2e^{-\beta n} + 4e^{-\eta_1\sqrt{n}}\right)\\
&\le ne^{-2\eta_5 t} + 7ne^{-2\eta_5\sqrt{n}}. \qquad (6.18)
\end{aligned}$$

Note that $t \ge t^* > \frac{4}{\beta}$, so the constants $\eta_2 = \eta_2(c) > 0$ and $\eta_3 > 0$ given by Lemma 2.5 satisfy

$$\begin{aligned}
\mathbb{P}\left(\left\{X_{t/2}\equiv Y_{t/2}\right\}\cap\left(\overline{F_t}\cup\overline{G_t}\right)\right) &\le \mathbb{P}\left(\left\{X_{t/2}\equiv Y_{t/2}\right\}\cap\left\{\left\|X_{t/2}\right\|_1 > cn\right\}\right)\\
&\quad + \mathbb{P}\left(\left\{X_{t/2}\equiv Y_{t/2}\right\}\cap\left\{\left\|X_{t/2}\right\|_\infty > \tfrac{1}{2}\beta t - 1\right\}\right)\\
&\le \mathbb{P}\left(\left\|Y_{t/2}\right\|_1 > cn\right) + \mathbb{P}\left(\left\|Y_{t/2}\right\|_\infty > \tfrac{1}{4}\beta t\right)\\
&\le e^{-\eta_2 n} + ne^{-\frac{1}{4}\eta_3\beta t}, \qquad (6.19)
\end{aligned}$$

and the constant $\eta_4 = \eta_4(c) > 0$ given by Theorem 3.14 satisfies

$$\mathbb{P}\left(X_{t/2}\not\equiv Y_{t/2}\right) \le ne^{-\frac{1}{2}\eta_4 t} + 2e^{-\eta_4 n} + \mathbb{P}\left(\|X_0\|_1 > cn\right) + \mathbb{P}\left(\|X_0\|_\infty > \tfrac{1}{2}\eta_4 t\right). \quad (6.20)$$

Hence, by (6.15) and (6.18)-(6.20), we have

$$\begin{aligned}
\mathbb{P}(X_t \ne Y_t) &\le ne^{-2\eta_5 t} + 7ne^{-2\eta_5\sqrt{n}} + e^{-\eta_2 n} + ne^{-\frac{1}{4}\eta_3\beta t}\\
&\quad + ne^{-\frac{1}{2}\eta_4 t} + 2e^{-\eta_4 n} + \mathbb{P}\left(\|X_0\|_1 > cn\right) + \mathbb{P}\left(\|X_0\|_\infty > \tfrac{1}{2}\eta_4 t\right)\\
&\le 3ne^{-2\eta_5 t} + (7n+3)e^{-2\eta_5\sqrt{n}} + \mathbb{P}\left(\|X_0\|_1 > cn\right) + \mathbb{P}\left(\|X_0\|_\infty > 2\eta_5 t\right).
\end{aligned}$$

Now $3 \le e^{\eta_5 t}$ (since $t \ge t^*$) and $7n+3 \le 2e^{\eta_5\sqrt{n}}$ (by (6.14)), so

$$\mathbb{P}(X_t \ne Y_t) \le ne^{-\eta_5 t} + 2e^{-\eta_5\sqrt{n}} + \mathbb{P}\left(\|X_0\|_1 > cn\right) + \mathbb{P}\left(\|X_0\|_\infty > \eta_5 t\right).$$

Hence the result follows if $\eta \le \eta_5$. □

Theorem 1.1 then follows from (1.12) and Theorem 6.15.

Chapter 7

Maximum queue length

In this chapter, we will analyse the equilibrium behaviour of the maximum queue length function $\|\cdot\|_\infty$, and then prove Theorem 1.2. This chapter is based on Chapter 7 of [10] by Luczak and McDiarmid.

Recall that in Section 1.2, we defined sequences $(a_i)_{i=0}^\infty$ and $(b_i)_{i=0}^\infty$ by setting $a_0 = b_0 = 1$ and

$$a_i := \lambda a_{i-1}^d b_{i-1}, \quad b_i := \frac{a_i^d b_{i-1}}{1 - d\,(a_{i-1} - a_i)\,a_i^{d-1}},$$

for all $i \geq 1$. We also let

$$i_n^* := \min\left\{i \geq 1 : a_i \leq \frac{\ln^2 n}{\sqrt{n}}\right\}, \quad \alpha := d + \tfrac{1}{2} + \sqrt{d^2 + \tfrac{1}{4}}.$$

Also recall that $u_i(x)$ gives the proportion of queues in $x \in \mathcal{Q}_n$ of length at least $i \geq 0$, and that $v(x)$ gives the length of the memory queue in x.

7.1 Equilibrium behaviour

In this section, we will prove Theorem 1.2. This result is analogous to Theorem 1.3 in [10] by Luczak and McDiarmid, which showed two-point concentration of the equilibrium maximum queue length. Here, we will show this by observing that in equilibrium, it is both unlikely for any single queue to be very long, and unlikely for the memory queue to be very long.

Lemma 7.1. *There exists $c_1 > 0$ such that the following holds. Let $n \geq 1$, and let X have the equilibrium distribution for the lengths process. For $i \geq 1$, let $\nu_i := \mathbb{E}\left[\mathbf{1}_{v(X)\geq i}\right]$. Then*

$$\nu_{i_n^*+z-1} \leq \left(\frac{c_1 \ln^{2d} n}{n^{\min(d/2,1)}}\right)^z \frac{\ln^2 n}{n^{d/(2\alpha)}},$$

for all $z \geq 1$.

