The Annals of Applied Probability
2013, Vol. 23, No. 3, 957–986
DOI: 10.1214/12-AAP861

AVERAGING OVER FAST VARIABLES IN THE FLUID LIMIT FOR MARKOV CHAINS: APPLICATION TO THE SUPERMARKET MODEL WITH MEMORY

BY M. J. LUCZAK[1] AND J. R. NORRIS[2]

London School of Economics and University of Cambridge

We set out a general procedure which allows the approximation of certain Markov chains by the solutions of differential equations. The chains considered have some components which oscillate rapidly and randomly, while others are close to deterministic. The limiting dynamics are obtained by averaging the drift of the latter with respect to a local equilibrium distribution of the former. Some general estimates are proved under a uniform mixing condition on the fast variable which give explicit error probabilities for the fluid approximation. Mitzenmacher, Prabhakar and Shah [In *Proc. 43rd Ann. Symp. Found. Comp. Sci.* (2002) 799–808, IEEE] introduced a variant with memory of the "join the shortest queue" or "supermarket" model, and obtained a limit picture for the case of a stable system in which the number of queues and the total arrival rate are large. In this limit, the empirical distribution of queue sizes satisfies a differential equation, while the memory of the system oscillates rapidly and randomly. We illustrate our general fluid limit estimate by giving a proof of this limit picture.

1. A general fluid limit estimate. We describe a general framework to allow the incorporation of averaging over fast variables into fluid limit estimates for Markov chains, building on the approach used in [2]. The main results of this section, Theorems 1.5 and 1.6, establish explicit error probabilities for the fluid approximation under assumptions which can be verified from knowledge of the transition rates of the Markov chain. Also see [1] for related results.

1.1. *Outline of the method.* Let $X = (X_t)_{t\geq 0}$ be a continuous-time Markov chain with countable state-space S and with generator matrix $Q = (q(\xi, \xi') : \xi, \xi' \in S)$. Assume that the total jump rate $q(\xi)$ is finite for all states ξ, and that X is nonexplosive. Then the law of X is determined uniquely by Q and the law of X_0. Make a choice of *fluid coordinates* $x^i : S \to \mathbb{R}$, for $i = 1, \dots, d$, and write $\mathbf{x} = (x^1, \dots, x^d) : S \to \mathbb{R}^d$. Consider the $\mathbb{R}^d$-valued process $\mathbf{X} = (\mathbf{X}_t)_{t\geq 0}$ given by

Received March 2010; revised August 2011.

[1]Supported in part by a STICERD grant at the LSE, and by EPSRC Leadership Fellowship EP/J004022/2. Part of this work was accomplished while visiting the Mittag-Leffler Institute.

[2]Supported by EPSRC Grant EP/E01772X/1.

MSC2010 subject classifications. Primary 60J28; secondary 60K25.

Key words and phrases. Join the shortest queue, supermarket model, supermarket model with memory, law of large numbers, exponential martingale inequalities, fast variables, correctors.

$\mathbf{X}_t = (X_t^1, \dots, X_t^d) = \mathbf{x}(X_t)$. Call $\mathbf{X}$ the *slow* or *fluid variable*. Define for each $\xi \in S$ the *drift vector*

$$\beta(\xi) = Q\mathbf{x}(\xi) = \sum_{\xi' \neq \xi} \big(\mathbf{x}(\xi') - \mathbf{x}(\xi)\big) q(\xi, \xi').$$

Also, make a choice of an *auxiliary coordinate* $y : S \to I$, for some countable set I, and set $Y_t = y(X_t)$. Call the process $Y = (Y_t)_{t \geq 0}$ the *fast variable*. For $\xi \in S$ and $y' \in I$ with $y' \neq y(\xi)$, write $\gamma(\xi, y')$ for the total rate at which Y jumps to y' when X is at ξ. Thus

$$\gamma(\xi, y') = \sum_{\xi' : y(\xi') = y'} q(\xi, \xi').$$

Choose a subset U of $\mathbb{R}^d$ and a function $b : U \times I \to \mathbb{R}^d$. Choose also, for each $x \in U$, a generator matrix $G_x = (g(x, y, y') : y, y' \in I)$ having a unique invariant distribution $\pi_x = (\pi(x, y) : y \in I)$. These choices are to be made so that $\beta(\xi)$ is close to $b(\mathbf{x}(\xi), y(\xi))$ and $\gamma(\xi, y')$ is close to $g(\mathbf{x}(\xi), y(\xi), y')$ whenever $\mathbf{x}(\xi) \in U$ and $y' \in I$. Define for $x \in U$

$$\bar{b}(x) = \sum_{y \in I} b(x, y) \pi(x, y).$$

Then, under regularity assumptions to be specified later, there exists a function $\chi : U \times I \to \mathbb{R}^d$ such that

$$G\chi(x, y) = \sum_{y' \in I} g(x, y, y') \chi(x, y') = b(x, y) - \bar{b}(x). \tag{1}$$

Make a choice of such a function χ. Call χ the *corrector for b*.

Fix $x_0 \in U$. We will assume that $\bar{b}$ is Lipschitz on U. Then the differential equation $\dot{x}_t = \bar{b}(x_t)$ has a unique maximal solution $(x_t)_{t<\zeta}$ in U starting from x_0. Fix $t_0 \in [0, \zeta)$. Then for $t \leq t_0$,

$$x_t = x_0 + \int_0^t \bar{b}(x_s)\, ds. \tag{2}$$

Define for $\xi \in S$ with $\mathbf{x}(\xi) \in U$

$$\bar{\mathbf{x}}(\xi) = \mathbf{x}(\xi) - \chi(\mathbf{x}(\xi), y(\xi)).$$

Let T be a stopping time such that $\mathbf{X}_t \in U$ for all $t \leq T$. Then, under regularity assumptions to be specified later, for $t \leq T$,

$$\bar{\mathbf{x}}(X_t) = \bar{\mathbf{x}}(X_0) + M_t + \int_0^t \bar{\beta}(X_s)\, ds, \tag{3}$$

where $M = M^{\bar{\mathbf{x}}}$ is a martingale and where

$$\bar{\beta} = Q\bar{\mathbf{x}} = \beta - Q(\chi(\mathbf{x}, y)). \tag{4}$$

On subtracting equations (2) and (3) we obtain for $t \le T \wedge t_0$

$$\begin{aligned} \mathbf{X}_t - x_t &= \mathbf{X}_0 - x_0 + \chi(\mathbf{X}_t, Y_t) - \chi(\mathbf{X}_0, Y_0) + M_t + \int_0^t \Delta(X_s)\,ds \\ &\quad + \int_0^t \big(\beta(X_s) - b(\mathbf{X}_s, Y_s)\big)\,ds + \int_0^t \big(\bar{b}(\mathbf{X}_s) - \bar{b}(x_s)\big)\,ds, \end{aligned} \tag{5}$$

where $\Delta = G\chi(\mathbf{x}, y) - Q(\chi(\mathbf{x}, y))$.