Proof. Since the left-hand side is bounded by 1 and $c_1 > 0$ may be arbitrarily large, it suffices to show the result for all sufficiently large n.

Let $\mathbf{X}$ be in equilibrium, and let k_1 denote the constant given by Lemma 5.10. Let

$$E := \left\{ \sup_{i \geq 1} |u_i(X_t) - a_i| < \frac{2k_1 \ln^2 n}{\sqrt{n}} \text{ for all } 0 \leq t \leq n \right\},$$

so that Lemma 5.10 (with $z = 2k_1$ and $r = 1$) gives $\eta > 0$ such that

$$\mathbb{P}\left(\overline{E}\right) \leq 2e^{-\eta \ln^2 n}. \tag{7.1}$$

Now let us consider $\mathbf{X}$ at the first event time $T > 0$. For $i \geq 1$, let A_i denote the event that T is an arrival where the customer selects only queues of length at least i. Let M denote the event that T is an arrival where the customer selects the memory queue, and let D denote the event that T is a potential departure. Then

$$\begin{aligned} \{v(X_T) \geq i\} \subseteq [\{v(X_{T-}) \geq i\} \cap A_{i-1}] \\ \cup [\{v(X_{T-}) = i-1\} \cap (A_i \cup M)] \cup [\{v(X_{T-}) \geq i\} \cap D]. \end{aligned} \tag{7.2}$$

To see the first two terms in (7.2), note that if T is an arrival, then a necessary condition for $v(X_T) \geq i$ is that $v(X_{T-}) \geq i-1$. In particular, if $v(X_{T-}) \geq i$, then as the candidates list necessarily contains only queues of length at least $i-1$, we deduce that A_{i-1} occurs. On the other hand, if $v(X_{T-}) = i-1$, then as the candidates list necessarily has the memory queue as its unique shortest queue, we deduce that $A_i \cup M$ occurs. To see the third term, note that if T is a potential departure, then a necessary condition for $v(X_T) \geq i$ is that $v(X_{T-}) \geq i$. Thus (7.2) holds.

Let C denote the event that T is an arrival where the customer selects only queues from the longest $\frac{1}{2^{1/d}}$ of the queues. Then

$$A_{i-1} \cap E \cap \{T \leq n\} \subseteq C \tag{7.3}$$

for all $i \geq i_n^*$, if n is sufficiently large. To see this inequality, note that the proportion of queues of length at least $i-1$ immediately before T satisfies

$$u_{i-1}(X_{T-}) \leq u_{i_n^*-1}(X_{T-}) \leq a_{i_n^*-1} + \frac{2k_1 \ln^2 n}{\sqrt{n}} \leq \frac{1}{2^{1/d}}, \quad \text{on } E \cap \{T \leq n\},$$

for all $i \geq i_n^*$, if n is sufficiently large (using the fact that $(a_i)_{i=1}^{\infty}$ is decreasing, by Lemma 5.5). Similarly, let C' denote the event that T is an arrival where the customer selects from only the longest $\frac{k_2 \ln^2 n}{\sqrt{n}}$ of the queues, where $k_2 := 2k_1 + 1$. Then

$$A_i \cap E \cap \{T \leq n\} \subseteq C', \tag{7.4}$$

for all $i \geq i_n^*$, since

$$u_i(X_{T-}) \leq u_{i_n^*}(X_{T-}) \leq a_{i_n^*} + \frac{2k_1 \ln^2 n}{\sqrt{n}} \leq \frac{k_2 \ln^2 n}{\sqrt{n}}, \quad \text{on } E \cap \{T \leq n\}.$$

Now let us apply the inequalities (7.1)-(7.4). Taking probabilities in (7.2), and noting

that T is the first event time, we have

$$\begin{aligned}\nu_i &= \mathbb{P}\left(v\left(X_T\right) \geq i\right) \\ &\leq \mathbb{P}\left(\left\{v\left(X_{T-}\right) \geq i\right\} \cap A_{i-1} \cap E \cap \{T \leq n\}\right) \\ &\quad + \mathbb{P}\left(\left\{v\left(X_{T-}\right) = i-1\right\} \cap A_i \cap E \cap \{T \leq n\}\right) \\ &\quad + \mathbb{P}\left(\left\{v\left(X_{T-}\right) = i-1\right\} \cap M\right) + \mathbb{P}\left(\left\{v\left(X_{T-}\right) \geq i\right\} \cap D\right) + \mathbb{P}\left(\overline{E}\right) + \mathbb{P}\left(T > n\right).\end{aligned}$$

By (7.1), (7.3), (7.4) and the fact that T is an exponential random variable with rate $(\lambda+1)\,n$, we have

$$\begin{aligned}\nu_i \leq\ & \mathbb{P}\left(\left\{v\left(X_{T-}\right) \geq i\right\} \cap C\right) + \mathbb{P}\left(\left\{v\left(X_{T-}\right) \geq i-1\right\} \cap C'\right) \\ & + \mathbb{P}\left(\left\{v\left(X_{T-}\right) \geq i-1\right\} \cap M\right) + \mathbb{P}\left(\left\{v\left(X_{T-}\right) \geq i\right\} \cap D\right) + 2e^{-\eta \ln^2 n} + e^{-(\lambda+1)n^2},\end{aligned}$$

for all $i \geq i_n^*$, if n is sufficiently large. The events C, C', M and D are all independent from events of the form $\{v\left(X_{T-}\right) \geq k\}$, where $k \geq 0$. To see this inequality, for C say, we may use the Tower Rule to argue that

$$\begin{aligned}\mathbb{P}\left(\left\{v\left(X_{T-}\right) \geq k\right\} \cap C\right) &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{v(X_{T-})\geq k}\mathbf{1}_C \mid \mathcal{F}_{T-}\right]\right] \\ &= \mathbb{E}\left[\mathbf{1}_{v(X_{T-})\geq k}\mathbb{E}\left[\mathbf{1}_C \mid \mathcal{F}_{T-}\right]\right] = \nu_k \mathbb{P}\left(C\right).\end{aligned}$$

Let $k_3 := 2\left(k_2^d + d\right)$ and $\beta := \min\left(\eta, \lambda+1\right)$, then

$$\begin{aligned}\nu_i &\leq \frac{\lambda}{\lambda+1}\left[\nu_i\left(\frac{1}{2^{1/d}}\right)^d + \nu_{i-1}\left(\frac{k_2 \ln^2 n}{\sqrt{n}}\right)^d + \nu_{i-1}\frac{d}{n}\right] + \frac{1}{\lambda+1}\nu_i + \frac{1}{2}\frac{\lambda}{\lambda+1}e^{-\beta \ln^2 n} \\ &\leq \frac{\lambda}{\lambda+1}\left[\frac{1}{2}\nu_i + \frac{k_3}{2}\frac{\ln^{2d} n}{n^{\min(d/2,1)}}\nu_{i-1}\right] + \frac{1}{\lambda+1}\nu_i + \frac{1}{2}\frac{\lambda}{\lambda+1}e^{-\beta \ln^2 n},\end{aligned}$$

for all $i \geq i_n^*$, if n is sufficiently large. Hence

$$\nu_i \leq \frac{k_3 \ln^{2d} n}{n^{\min(d/2,1)}}\nu_{i-1} + e^{-\beta \ln^2 n},$$

for all $i \geq i_n^*$, if n is sufficiently large. By induction, we have

$$\nu_{i_n^*+z-1} \leq \left(\frac{k_3 \ln^{2d} n}{n^{\min(d/2,1)}}\right)^z \nu_{i_n^*-1} + z e^{-\beta \ln^2 n} \tag{7.5}$$

for all $z \geq 1$, if n is sufficiently large.

Now let us bound $\nu_{i_n^*-1}$. Let $k_4 > 1$ denote the constant given by Lemma 5.7, so that $a_{i-1}^\alpha \leq k_4 a_i$ for all $i \geq 1$. Then

$$b_{i_n^*-1} = \frac{a_{i_n^*-1}^d b_{i_n^*-2}}{1 - d\left(a_{i_n^*-2} - a_{i_n^*-1}\right) a_{i_n^*-1}^{d-1}} \leq 2a_{i_n^*-1}^d \leq 2\left(k_4 a_{i_n^*}\right)^{d/\alpha} \leq \frac{2k_4^{d/\alpha} \ln^{2d/\alpha} n}{n^{d/(2\alpha)}},$$

if n is sufficiently large. Let $k_5 > 0$ denote the constant given by Lemma 5.9, then let

$k_6 := 2k_4^{d/\alpha} + k_5$. This gives

$$\nu_{i_n^*-1} \leq b_{i_n^*-1} + \frac{k_5 \ln^2 n}{\sqrt{n}} \leq \frac{k_6 \ln^2 n}{n^{d/(2\alpha)}}, \tag{7.6}$$

if n is sufficiently large.

Hence, by (7.5) and (7.6), we have

$$\nu_{i_n^*+z-1} \leq \left(\frac{k_3 \ln^{2d} n}{n^{\min(d/2,1)}} \right)^z \frac{k_6 \ln^2 n}{n^{d/(2\alpha)}} + ze^{-\beta \ln^2 n},$$

if n is sufficiently large. The lemma easily follows. □

We will say that a customer is *new* if he/she arrived after time 0, and *initial* otherwise. The following lemma says that, with high probability, every customer in the system after a long period of time is new.

Lemma 7.2. *There exists $\eta > 0$ such that the following holds. Let $n \geq 1$, and let $\mathbf{X}$ be in equilibrium. For $t \geq 0$, let N_t denote the event that every customer at time t is new. Then*

$$\mathbb{P}\left(\overline{N_t}\right) \leq 3ne^{-\eta t}$$

for all $t \geq 0$.

Remark. This proof is essentially the same as the proof of Lemma 7.2 in [10], the analogous result for the standard supermarket model. The only difference is that here we use the results for the equilibrium distribution for the supermarket model with memory instead of Markov's inequality.

Proof. Since the left-hand side is bounded by 1 and $\eta > 0$ may be arbitrarily small, it suffices to show the result for all sufficiently large t.