The discussion in the present paragraph is intended for orientation only, and will play no essential role in the derivation of our results. Fix $U_0 \subseteq U$ such that for all $\xi, \xi' \in S$ with $\mathbf{x}(\xi) \in U_0$ and $q(\xi, \xi') > 0$ we have $\mathbf{x}(\xi') \in U$. Assume that T is chosen so that $\mathbf{X}_t \in U_0$ for all $t \le T$. Define for $\xi \in S$ with $\mathbf{x}(\xi) \in U_0$ the *diffusivity tensor* $\alpha(\xi) \in \mathbb{R}^d \otimes \mathbb{R}^d$ by

$$\alpha^{ij}(\xi) = \sum_{\xi' \neq \xi} \big(\bar{\mathbf{x}}^i(\xi') - \bar{\mathbf{x}}^i(\xi)\big)\big(\bar{\mathbf{x}}^j(\xi') - \bar{\mathbf{x}}^j(\xi)\big) q(\xi, \xi') \tag{6}$$

and define for $t \le T$

$$N_t = M_t \otimes M_t - \int_0^t \alpha(X_s)\,ds.$$

Then, under regularity assumptions, N is a martingale in $\mathbb{R}^d \otimes \mathbb{R}^d$. Choose a function $a : U_0 \times I \to \mathbb{R}^d \otimes \mathbb{R}^d$ and set

$$\bar{a}(x) = \sum_{y \in I} a(x, y)\pi(x, y).$$

This choice is to be made so that $\alpha(\xi)$ is close to $a(\mathbf{x}(\xi), y(\xi))$ whenever $\mathbf{x}(\xi) \in U_0$. Suppose we can also find a corrector for a, that is, a function $\tilde{\chi} : U_0 \times I \to \mathbb{R}^d \otimes \mathbb{R}^d$ such that

$$G\tilde{\chi}(x, y) = a(x, y) - \bar{a}(x). \tag{7}$$

Then, for $t \le T$,

$$\begin{aligned} \int_0^t \alpha(X_s)\,ds &= \tilde{\chi}(\mathbf{X}_t, Y_t) - \tilde{\chi}(\mathbf{X}_0, Y_0) - \tilde{M}_t + \int_0^t \tilde{\Delta}(X_s)\,ds \\ &\quad + \int_0^t \big(\alpha(X_s) - a(\mathbf{X}_s, Y_s)\big)\,ds + \int_0^t \bar{a}(\mathbf{X}_s)\,ds, \end{aligned} \tag{8}$$

where $\tilde{\Delta} = G\tilde{\chi}(\mathbf{x}, y) - Q(\tilde{\chi}(\mathbf{x}, y))$ and, under suitable regularity conditions, $\tilde{M} = M^{\tilde{\chi}}$ is a martingale up to T.

The martingale terms M and $\tilde{M}$ in (5) and (8) can be shown to be small, under suitable conditions, using the following standard type of exponential martingale inequality. In the form given here it may be deduced, for example, from [2], Proposition 8.8, by setting $f = \theta\phi$, $A = \theta^2 e^{\theta J}\varepsilon/2$ and $B = \theta\delta$.

PROPOSITION 1.1. *Let ϕ be a function on S. Define*

$$M_t = M_t^\phi = \phi(X_t) - \phi(X_0) - \int_0^t Q\phi(X_s)\,ds.$$

Write $J = J(\phi)$ for the maximum possible jump in $\phi(X)$, thus

$$J = \sup_{\xi,\xi'\in S, q(\xi,\xi')>0} |\phi(\xi') - \phi(\xi)|.$$

Define a function $\alpha = \alpha^\phi$ on S by

$$\alpha(\xi) = \sum_{\xi'\neq\xi} \{\phi(\xi') - \phi(\xi)\}^2 q(\xi,\xi').$$

Then, for all $\delta, \varepsilon \in (0,\infty)$ and all stopping times T, we have

$$\mathbb{P}\left(\sup_{t\le T} M_t \ge \delta \text{ and } \int_0^T \alpha(X_t)\,dt \le \varepsilon\right) \le \exp\{-\delta^2/(2\varepsilon e^{\theta J})\},$$

where $\theta \in (0,\infty)$ is determined by $\theta e^{\theta J} = \delta/\varepsilon$.

Now, if β, γ, α are well approximated by b, g, a and if we can show that the corrector terms in (5) and (8) are insignificant, then we may hope to use these equations to show that the path $(x_t : t \le t_0)$ provides a good (first order) approximation to $(\mathbf{X}_t : t \le t_0)$ and, moreover, that the fluctuation process $(\mathbf{X}_t - x_t : t \le t_0)$ is approximated (to second order) by a Gaussian process $(F_t : t \le t_0)$ given by

$$F_t = F_0 + B_t + \int_0^t \nabla\bar{b}(x_s) F_s\,ds,$$

where $(B_t : t \le t_0)$ is a zero-mean Gaussian process in $\mathbb{R}^d$ with covariance

$$\mathbb{E}(B_s \otimes B_t) = \int_0^{s\wedge t} \bar{a}(x_r)\,dr.$$

Our aim in the rest of this section is to give an explicit form of the first order approximation with optimal error scale, that is, of the same order as the scale of deviation predicted by the second order approximation. The next subsection contains some preparatory material on correctors. A reader who wishes to understand only the statement of the fluid limit estimate can skip directly to Section 1.3.

1.2. *Correctors.* In order to implement the method just outlined, it is necessary either to come up with explicit correctors or to appeal to a general result which guarantees the existence, subject to verifiable conditions, of correctors with good properties. In this subsection we obtain such a general result. In fact, we shall find conditions which guarantee the existence, for each bounded measurable function f on $U \times I$, of a good *corrector for* f, that is to say, a function $\chi = \chi_f$ on $U \times I$ such that

$$G\chi(x,y) = f(x,y) - \bar{f}(x),$$

where

$$\bar{f}(x) = \sum_{y \in I} f(x, y)\pi(x, y).$$

Moreover, we shall see that χ_f depends linearly on f and we shall obtain a uniform bound and a continuity estimate for χ_f.

Assume that there is a constant $\nu \in (0, \infty)$ such that, for all $x \in U$ and all $y \in I$, the total rate of jumping from y under G_x does not exceed ν. Then we can choose an auxiliary measurable space E, with a σ-field $\mathcal{E}$, a family of probability measures $\mu = (\mu_x : x \in U)$ on $(E, \mathcal{E})$ and a measurable function $F : I \times E \to I$ such that, for all $x \in U$ and all $y, y' \in I$ distinct,

$$g(x, y, y') = \nu\mu_x\bigl(\{v \in E : F(y, v) = y'\}\bigr). \tag{9}$$

Let $N = (N(t) : t \geq 0)$ be a Poisson process of rate ν. Fix $x \in U$ and let $V = (V_n : n \in \mathbb{N})$ be a sequence of independent random variables in E, all with law μ_x. Thus

$$g(x, y, y') = \nu\mathbb{P}\bigl(F(y, V_n) = y'\bigr)$$

for all pairs of distinct states y, y' and all n. Fix a reference state $\bar{y} \in I$. Given $y \in I$, set $Z_0 = y$ and $\bar{Z}_0 = \bar{y}$ and define recursively for $n \geq 0$,

$$Z_{n+1} = F(Z_n, V_{n+1}), \qquad \bar{Z}_{n+1} = F(\bar{Z}_n, V_{n+1}).$$

Set $Y_t = Z_{N(t)}$ and $\bar{Y}_t = \bar{Z}_{N(t)}$. Then $Y = (Y_t)_{t \geq 0}$ and $\bar{Y} = (\bar{Y}_t)_{t \geq 0}$ are both Markov chains in I with generator matrix G_x, starting from y and $\bar{y}$, respectively,[3] and are realized on the same probability space. We call the triple (ν, μ, F) a *coupling mechanism*. Define the *coupling time*

$$T_c = \inf\{t \geq 0 : Y_t = \bar{Y}_t\}.$$

Assume that, for some positive constant τ, for all $x \in U$ and all $y, \bar{y} \in I$,

$$m(x, y, \bar{y}) = \mathbb{E}_{(x,y,\bar{y})}(T_c) \leq \tau. \tag{10}$$

Fix a bounded measurable function f on $U \times I$ and set

$$\chi(x, y) = \mathbb{E}_{(x,y)} \int_0^{T_c} \bigl(f(x, Y_t) - f(x, \bar{Y}_t)\bigr)\, dt. \tag{11}$$

Then χ is well defined and, for all $x \in U$ and all $y \in I$,

$$|\chi(x, y)| \leq 2\tau\|f\|_\infty. \tag{12}$$

PROPOSITION 1.2. *The function χ is a corrector for f.*

---

[3]The process Y introduced here is not the fast variable, also denoted Y in the rest of the paper: the current Y is to be considered as a local approximation of the fast variable.

PROOF. In the proof we suppress the variable x. Note first that, if instead of taking $\bar{Z}_0 = \bar{y}$, we start $\bar{Z}$ randomly with the invariant distribution π, then we change the value of χ by a constant independent of y. Hence, it will suffice to establish the corrector equation $G\chi = f - \bar{f}$ in this case. Fix $\lambda > 0$ and define

$$\phi^\lambda(y) = \mathbb{E}\int_0^{T_\lambda} f(Y_t)\,dt, \qquad \bar{\phi}^\lambda = \mathbb{E}\int_0^{T_\lambda} f(\bar{Y}_t)\,dt,$$

where $T_\lambda = T_1/\lambda$, with T_1 an independent exponential random variable of parameter 1. Then, since Y and $\bar{Y}$ coincide after T_c,

$$\bar{\phi}^\lambda - \phi^\lambda(y) = \mathbb{E}\int_0^{T_\lambda \wedge T_c} \big(f(\bar{Y}_t) - f(Y_t)\big)\,dt \to \chi(y)$$

as $\lambda \to 0$. By elementary conditioning arguments, $(G - \lambda)\phi^\lambda + f = 0$ and $\lambda\bar{\phi}^\lambda = \bar{f}$, so

$$(G - \lambda)(\bar{\phi}^\lambda - \phi^\lambda) = f - \bar{f}.$$

On passing to the limit $\lambda \to 0$ in this equation, using bounded convergence we find that $G\chi = f - \bar{f}$, as required. $\square$

We remark that the corrector $\chi(x, \cdot)$ in fact depends only on f, G_x and the choice of $\bar{y}$, as the preceding proof makes clear. The further choice of a coupling mechanism is a way to obtain estimates on χ.

The following estimate will be used in dealing with the Δ term in (5). We write $\|\mu_x - \mu_{x'}\|$ for the total variation distance between μ_x and $\mu_{x'}$.

PROPOSITION 1.3. *For all $x, x' \in U$ and all $y \in I$,*

$$|\chi(x, y) - \chi(x', y)| \le 2\tau \sup_{z \in I} |f(x, z) - f(x', z)| \tag{13}$$
$$+ 2\nu\tau^2 \|f\|_\infty \|\mu_x - \mu_{x'}\|.$$

PROOF. By a standard construction (maximal coupling) there exists a sequence of independent random variables $((V_n, V_n') : n \in \mathbb{N})$ in $E \times E$ such that V_n has distribution μ_x, V_n' has distribution $\mu_{x'}$ and $\mathbb{P}(V_n \ne V_n') = \frac{1}{2}\|\mu_x - \mu_{x'}\| = \sup_{A \in \mathcal{E}} |\mu_x(A) - \mu_{x'}(A)|$, for all n. Write $(\mathcal{F}_t)_{t \ge 0}$ for the filtration of the marked Poisson process obtained by marking N with the random variables (V_n, V_n'). Construct $(Y, \bar{Y})$ from N and $(V_n : n \in \mathbb{N})$ as above. Similarly construct $(Y', \bar{Y}')$ from N and $(V_n' : n \in \mathbb{N})$. Recall that $T_c = \inf\{t \ge 0 : Y_t = \bar{Y}_t\}$ and set $T_c' = \inf\{t \ge 0 : Y_t' = \bar{Y}_t'\}$. Set $\lambda = \frac{1}{2}\nu\|\mu_x - \mu_{x'}\|$ and set

$$D = \inf\{t \ge 0 : (Y_t, \bar{Y}_t) \ne (Y_t', \bar{Y}_t')\}.$$

Then the process $t \mapsto 1_{\{D \le t\}} - \lambda t$ is an $(\mathcal{F}_t)_{t \ge 0}$-supermartingale and T_c is an $(\mathcal{F}_t)_{t \ge 0}$-stopping time. So, by optional stopping, we have $\mathbb{P}(D \le T_c) \le \lambda\mathbb{E}(T_c) \le$