Let $t \geq 2$, so that $k := \left\lfloor \frac{1}{2}t \right\rfloor \geq \frac{1}{4}t$. By Lemma 2.5, there exists $\eta_1 > 0$ such that

$$\begin{aligned} \mathbb{P}\left(\overline{N_t}\right) &\leq \mathbb{P}\left(\overline{N_t} \cap \{\|X_0\|_\infty \leq k\}\right) + \mathbb{P}\left(\|X_0\|_\infty > k\right) \\ &\leq n\mathbb{P}\left(\mathrm{Po}\left(t\right) \leq \tfrac{1}{2}t\right) + ne^{-\eta_1 k}. \end{aligned}$$

To see this inequality, note that if some queue still has an initial customer at time t, then this queue has at most k departures over $[0, t]$. There are n choices of such a queue, and the number of potential departures from any given queue in an interval of length t is $\mathrm{Po}\left(t\right)$. Let $\eta := \min\left(\frac{1}{12}, \frac{1}{4}\eta_1\right)$, then Lemma 1.4 gives

$$\mathbb{P}\left(\overline{N_t}\right) \leq 2ne^{-\frac{1}{12}t} + ne^{-\frac{1}{4}\eta_1 t} \leq 3ne^{-\eta t}.$$

□

Now we will bound the probability that the equilibrium maximum queue length is at most $i_n^* - 2$, and at least $i_n^* + z$ for all $z \geq 1$. This will immediately imply two-point concentration on the values $i_n^* - 1$ and i_n^*.

Lemma 7.3. *There exist $\eta > 0$ and $c_2 > 0$ such that the following holds. Let $n \geq 1$, and let X have the equilibrium distribution for the lengths process. Then*

$$\mathbb{P}\left(\|X\|_\infty \leq i_n^* - 2\right) \leq 2e^{-\eta \ln^2 n},$$
$$\mathbb{P}\left(\|X\|_\infty \geq i_n^* + z\right) \leq \left(\frac{c_2 \ln^{2d} n}{n^{\min(d/2,1)}}\right)^z \frac{\ln^{2d+4} n}{n^{d/2+d/(2\alpha)-1}},$$

for all $z \geq 1$.

Proof. Let $\mathbf{X}$ be in equilibrium, and let k_1 denote the constant given by Lemma 5.10. Let

$$E := \left\{\sup_{i\geq 1} |u_i(X_t) - a_i| < \frac{2k_1 \ln^2 n}{\sqrt{n}} \text{ for all } 0 \leq t \leq n^2\right\},$$

so that Lemma 5.10 (with $z = 2k_1$ and $r = 2$) gives $\eta > 0$ such that

$$\mathbb{P}\left(\overline{E}\right) \leq 2e^{-\eta \ln^2 n}. \tag{7.7}$$

For the first inequality, let us consider $\mathbf{X}$ at any time $0 \leq t \leq n^2$. Then

$$\begin{aligned}\mathbb{P}\left(\|X_t\|_\infty \leq i_n^* - 2\right) &\leq \mathbb{P}\left(u_{i_n^*-1}(X_t) = 0\right) \\ &\leq \mathbb{P}\left(\left|u_{i_n^*-1}(X_t) - a_{i_n^*-1}\right| \geq \frac{\ln^2 n}{\sqrt{n}}\right) \\ &\leq \mathbb{P}\left(\overline{E}\right) \leq 2e^{-\eta \ln^2 n}.\end{aligned}$$

For the second inequality, let us consider $\mathbf{X}$ at time $\ln^2 n$. First note that since the left-hand side is bounded by 1 and $c_2 > 0$ may be arbitrarily large, it suffices to show the result for all sufficiently large n.

Let N denote the event that every customer at time $\ln^2 n$ is new. By Lemma 7.2, there exists $\gamma > 0$ such that

$$\mathbb{P}\left(\overline{N}\right) \leq 3ne^{-\gamma \ln^2 n}. \tag{7.8}$$

Let $m := \lceil 2(\lambda+1) n \ln^2 n \rceil$. For $i, z \geq 1$, let $A_{i,z}$ denote the event that T_i is an arrival where the customer joins a queue of length $i_n^* + z - 1$ (hence making it a queue of length $i_n^* + z$). Then

$$\left\{\|X_{\ln^2 n}\|_\infty \geq i_n^* + z\right\} \cap N \cap \left\{T_{m+1} > \ln^2 n\right\} \subseteq \bigcup_{i=1}^m A_{i,z}. \tag{7.9}$$

To see this inequality, note that if at time $\ln^2 n$, there is a queue of length at least $i_n^* + z$ consisting entirely of new customers, and if there have been at most m events by time $\ln^2 n$, then at least one of $T_1, \ldots, T_m$ is an arrival where the customer joins a queue of length $i_n^* + z$. Thus (7.9) holds.

For $i \geq 1$, let B_i denote the event that T_i is an arrival where the customer selects only queues of length at least i_n^*. Then

$$A_{i,z} \subseteq \left\{v(X_{T_i-}) \geq i_n^* + z - 1\right\} \cap B_i, \tag{7.10}$$

for all $i, z \geq 1$. To see this inequality, note that a necessary condition for $A_{i,z}$ is that the candidates list only contains queues of length at least $i_n^* + z - 1$.

For $i \geq 1$, let C_i denote the event that T_i is an arrival where the customer selects only queues from the longest $\frac{k_2 \ln^2 n}{\sqrt{n}}$ of the queues, where $k_2 := 2k_1 + 1$. Then

$$B_i \cap E \cap \left\{T_i \leq n^2\right\} \subseteq C_i \tag{7.11}$$

for all $i \geq 1$. To see this inequality, note that the proportion of queues of length at least i_n^* immediately before T_i satisfies

$$u_{i_n^*}\left(X_{T_i-}\right) \leq a_{i_n^*} + \frac{2k_1 \ln^2 n}{\sqrt{n}} \leq \frac{k_2 \ln^2 n}{\sqrt{n}}, \quad \text{on } E \cap \left\{T_i \leq n^2\right\}.$$

Now let us apply the inequalities (7.7)-(7.11). Taking probabilities in (7.9), we have

$$\begin{aligned}\mathbb{P}\left(\left\|X_{\ln^2 n}\right\|_\infty \geq i_n^* + z\right) \leq{}& \mathbb{P}\left(\left\{\left\|X_{\ln^2 n}\right\|_\infty \geq i_n^* + z\right\} \cap E \cap N \cap \left\{\ln^2 n < T_{m+1} \leq n^2\right\}\right) \\ &+ \mathbb{P}\left(\overline{E}\right) + \mathbb{P}\left(\overline{N}\right) + \mathbb{P}\left(T_{m+1} \leq \ln^2 n\right) + \mathbb{P}\left(T_{m+1} > n^2\right),\end{aligned}$$

for all $z \geq 1$. By (7.7)-(7.9) and the fact that the number of events in $[0, t]$ is $\mathrm{Po}\left((\lambda+1)nt\right)$, we have

$$\begin{aligned}\mathbb{P}\left(\left\|X_{\ln^2 n}\right\|_\infty \geq i_n^* + z\right) \leq{}& \mathbb{P}\left(\bigcup_{i=1}^{m}\left(A_{i,z} \cap E \cap \left\{T_i \leq n^2\right\}\right)\right) + 2e^{-\eta \ln^2 n} + 3ne^{-\gamma \ln^2 n} \\ &+ \mathbb{P}\left(\mathrm{Po}\left((\lambda+1)\, n \ln^2 n\right) > m\right) + \mathbb{P}\left(\mathrm{Po}\left((\lambda+1)\, n^3\right) \leq m\right),\end{aligned}$$

for all $z \geq 1$. By (7.10) and Lemma 1.4 (once with $(\varepsilon, \mu) = \left(1, (\lambda+1)\, n \ln^2 n\right)$ and once with $(\varepsilon, \mu) = \left(\frac{1}{2}, (\lambda+1)\, n^3\right)$), we have

$$\begin{aligned}\mathbb{P}\left(\left\|X_{\ln^2 n}\right\|_\infty \geq i_n^* + z\right) \leq{}& \sum_{i=1}^{m} \mathbb{P}\left(\left\{v\left(X_{T_i-}\right) \geq i_n^* + z - 1\right\} \cap B_i \cap E \cap \left\{T_i \leq n^2\right\}\right) \\ &+ 2e^{-\eta \ln^2 n} + 3ne^{-\gamma \ln^2 n} + 2e^{-\frac{2}{3}(\lambda+1) n \ln^2 n} + 2e^{-\frac{1}{12}(\lambda+1) n^3},\end{aligned}$$

for all $z \geq 1$, if n is sufficiently large. Let $\beta := \min\left(\eta, \gamma, \frac{1}{12}(\lambda+1)\right)$, then (7.11) gives

$$\mathbb{P}\left(\left\|X_{\ln^2 n}\right\|_\infty \geq i_n^* + z\right) \leq \sum_{i=1}^{m} \mathbb{P}\left(\left\{v\left(X_{T_i-}\right) \geq i_n^* + z - 1\right\} \cap C_i\right) + 9ne^{-\beta \ln^2 n},$$

for all $z \geq 1$, if n is sufficiently large. The event C_i is independent from $\left\{v\left(X_{T_i-}\right) \geq i_n^* + z - 1\right\}$, since we may use the Tower Rule to argue that

$$\begin{aligned}\mathbb{P}\left(\left\{v\left(X_{T_i-}\right) \geq i_n^* + z - 1\right\} \cap C_i\right) &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{v\left(X_{T_i-}\right) \geq i_n^* + z - 1} \mathbf{1}_{C_i} \mid \mathcal{F}_{T_i-}\right]\right] \\ &= \mathbb{E}\left[\mathbf{1}_{v\left(X_{T_i-}\right) \geq i_n^* + z - 1} \mathbb{E}\left[\mathbf{1}_{C_i} \mid \mathcal{F}_{T_i-}\right]\right] = \nu_{i_n^* + z - 1} \mathbb{P}\left(C_i\right).\end{aligned}$$

By Lemma 7.1, there exists $c_1 > 0$ such that

$$\begin{aligned}\mathbb{P}\left(\|X_{\ln^2 n}\|_\infty \geq i_n^* + z\right) &\leq m\nu_{i_n^*+z-1}\left(\frac{k_2 \ln^2 n}{\sqrt{n}}\right)^d + 9ne^{-\beta \ln^2 n} \\ &\leq 8n \ln^2 n \left[\left(\frac{c_1 \ln^{2d} n}{n^{\min(d/2,1)}}\right)^z \frac{\ln^2 n}{n^{d/(2\alpha)}}\right] \frac{k_2 \ln^{2d} n}{n^{d/2}} + 9ne^{-\beta \ln^2 n},\end{aligned}$$

for all $z \geq 1$, if n is sufficiently large. The second inequality in the statement of the lemma easily follows. □

Theorem 1.2 then follows from Lemma 7.3.

Proof of Theorem 1.2. Since the left-hand side is bounded by 1 and $c > 0$ may be arbitrarily large, it suffices to show the result for all sufficiently large n.