$\lambda\tau$. Moreover, by the strong Markov property, on $\{D \le T_c\}$, we have $\mathbb{E}(T_c - D|\mathcal{F}_D) = m(x, Y_D, \bar{Y}_D) \le \tau$ so, for any function $g : I \to \mathbb{R}^d$, with $|g| \le \|f\|_\infty$,

$$\mathbb{E}\left|\int_{D\wedge T_c}^{T_c} g(Y_t)\,dt\right| \le \tau\|f\|_\infty \mathbb{P}(D \le T_c) \le \lambda\tau^2\|f\|_\infty.$$

On the other hand,

$$\int_0^{D\wedge T_c} g(Y_t)\,dt = \int_0^{D\wedge T'_c} g(Y'_t)\,dt$$

so

$$\left|\mathbb{E}\int_0^{T_c} g(Y_t)\,dt - \mathbb{E}\int_0^{T'_c} g(Y'_t)\,dt\right| \le 2\lambda\tau^2\|f\|_\infty = \nu\tau^2\|f\|_\infty\|\mu_x - \mu_{x'}\|.$$

We apply this estimate with $g = f(x, \cdot)$ to obtain

$$\begin{aligned}
&|\chi(x, y) - \chi(x', y)| \\
&\quad = \left|\mathbb{E}\int_0^{T_c} \big(f(x, \bar{Y}_t) - f(x, Y_t)\big)\,dt - \mathbb{E}\int_0^{T'_c} \big(f(x', \bar{Y}'_t) - f(x', Y'_t)\big)\,dt\right| \\
&\quad \le 2\tau \sup_{z\in I}|f(x, z) - f(x', z)| + \left|\mathbb{E}\int_0^{T_c} f(x, \bar{Y}_t)\,dt - \mathbb{E}\int_0^{T'_c} f(x, \bar{Y}'_t)\,dt\right| \\
&\qquad + \left|\mathbb{E}\int_0^{T_c} f(x, Y_t)\,dt - \mathbb{E}\int_0^{T'_c} f(x, Y'_t)\,dt\right| \\
&\quad \le 2\tau \sup_{z\in I}|f(x, z) - f(x', z)| + 2\nu\tau^2\|f\|_\infty\|\mu_x - \mu_{x'}\|
\end{aligned}$$

as required. □

To summarize, we have shown the following proposition.

PROPOSITION 1.4. *Assume conditions* (9) *and* (10). *Then*, *for any bounded measurable function* f *on* $U \times I$, *there exists a corrector* χ_f *for* f *satisfying the estimates* (12) *and* (13).

1.3. *Statement of the estimates*. Recall the context of Section 1.1. We consider a continuous-time Markov chain X with countable state-space S and generator matrix Q. We choose fluid coordinates $\mathbf{x} : S \to \mathbb{R}^d$ and an auxiliary coordinate $y : S \to I$. We choose also a subset $U \subseteq \mathbb{R}^d$, which provides a means of localization, together with a map $b : U \times I \to \mathbb{R}^d$, and a family $G = (G_x : x \in U)$ of generator matrices on I, each having a unique invariant distribution π_x. Also choose, as in the preceding subsection, a coupling mechanism for G. This comprises a constant $\nu > 0$, an auxiliary space E, a function $F : I \times E \to I$ and a family of probability distributions $\mu = (\mu_x : x \in U)$ on E such that

$$g(x, y, y') = \nu\mu_x\big(\{v \in E : F(y, v) = y'\}\big), \qquad x \in U, y, y' \in I \text{ distinct.}$$

Define for $x \in U$

$$\bar{b}(x) = \sum_{y \in I} b(x, y)\pi(x, y).$$

Write $\mathbf{X}_t = \mathbf{x}(X_t)$ and assume that $(x_t)_{0 \le t \le t_0}$ is a solution in U to $\dot{x}_t = \bar{b}(x_t)$. We use a scaled supremum norm on $\mathbb{R}^d$: fix positive constants $\sigma_1, \dots, \sigma_d$ and define for $x \in \mathbb{R}^d$

$$\|x\| = \max_{1 \le i \le d} |x_i|/\sigma_i.$$

We now introduce some constants $\Lambda, B, \tau, J, J_1(b), J(\mu), K$ which characterize certain regularity properties of Q, b and G. Assume that, for all $\xi \in S$, all $x \in U$ and all $y, y' \in I$,

$$q(\xi) \le \Lambda, \qquad \|b(x, y)\| \le B, \qquad m(x, y, y') \le \tau. \tag{14}$$

Here $m(x, y, y')$ is the mean coupling time for G_x starting from y and y', defined in the preceding subsection, which depends on the choice of coupling mechanism. Write $\mathcal{J}$ for the set of pairs of points in U between which $\mathbf{X}$ can jump, thus

$$\mathcal{J} = \{(x, x') \in U \times U : x = \mathbf{x}(\xi), x' = \mathbf{x}(\xi') \text{ for some } \xi, \xi' \in S \text{ with } q(\xi, \xi') > 0\}.$$

Set

$$\begin{aligned} J &= \sup_{(x,x') \in \mathcal{J}} \|x - x'\|, \\ J_1(b) &= \sup_{(x,x') \in \mathcal{J}, y \in I} \|b(x, y) - b(x', y)\|, \\ J(\mu) &= \sup_{(x,x') \in \mathcal{J}} \|\mu_x - \mu_{x'}\|. \end{aligned}$$

Write K for the Lipschitz constant of $\bar{b}$ on U; thus, for all $x, x' \in U$,

$$\|\bar{b}(x) - \bar{b}(x')\| \le K\|x - x'\|. \tag{15}$$

Recall from Section 1.1 the definitions of the drift vector β for $\mathbf{x}$ and the jump rate γ for y. Define

$$T = \inf\{t \ge 0 : \mathbf{X}_t \notin U\}.$$

Fix constants $\delta(\beta, b), \delta(\gamma, g) \in (0, \infty)$ and consider the events

$$\Omega(\beta, b) = \left\{ \int_0^{T \wedge t_0} \|\beta(X_t) - b(\mathbf{x}(X_t), y(X_t))\| \, dt \le \delta(\beta, b) \right\} \tag{16}$$

and

$$\Omega(\gamma, g) = \left\{ \int_0^{T \wedge t_0} \sum_{y' \ne y(X_t)} |\gamma(X_t, y') - g(\mathbf{x}(X_t), y(X_t), y')| \, dt \le \delta(\gamma, g) \right\}. \tag{17}$$

THEOREM 1.5. *Let $\varepsilon > 0$ be given and set $\delta = \varepsilon e^{-Kt_0}/7$. Assume that $J \le \varepsilon$ and*

$$\max\{\|\mathbf{X}_0 - x_0\|, \delta(\beta, b), 2\tau B\delta(\gamma, g), 2\tau B, 2\Lambda t_0(\tau J_1(b) + \nu\tau^2 BJ(\mu))\} \le \delta.$$

Set $\bar{J} = J + 4\tau B$ and assume that $\delta \le \Lambda \bar{J} t_0/4$. Further assume that the following tube condition holds:

$$\textit{for } \xi \in S \textit{ and } t \le t_0 \qquad \|\mathbf{x}(\xi) - x_t\| \le 2\varepsilon \quad \Longrightarrow \quad \mathbf{x}(\xi) \in U.$$

Then

$$\mathbb{P}\Big(\sup_{t \le t_0} \|\mathbf{X}_t - x_t\| > \varepsilon\Big) \le 2de^{-\delta^2/(4\Lambda \bar{J}^2 t_0)} + \mathbb{P}\big(\Omega(\beta, b)^c \cup \Omega(\gamma, g)^c\big).$$

The proof of this result follows the initial stages of the proof of the more elaborate Theorem 1.6 below. We will not write it out separately but give further indications immediately before the statement of Theorem 1.6. The reader will understand clearly the role of the inequalities which appear as hypotheses by following the proof. Here is an informal guide to their meanings. The tube condition, together with $J \le \varepsilon$, allows us to localize the other hypotheses to U by trapping the process inside a tube around the limit path; these conditions can be satisfied by choosing U sufficiently large. The conditions $\|\mathbf{X}_0 - x_0\| \le \delta$ and $\delta(\beta, b) \le \delta$ enforce that the initial conditions and drift fields match closely. This requires, in particular, that the fluid and auxiliary coordinates provide sufficient information to nearly determine β. The condition on $\delta(\gamma, g)$ forces a close match between the local behavior of the fast variable and the idealized fast process used to compute the corrector. The condition $2\tau B \le \delta$ allows us to control the size of the corrector, balancing the mean recurrence time of the fast variable τ against the range of the drift field b. The condition on $2\Lambda t_0(\tau J_1(b) + \nu\tau^2 BJ(\mu))$ is needed for local regularity of the corrector, allowing us to pass back from the idealized fast process at one point x to the actual fast variable when the fluid variable is near x. Finally, the condition $\delta \le \Lambda \bar{J} t_0/4$ ensures we are in the "Gaussian regime" of the exponential martingale inequality, where bad events cannot occur by a small number of large jumps. For a nontrivial limiting dynamics, ΛJ should be of order 1, while for a useful estimate ΛJ^2 should be small; thus, as expected, we can attempt to use the result when the Markov chain takes small jumps at a high rate.

It is sometimes possible to improve on the constant $\Lambda \bar{J}^2$ appearing in the preceding estimate, thereby obtaining useful probability bounds for smaller choices of ε. However, to do this we have to make hypotheses expressed in terms of a corrector. Fix $\bar{y} \in I$ and denote by χ the corrector for b given by (11). Define for $\xi \in S$ with $\mathbf{x}(\xi) \in U$

$$\bar{\mathbf{x}}(\xi) = \mathbf{x}(\xi) - \chi(\mathbf{x}(\xi), y(\xi)).$$

Define, for $\xi \in S$ such that $\mathbf{x}(\xi) \in U$ and $\mathbf{x}(\xi') \in U$ whenever $q(\xi, \xi') > 0$,

$$\alpha^i(\xi) = \sum_{\xi' \neq \xi} \{\bar{\mathbf{x}}^i(\xi') - \bar{\mathbf{x}}^i(\xi)\}^2 q(\xi, \xi'), \qquad i = 1, \ldots, d.$$

Note that, since we shall be interested only in upper bounds, we deal here only with the diagonal terms of the diffusivity tensor defined at (6). Choose functions $a^i : I \to [0, \infty)$ such that, for all $\xi \in U$ where $\alpha^i(\xi)$ is defined,

$$\alpha^i(\xi) \le a^i(y(\xi)), \qquad i = 1, \ldots, d. \tag{18}$$

For simplicity, we do not allow a to depend on the fluid variable $\mathbf{x}(\xi)$. Since we can localize our hypotheses near the (compact) limit path, we do not expect to lose much precision by this simplification. On the other hand, by permitting a dependence on the fast variable we can sometimes do significantly better than Theorem 1.5, as we shall see in Section 2. Set

$$\bar{a}(x) = \sum_{y \in I} a(y)\pi(x, y), \qquad x \in U.$$

We introduce two further constants A and $\bar{A}$, with $\bar{A} \le A \le \Lambda \bar{J}^2$. Assume that, for all $x \in U$ and all $y \in I$,

$$a^i(y) \le A\sigma_i^2, \qquad \bar{a}^i(x) \le \bar{A}\sigma_i^2, \qquad i = 1, \ldots, d. \tag{19}$$

Note that the corrector bound (12) gives $\|\chi(\mathbf{x}(\xi), y(\xi))\| \le 2\tau B$, so $\alpha^i(\xi) \le \Lambda \bar{J}^2 \sigma_i^2$ and so (19) holds with $A = \bar{A} = \Lambda \bar{J}^2$ and $a^i(y) = A\sigma_i^2$ and $\bar{a}^i(x) = A\sigma_i^2$. Thus Theorem 1.5 follows directly from (20) below. The new inequalities required on the left-hand side of (21) can be understood roughly as imposing that the ratio of the averaged diffusivity to a uniform bound on the diffusivity is not too small compared to the mean recurrence time of the fast variable; so an effective averaging takes place.

THEOREM 1.6. *Assume that the hypotheses of Theorem 1.5 hold and that $\delta \bar{J} \le At_0/4$. Then*

$$\mathbb{P}\Big(\sup_{t \le t_0} \|\mathbf{X}_t - x_t\| > \varepsilon\Big) \le 2de^{-\delta^2/(4At_0)} + \mathbb{P}\big(\Omega(\beta, b)^c \cup \Omega(\gamma, g)^c\big). \tag{20}$$

Moreover, under the further conditions $\delta \bar{J} \le \bar{A}t_0/4$ and

$$\frac{1}{t_0} \max\{\tau, \tau\delta(\gamma, g), \Lambda t_0 \nu \tau^2 J(\mu)\} \le \bar{A}/(20A) \le \Lambda\tau, \tag{21}$$

we have

$$\begin{aligned} \mathbb{P}\Big(\sup_{t \le t_0} \|\mathbf{X}_t - x_t\| > \varepsilon\Big) &\le 2de^{-\delta^2/(4\bar{A}t_0)} + 2de^{-(\bar{A}/A)^2 t_0/(6400\Lambda\tau^2)} \\ &\quad + \mathbb{P}\big(\Omega(\beta, b)^c \cup \Omega(\gamma, g)^c\big). \end{aligned} \tag{22}$$

PROOF. Consider the stopping time

$$T_0 = \inf\{t \geq 0 : \|\mathbf{X}_t - x_t\| > \varepsilon\}.$$

By the tube condition, we have $T_0 \leq T$. Moreover, for any $t < T_0$ and any $\xi' \in S$ such that $q(X_t, \xi') > 0$, we have

$$\|\mathbf{x}(\xi') - x_t\| \leq J + \|\mathbf{X}_t - x_t\| \leq 2\varepsilon$$

so by the tube condition $\mathbf{x}(\xi') \in U$.