Let $\eta > 0$ and $c_2 > 0$ denote the constants given by Lemma 7.3. Taking $z = 1$ gives

$$\begin{aligned}\mathbb{P}\left(\|X\|_\infty \neq i_n^* - 1 \text{ or } i_n^*\right) &\leq \mathbb{P}\left(\|X\|_\infty \leq i_n^* - 2\right) + \mathbb{P}\left(\|X\|_\infty \geq i_n^* + 1\right) \\ &\leq 2e^{-\eta \ln^2 n} + \left(\frac{c_2 \ln^{2d} n}{n^{\min(d/2,1)}}\right) \frac{\ln^{2d+4}}{n^{d/2+d/(2\alpha)-1}},\end{aligned}$$

if n is sufficiently large. The theorem easily follows. □

7.2 Long-term behaviour

In this section, we will show that the equilibrium maximum queue length is concentrated around $\frac{\ln \ln n}{\ln \alpha}$ for long periods of time. We will show that $i_n^* = \frac{\ln \ln n}{\ln \alpha} + O(1)$, as claimed in Section 1.2. The calculation is given in full detail for the sake of completeness, but is routine and easy.

Lemma 7.4. *There exist $c_3 > 0$ and $n^* \geq 1$ such that the following holds. Let $n \geq n^*$, then*

$$\left| i_n^* - \frac{\ln \ln n}{\ln \alpha} \right| \leq c_3.$$

Proof. Throughout this proof, we will write $i = i_n^*$ for brevity. By Lemma 5.8, there exist $0 < \sigma < \tau < 1$ such that

$$\sigma^{\alpha^i} < a_i \leq \frac{\ln^2 n}{\sqrt{n}} < a_{i-1} < \tau^{\alpha^{i-1}}. \tag{7.12}$$

We will assume, without loss of generality, that $0 < \sigma < 1$ is sufficiently small so that $\ln \ln \sigma^{-4} > 0$, and that $0 < \tau < 1$ is sufficiently large so that $\ln \ln \tau^{-2/\alpha} < 0$. Thus

$$c_3 := \frac{\max\left(\ln \ln \sigma^{-4}, -\ln \ln \tau^{-2/\alpha}\right)}{\ln \alpha} > 0.$$

Now, taking logarithms in (7.12) gives

$$\alpha^i \ln \sigma < 2 \ln \ln n - \tfrac{1}{2} \ln n < \alpha^{i-1} \ln \tau,$$

and thus

$$\alpha^i \ln \sigma < -\tfrac{1}{4} \ln n, \quad \alpha^{i-1} \ln \tau > -\tfrac{1}{2} \ln n,$$

if n^* is sufficiently large. Rearranging gives

$$\ln \tau^{-2/\alpha} < \frac{\ln n}{\alpha^i} < \ln \sigma^{-4},$$

and thus

$$-c_3 \ln \alpha \leq \ln \ln \tau^{-2/\alpha} < \ln \ln n - i \ln \alpha < \ln \ln \sigma^{-4} \leq c_3 \ln \alpha,$$

if n^* is sufficiently large. The result follows. □

Now we will need the following result which extends bounds on the maximum queue length from instants to polynomial periods of time.

Lemma 7.5. *Let $n \geq 1$, and let $\mathbf{X}$ be in equilibrium. Then*

$$\mathbb{P}\left(\bigcup_{0 \leq t \leq n^r} \{ \|X_t\|_\infty \leq k \} \right) \leq 3n^{r+2} \left(\mathbb{P}\left(\|X\|_\infty \leq k + l \right) + \frac{1}{n^l} \right), \tag{7.13}$$

$$\mathbb{P}\left(\bigcup_{0 \leq t \leq n^r} \{ \|X_t\|_\infty \geq k + l \} \right) \leq 3n^{r+2} \left(\mathbb{P}\left(\|X\|_\infty \geq k \right) + \frac{1}{n^l} \right), \tag{7.14}$$

for all $k, l \geq 0$ and $r > 0$.

Remark. This proof is similar to the proof of Lemma 7.1 in [10], the analogous result for the standard supermarket model. The only difference is that here we use potential departures instead of arrivals.

Proof. For $t, h \geq 0$, let

$$M_{t,h} := \{ \|X_t\|_\infty \leq h \}, \quad L_{t,h} := \{ \|X_t\|_\infty \geq h \}.$$

Consider covering the interval $[0, n^r]$ with sub-intervals of length $\delta = \frac{1}{n}$; clearly $m := \lceil n^{r+1} \rceil$ such sub-intervals will cover $[0, n^r]$. For $i \geq 0$, let $t_i := i\delta$, then

$$\mathbb{P}\left(\bigcup_{0 \leq t \leq n^r} M_{t,k} \right) \leq \sum_{i=0}^{m} \mathbb{P}\left(M_{t_i, k+l} \right) + mn\mathbb{P}\left(\mathrm{Po}\left(\delta \right) > l \right).$$

To see the last term in this inequality, suppose that $\overline{M_{t_i,k+l}}$ holds for all end-points t_i. Then there exists a sub-interval $\mathcal{I}_r := [t_{r-1}, t_r)$ containing t. Since $\overline{M_{t_{r-1},k+l}}$ and $M_{t,k}$ hold, we deduce that over $\mathcal{I}_r$ the maximum queue length decreases by more than l, and thus over $\mathcal{I}_r$, we have more than l potential departures from a specified queue (with n choices for such a queue). However, the number of potential departures from a specified queue over $\mathcal{I}_r$, an interval of length δ, is Poisson with mean $n \cdot \frac{1}{n} \cdot \delta = \delta$. Now

$$\mathbb{P}\left(\mathrm{Po}\left(\delta \right) > l \right) = \sum_{i=l+1}^{\infty} \frac{e^{-\delta} \delta^i}{i!} \leq \sum_{i=l+1}^{\infty} \delta^i = \frac{\delta^{l+1}}{1-\delta} = \frac{1}{n^{l+1}\left(1 - \frac{1}{n}\right)} \leq \frac{1}{n^l},$$

so

$$\mathbb{P}\left(\bigcup_{0\le t\le n^r} M_{t,k}\right) \le n(m+1)\left(\mathbb{P}\left(\|X\|_\infty \le k+l\right) + \frac{1}{n^l}\right)$$
$$\le 3n^{r+2}\left(\mathbb{P}\left(\|X\|_\infty \le k+l\right) + \frac{1}{n^l}\right).$$

Similarly, we may write

$$\mathbb{P}\left(\bigcup_{0\le t\le n^r} L_{t,k+l}\right) \le \sum_{i=0}^{m} \mathbb{P}\left(L_{t_i,k}\right) + mn\mathbb{P}\left(\mathrm{Po}\left(\delta\right) > l\right).$$

By arguing as above, the result follows. □

Now we will show that the maximum queue length is concentrated around $\frac{\ln\ln n}{\ln\alpha}$ for long periods of time.

Theorem 7.6. *Let $r > 0$. Then there exists $c = c(r) > 0$ such that the following holds. Let $n \ge 1$, and let $\mathbf{X}$ be in equilibrium. Then*

$$\mathbb{P}\left(\left|\|X_t\|_\infty - \frac{\ln\ln n}{\ln\alpha}\right| > c \text{ for some } 0 \le t \le n^r\right) \le \frac{c}{n^r}.$$

Proof. Since the left-hand side is bounded by 1 and $c > 0$ may be arbitrarily large, it suffices to show the result for all sufficiently large n.

Let $c_3 > 0$ denote the constant given by Lemma 7.4, so that

$$\left|i_n^* - \frac{\ln\ln n}{\ln\alpha}\right| \le c_3,$$

if n is sufficiently large. Let $\delta := \frac{1}{2}\min\left(\frac{1}{2}d, 1\right)$, then let

$$z = z(r) := \left\lceil \frac{2r+3}{\delta} \right\rceil + 1, \quad c = c(r) := c_3 + 2z + 2.$$

For brevity, let

$$E_r := \left\{\left|\|X_t\|_\infty - \frac{\ln\ln n}{\ln\alpha}\right| > c \text{ for some } 0 \le t \le n^r\right\}.$$

If E_r holds, then there exists $0 \le t \le n^r$ such that

$$\|X_t\|_\infty < \frac{\ln\ln n}{\ln\alpha} - c = \frac{\ln\ln n}{\ln\alpha} - c_3 - 2z - 2 \le i_n^* - z - 2,$$

or

$$\|X_t\|_\infty > \frac{\ln\ln n}{\ln\alpha} + c = \frac{\ln\ln n}{\ln\alpha} + c_3 + 2z + 2 \ge i_n^* + 2z.$$

Hence, by (7.13) (with $k = i_n^* - z - 2$ and $l = z$) and (7.14) (with $k = i_n^* + z$ and $l = z$),

we have

$$\begin{aligned}\mathbb{P}\left(E_r\right) &\leq \mathbb{P}\left(\bigcup_{0\leq t\leq n^r}\left\{\|X_t\|_\infty \leq i_n^* - z - 2\right\}\right) + \mathbb{P}\left(\bigcup_{0\leq t\leq n^r}\left\{\|X_t\|_\infty \geq i_n^* + 2z\right\}\right)\\ &\leq 3n^{r+2}\left(\mathbb{P}\left(\|X\|_\infty \leq i_n^* - 2\right) + \frac{1}{n^z}\right) + 3n^{r+2}\left(\mathbb{P}\left(\|X\|_\infty \geq i_n^* + z\right) + \frac{1}{n^z}\right).\end{aligned}$$

Let $\eta > 0$ and $c_2 > 0$ denote the constants given by Lemma 7.3. Since $z \geq 2$, we may use (1.4) to write

$$\mathbb{P}\left(\|X\|_\infty \geq i_n^* + z\right) \leq \left(\frac{c_2 \ln^{2d} n}{n^{2\delta}}\right)^{z-1} \frac{c_2 \ln^{4d+4} n}{n^{2\delta + d/2 + d/(2\alpha) - 1}} \leq \left(\frac{1}{n^\delta}\right)^{z-1} \frac{c_2 \ln^{4d+4} n}{n^{d/(2\alpha)}} \leq \frac{1}{n^{\delta(z-1)}},$$

if n is sufficiently large. Hence

$$\begin{aligned}\mathbb{P}\left(E_r\right) &\leq 3n^{r+2}\left(2e^{-\eta \ln^2 n} + \frac{1}{n^z}\right) + 3n^{r+2}\left(\frac{1}{n^{\delta(z-1)}} + \frac{1}{n^z}\right).\\ &\leq 3n^{r+2}\left(\frac{4}{n^{\delta(z-1)}}\right) \leq \frac{12n^{r+2}}{n^{2r+3}} \leq \frac{1}{n^r},\end{aligned}$$

if n is sufficiently large. □

Chapter 8

Further ideas

We have investigated the supermarket model with memory. We have shown that the system is rapidly mixing. That is, with reasonable initial conditions, the convergence to equilibrium is very fast. We have also shown that with probability tending to 1 as $n \to \infty$, the maximum queue length in equilibrium is concentrated on two consecutive values which are $\frac{\ln \ln n}{\ln \alpha} + O(1)$, where $\alpha := d + \frac{1}{2} + \sqrt{d^2 + \frac{1}{4}}$.