Recall that χ is the corrector for b given by (11). For the proof of (22), we shall use (11) to also construct a corrector $\tilde{\chi}$ for a. Set $\tilde{\delta} = \bar{A}t_0/10$. Note from (12) the bounds

$$\|\chi(x, y)\| \leq 2\tau B \leq \delta, \qquad |\tilde{\chi}^i(x, y)| \leq 2\tau A\sigma_i^2 \leq \tilde{\delta}\sigma_i^2.$$

The inequality involving $\tilde{\delta}$ and further such inequalities below, which depend on the first inequality in assumption (21), will not be used in the proof of (20). Write $\Delta = G\chi(\mathbf{x}, y) - Q(\chi(\mathbf{x}, y)) = \Delta_1 + \Delta_2$ and $\tilde{\Delta} = G\tilde{\chi}(\mathbf{x}, y) - Q(\tilde{\chi}(\mathbf{x}, y)) = \tilde{\Delta}_1 + \tilde{\Delta}_2$, where

$$\Delta_1(\xi) = \sum_{y' \neq y(\xi)} \{g(\mathbf{x}(\xi), y(\xi), y') - \gamma(\xi, y')\}\chi(\mathbf{x}(\xi), y') \tag{23}$$

and

$$\Delta_2(\xi) = \sum_{\xi' \neq \xi} q(\xi, \xi')\{\chi(\mathbf{x}(\xi), y(\xi')) - \chi(\mathbf{x}(\xi'), y(\xi'))\} \tag{24}$$

and where $\tilde{\Delta}_1$ and $\tilde{\Delta}_2$ are defined analogously. Then, on $\Omega(\gamma, g)$, for $t \leq T \wedge t_0$,

$$\left\| \int_0^t \Delta_1(X_s)\, ds \right\| \leq 2\tau B\delta(\gamma, g) \leq \delta$$

and, using Proposition 1.3,

$$\left\| \int_0^t \Delta_2(X_s)\, ds \right\| \leq 2\Lambda t_0\big(\tau J_1(b) + \nu\tau^2 BJ(\mu)\big) \leq \delta.$$

Similarly, for $t \leq T \wedge t_0$,

$$\left| \int_0^t \tilde{\Delta}_1^i(X_s)\, ds \right| \leq 2\tau A\delta(\gamma, g)\sigma_i^2 \leq \tilde{\delta}\sigma_i^2$$

and

$$\left| \int_0^t \tilde{\Delta}_2^i(X_s)\, ds \right| \leq 2\Lambda t_0 \nu\tau^2 AJ(\mu)\sigma_i^2 \leq \tilde{\delta}\sigma_i^2.$$

Take $M = M^{\bar{\mathbf{x}}}$ as in equations (3) and (5) and consider the event

$$\Omega(M) = \Big\{ \sup_{t \leq T_0 \wedge t_0} \|M_t\| \leq \delta \Big\}.$$

Then, on $\Omega(\beta, b) \cap \Omega(\gamma, g) \cap \Omega(M)$, we can estimate the terms in (5) to obtain for $t \le T_0 \wedge t_0$,

$$\|\mathbf{X}_t - x_t\| \le 7\delta + K \int_0^t \|\mathbf{X}_s - x_s\| \, ds,$$

so that $\|\mathbf{X}_t - x_t\| \le \varepsilon$ by Gronwall's lemma. Note that this forces $T_0 \ge t_0$ and hence, $\sup_{t \le t_0} \|\mathbf{X}_t - x_t\| \le \varepsilon$. Set $\rho = 3\bar{A}/2$ and consider the event

$$\Omega(a) = \left\{ \int_0^{T_0 \wedge t_0} a^i(Y_s) \, ds \le \rho t_0 \sigma_i^2 \text{ for } i = 1, \ldots, d \right\}.$$

By condition (18), on $\Omega(a)$ we have

$$\int_0^{T_0 \wedge t_0} \alpha^i(X_s) \, ds \le \rho t_0 \sigma_i^2.$$

Set

$$J_i = J(\bar{\mathbf{x}}^i) = \sup_{\xi, \xi' \in S, \mathbf{x}(\xi), \mathbf{x}(\xi') \in U, q(\xi, \xi') > 0} |\bar{\mathbf{x}}^i(\xi) - \bar{\mathbf{x}}^i(\xi')|, \qquad i = 1, \ldots, d,$$

and use (12) to see that $J_i \le \bar{J}\sigma_i$. Determine $\theta_i \in (0, \infty)$ by $\theta_i e^{\theta_i J_i} = \delta/(\rho t_0 \sigma_i)$; then $\theta_i \le \delta/(\rho t_0 \sigma_i)$, so $\theta_i J_i \le 2\delta \bar{J}/(3\bar{A} t_0) \le 1/4$, since we assumed that $\delta \bar{J} \le \bar{A} t_0/4$. Since $e^{1/4} \le 4/3$, we have $\rho e^{\theta_i J_i} \le 2\bar{A}$. We now apply the exponential martingale inequality, Proposition 1.1, substituting $\pm \bar{\mathbf{x}}^i$ for ϕ for $i = 1, \ldots, d$ and substituting $\delta \sigma_i$ for δ and $\rho t_0 \sigma_i^2$ for ε. We thus obtain

$$\mathbb{P}\big(\Omega(M)^c \cap \Omega(a)\big) \le 2d e^{-\delta^2/(4\bar{A} t_0)}.$$

If we take $\bar{A} = A$, then, using (18) and (19), we have $\Omega(a) = \Omega$, so the proof of (20) is now complete.

Set $\eta = 16\Lambda \tau^2 A^2$. We shall complete the proof of (22) by showing that

$$\mathbb{P}\big(\Omega(a)^c \cap \Omega(\gamma, g)\big) \le 2d e^{-\tilde{\delta}^2/(4\eta t_0)}.$$

Take $\tilde{M}$ as in (8), with a as in (18). Then, for $t \le T$,

$$\begin{aligned} \int_0^t a(Y_s) \, ds &= \tilde{\chi}(\mathbf{X}_t, Y_t) - \tilde{\chi}(\mathbf{X}_0, Y_0) - \tilde{M}_t \\ &\quad + \int_0^t \tilde{\Delta}(X_s) \, ds + \int_0^t \bar{a}(\mathbf{X}_s) \, ds, \end{aligned} \tag{25}$$

where $\tilde{\Delta} = G\tilde{\chi}(\mathbf{x}, y) - Q(\tilde{\chi}(\mathbf{x}, y))$. Consider the event

$$\Omega(\tilde{M}) = \left\{ \sup_{t \le T_0 \wedge t_0} |\tilde{M}_t^i| \le \tilde{\delta} \sigma_i^2 \text{ for } i = 1, \ldots, d \right\}.$$

Then, on $\Omega(\gamma, g) \cap \Omega(\tilde{M})$, we can estimate the terms in (25) to obtain

$$\int_0^{T_0 \wedge t_0} a^i(Y_s) \, ds \le (5\tilde{\delta} + \bar{A} t_0) \sigma_i^2 \le \rho t_0 \sigma_i^2.$$

Hence, it will suffice to show that

$$\mathbb{P}(\Omega(\tilde{M})^c) \le 2de^{-\tilde{\delta}^2/(4\eta t_0)}.$$

For this, we again use the exponential martingale inequality. Take $\phi(\xi) = \pm\tilde{\chi}^i(\mathbf{x}(\xi), y(\xi))$ in Proposition 1.1 and note that $\alpha^\phi(\xi) \le 16\Lambda\tau^2 A^2\sigma_i^4$, so

$$\int_0^{T_0\wedge t_0} \alpha^\phi(X_s)\,ds \le 16\Lambda\tau^2 A^2\sigma_i^4 t_0 = \eta t_0\sigma_i^4.$$

Set

$$\tilde{J}_i = J(\phi) = \sup_{\xi,\xi'\in S,\mathbf{x}(\xi),\mathbf{x}(\xi')\in U,q(\xi,\xi')>0} |\phi(\xi) - \phi(\xi')|, \qquad i = 1,\dots,d,$$

then $\tilde{J}_i \le 4\tau A\sigma_i^2$. Determine $\tilde{\theta}_i \in (0,\infty)$ by $\tilde{\theta}_i e^{\tilde{\theta}_i\tilde{J}_i} = \tilde{\delta}/(\eta t_0\sigma_i^2)$. Then $\tilde{\theta}_i \le \tilde{\delta}/(\eta t_0\sigma_i^2)$ so $\tilde{\theta}_i\tilde{J}_i \le \bar{A}/(40\Lambda\tau A) \le 1/2$ and so $e^{\tilde{\theta}_i\tilde{J}_i} \le 2$. Hence,

$$\mathbb{P}(\Omega(\tilde{M})^c) \le 2d\exp\{-\tilde{\delta}^2/(2\eta t_0 e^{\tilde{\theta}_i\tilde{J}_i})\} \le 2de^{-\tilde{\delta}^2/(4\eta t_0)}$$

as required. □

2. The supermarket model with memory. The supermarket model with memory is a variant, introduced in [8], of the "join the shortest queue" model, which has been widely studied [3–7, 9]. We shall rigorously verify the asymptotic picture for large numbers of queues derived in [8]. This will serve as an example to illustrate the general theory of the preceding sections. The explicit form of the error probabilities in Theorem 1.6 is used to advantage in dealing with the infinite-dimensional character of the limit model.

Fix $\lambda \in (0,1)$ and an integer $n \ge 1$. We shall consider the limiting behavior as $N \to \infty$ of the following queueing system. Customers arrive as a Poisson process of rate $N\lambda$ at a system of N single-server queues. At any given time, the length of one of the queues is kept under observation. This queue is called the *memory queue*. On each arrival, an independent random sample of size n is chosen from the set of all N queues. For simplicity, we sample with replacement, allowing repeats and allowing the choice of the memory queue. The customer joins whichever of the memory queue or the sampled queues is shortest, choosing randomly in the event of a tie. Immediately after the customer has joined a queue, we switch the memory queue, if necessary, so that it is the currently shortest queue among the queues just sampled and the previous memory queue. The service requirements of all customers are assumed independent and exponentially distributed of mean 1.

Write $Z_t^k = Z_t^{N,k}$ for the proportion of queues having at least k customers at time t, and write Y_t for length of the memory queue at time t. Set $Z_t = (Z_t^k : k \in \mathbb{N})$ and $X_t = (Z_t, Y_t)$. Then $X = (X_t)_{t\ge0}$ is a Markov chain, taking values in $S = S_0 \times \mathbb{Z}^+$, where S_0 is the set of nonincreasing sequences in $N^{-1}\{0,1,\dots,N\}$ with finitely many nonzero terms. We shall treat Y as a fast variable and prove a fluid limit for Z as $N \to \infty$.

2.1. *Statement of results.* Let D be the set of nonincreasing sequences[4] $z = (z_k : k \in \mathbb{N})$ in the interval $[0, 1]$ such that

$$m(z) := \sum_k z_k < \infty.$$

Define for $z \in D$ and $k \in \mathbb{N}$

$$\mu(z,k) = \prod_{j=1}^{k} \frac{z_j^n}{1 - p_{j-1}(z)}, \tag{26}$$

where

$$p_{k-1}(z) = n(z_{k-1} - z_k) z_k^{n-1}$$

and where we take $z_0 = 1$. Set $\mu(z, 0) = 1$ for all z. An elementary calculation (maximizing over z_k while keeping z_{k-1} fixed) shows that in the case $n \geq 2$,

$$p_{k-1}(z) \leq z_{k-1}^n (1 - 1/n)^{n-1} \leq (1 - 1/n)^{n/2} \leq e^{-1/2} < 1. \tag{27}$$

In the case $n = 1$ we have $p_{k-1}(z) = z_{k-1} - z_k \leq 1$ and it is possible that $0/0$ appears in the product (26). For definiteness we agree to set $0/0 = 1$ in this case. Note that $\mu(z, k) \geq \mu(z, k + 1)$ for all $k \geq 0$. Define for $z \in D$

$$v_k(z) = \lambda z_{k-1}^n \mu(z, k-1) - \lambda z_k^n \mu(z, k) - (z_k - z_{k+1})$$

and consider the differential equation

$$\dot{z}(t) = v(z(t)), \qquad t \geq 0. \tag{28}$$

By a *solution* to (28) in D we mean a family of differentiable functions $z_k : [0, \infty) \to [0, 1]$ such that for all $t \geq 0$ and $k \in \mathbb{N}$ we have $(z_k(t) : k \in \mathbb{N}) \in D$ and

$$\dot{z}_k(t) = v_k(z(t)).$$

THEOREM 2.1. *For all $z(0) \in D$, the differential equation $\dot{z}(t) = v(z(t))$ has a unique solution in D starting from $z(0)$. Moreover, if $(w(t) : t \in D)$ is another solution in D with $z_k(0) \leq w_k(0)$ for all k, then $z_k(t) \leq w_k(t)$ for all k and all $t \geq 0$.*

There is a fixed point of these dynamics $a \in D$ given by setting $a_0 = 1$ and defining

$$a_{k+1} = \lambda a_k^n \mu(a, k), \qquad k \geq 0. \tag{29}$$

[4]To lighten the notation, we shall sometimes move the coordinate index from a superscript to a subscript, allowing the nth power of the kth coordinate to be written z_k^n. We shall also write the time variable sometimes as a subscript, sometimes as an argument.