A desired result, which has an analogue for the standard supermarket model in [10] by Luczak and McDiarmid, is that the upper bounds on the mixing times in Theorem 1.1 are of the right order. That is, we wish to show that there exists $c > 0$ such that if $t \leq c \ln n$, then

$$d_{\mathrm{TV}}\left(\mathcal{L}\left(X_t\right), \Pi\right) = 1 - e^{-\Omega\left(\ln^2 n\right)}.$$

The analogous result is shown by using the fact that the proportion of non-empty queues in a system is close to its mean. For a standard lengths process in equilibrium, this mean is λ, whilst for a standard lengths process started from the empty state $\mathbf{0} \in \mathbb{Z}_+^n$, this mean is at most $\lambda - c$ at a time $t \leq c \ln n$. Hence, the two distributions are far apart at such a time. Our analysis would take us through (5.7) and (5.8), but we have been unable to complete this.

A further line of investigation could be the generalised supermarket model with memory where the memory saves a set of $m \geq 1$ queues, rather than just $m = 1$. It is unclear which parts of our arguments will easily extend to such a model, though we suspect that the random walk lemmas of Section 2.3 will need generalising.

Bibliography

[1] Graham Brightwell and Malwina Luczak. The supermarket model with arrival rate tending to one. January 2012.

[2] Russ Bubley and Martin Dyer. Path coupling: a technique for proving rapid mixing in Markov chains. In *Proceedings of the 38th Annual Symposium on Foundations of Computer Science*, page 223, 1997.

[3] Michael Dahlin. Interpreting stale load information. *IEEE Transactions on Parallel and Distributed Systems*, 11(10):1033–1047, October 2000.

[4] Derek L. Eager, Edward D. Lazowska, and John Zahorjan. A comparison of receiver-initiated and sender-initiated adaptive load sharing (extended abstract). *ACM*, 13(2):1–3, August 1985.

[5] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterisation and Convergence*. Wiley, 1986.

[6] Marianne Fairthorne. *The supermarket model with system-size dependent parameters*. PhD thesis, London School of Economics and Political Science, June 2011.

[7] Carl Graham. Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journal of Applied Probability*, 37:198–211, 2000.

[8] Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001.

[9] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematics Society, 2009.

[10] Malwina J. Luczak and Colin McDiarmid. On the maximum queue length in the supermarket model. *The Annals of Probability*, 34(2):493–527, 2006.

[11] Malwina J. Luczak and Colin McDiarmid. Asymptotic distributions and chaos for the supermarket model. *EJP*, 12:75–99, December 2007.

[12] Malwina J. Luczak and James Norris. Strong approximation for the supermarket model. May 2004.

[13] M.J. Luczak and J.R. Norris. Averaging over fast variables in the fluid limit for Markov chains: application to the supermarket model with memory. *The Annals of Applied Probability (to appear)*, January 2010.

[14] J. B. Martin and Yu. M. Suhov. Fast Jackson networks. *The Annals of Applied Probability*, 9(3):854–870, 1999.

[15] Colin McDiarmid. Concentration. *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248, 1998.

[16] Ravi Mirchandaney, Don Towsley, and John A. Stankovic. Analysis of the effects of delays on load sharing. *IEEE Transactions on Computers*, 38(11):1513–1525, November 1989.

[17] Michael Mitzenmacher. Load balancing and density dependent jump Markov processes (extended abstract). In *Proceedings of the 37th Annual Symposium on Foundations of Computer Science*, page 213, 1996.

[18] Michael Mitzenmacher. On the analysis of randomized load balancing schemes. In *Proceedings of the 9th Annual ACM symposium on Parallel Algorithms and Architectures*, pages 292–301, 1997.

[19] Michael Mitzenmacher. Analyses of load stealing models based on differential equations. In *Proceedings of the 10th Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 212–221, 1998.

[20] Michael Mitzenmacher. How useful is old information? *IEEE Transactions on Parallel and Distributed Systems*, 11(1):6–20, January 2000.

[21] Michael Mitzenmacher, Balaji Prabhakar, and Devavrat Shah. Load balancing with memory. In *Proceedings of the 43rd Symposium on Foundations of Computer Science*, pages 799–808, 2002.

[22] Michael Mitzenmacher, Andréa W. Richa, and Ramesh Sitaraman. *The Power of Two Random Choices: A Survey of Techniques and Results*, volume 1 of *Handbook of Randomized Computing*, chapter 9, pages 255–312. Springer, 2001.

[23] Michael David Mitzenmacher. *The Power of Two Choices in Randomized Load Balancing*. PhD thesis, University of California at Berkeley, 1996.

[24] J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[25] Devavrat Shah and Balaji Prabhakar. The use of memory in randomized load balancing. *International Symposium on Information Theory*, page 125, 2002.

[26] Berthold Vöcking. How asymmetry helps load balancing. *Journal of the ACM*, 50(4):568–589, July 2003.

[27] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Probl. Peredachi Inf.*, 32:20–34, 1996.

[28] Nikita D. Vvedenskaya and Yuri M. Suhov. Dobrushin's mean-field approximation for a queue with dynamic routing. *Markov Processes and Related Fields*, 3(4):493–526, 1997.