The components of a decay super-geometrically. Set

$$\alpha = n + \tfrac{1}{2} + \sqrt{n^2 + \tfrac{1}{4}}. \tag{30}$$

Then $\alpha \in (2n, 2n+1)$.

THEOREM 2.2. *We have*

$$\lim_{k\to\infty} \frac{1}{k} \log\log\left(\frac{1}{a_k}\right) = \alpha.$$

Assume for simplicity that we start the queueing system from the state where all queues, except the memory queue, are empty and where the memory queue has exactly one customer. Write $(z(t): t \geq 0)$ for the solution to (28) starting from 0. Then $z_k(t) \leq a_k$ for all k and t. Our main result shows that $(z(t): t \geq 0)$ is a good approximation to the process of empirical distributions of queue lengths $(Z^N(t): t \geq 0)$ for large N. The sense of this approximation is reasonably sharp. In particular, as a straightforward corollary, we obtain that, on a given time interval $[0, t_0]$, for any $r > \alpha^{-1}$, with high probability as $N \to \infty$, no queue length exceeds $r \log\log N$.

THEOREM 2.3. *Set* $\kappa = (2\alpha)^{-1}$ *and define*

$$d = d(N) = \sup\{k \in \mathbb{N} : N a_k > N^\kappa\}.$$

Fix a function ϕ *on* $\mathbb{N}$ *such that* $\phi(N)/N^\kappa \to 0$ *and* $\log\phi(N)/\log\log N \to \infty$ *as* $N \to \infty$. *Set* $\rho = 4/(1-\lambda)$ *when* $n = 1$ *and set* $\rho = 2^n/(1 - e^{-1/2})$ *when* $n \geq 2$. *Set* $\tilde{a}_{d+1} = N^{-1} a_d^n + \rho^d a_{d+1}$. *Then*

$$\lim_{N\to\infty} d(N)/\log\log N = 1/\alpha. \tag{31}$$

Moreover, for all $t_0 \geq 0$, *we have*

$$\lim_{N\to\infty} \mathbb{P}\left(\sup_{t\leq t_0}\sup_{k\leq d} \frac{|Z_t^{N,k} - z_t^k|}{\sqrt{a_k}} \geq \sqrt{\frac{\phi(N)}{N}}\right) = 0 \tag{32}$$

and

$$\lim_{R\to\infty} \limsup_{N\to\infty} \mathbb{P}(Z_t^{N,d+1} \geq R\tilde{a}_{d+1} \textit{ for some } t \leq t_0) = 0 \tag{33}$$

and

$$\lim_{N\to\infty} \mathbb{P}(Z_t^{N,d+2} = 0 \textit{ for all } t \leq t_0) = 1. \tag{34}$$

The argument used to prove this result would apply without modification starting from any initial condition $z(0)$ for the limit dynamics (28) such that $z_k(0) \le a_k$ for all k, with suitable conditions on the convergence of $Z^N(0)$ to $z(0)$. It may be harder to move beyond initial conditions which do not lie below the fixed point. We do note here, however, a family of long-time upper bounds for the limit dynamics which might prove useful for such an extension. Fix $j \in \mathbb{N}$ and define $a_k^{(j)} = a_{(k-j)^+}$ for each $k \in \mathbb{Z}^+$; then $a^{(j)}$ is a fixed point of the modified equation

$$\dot{w}_k(t) = v_k(w(t)) + \big(w_j(t) - w_{j+1}(t)\big)1_{\{k=j\}}.$$

Since the added term is always nonnegative, a similar argument to that used to prove Theorem 2.1 in the next subsection also shows that, if $z(0) \le a^{(j)}$ and $(z(t) : t \ge 0)$ is a solution of the original equation, then $z(t) \le a^{(j)}$ for all t.

2.2. *Existence and monotonicity of the limit dynamics.* The differential equation (28) characterizes the limit dynamics for the fluid variables in our queueing model. Our analysis of its space of solutions will rest on the exploitation of certain nonnegativity properties which have a natural probabilistic interpretation. We shall use the following standard property of differential equations: if $b = (b^1, \dots, b^d)$ is a Lipschitz vector field on $\mathbb{R}^d$ such that $b^1(x) \ge 0$ whenever $x = (x^1, \dots, x^d)$ with $x^1 = 0$, and if $(x(t) : t \le t_0)$ is a solution to $\dot{x}(t) = b(x(t))$ with $x^1(0) \ge 0$, then $x^1(t) \ge 0$ for all $t \le t_0$.

We consider first a truncated, finite-dimensional system. Fix $d \in \mathbb{N}$ and define a vector field $u = u^{(d)}$ on D by setting $u_k(z) = v_k(z)$ for $k \le d-1$ and

$$u_d(z) = \lambda z_{d-1}^n \mu(z, d-1) - \lambda z_d^n \mu(z, d) - z_d \tag{35}$$

and $u_k(z) = 0$ for $k \ge d+1$. Set $D(d) = \{(x_1, \dots, x_d, 0, 0, \dots) : 0 \le x_d \le \dots \le x_1 \le 1\}$.

PROPOSITION 2.4. *For all $x(0) \in D(d)$, the differential equation $\dot{x}(t) = u(x(t))$ has a unique solution $(x(t) : t \ge 0)$ in $D(d)$ starting from $x(0)$.*

PROOF. In the proof, we consider $D(d)$ as a subset of $\mathbb{R}^d$. The function u is continuous on $D(d)$ and is differentiable in the interior of $D(d)$ with bounded partial derivatives. [In the case $n = 1$, the singularity in $(\partial/\partial x_j)\mu(x, k)$ for $j \le k$ as $x_{j-1} - x_j \to 1$ is canceled by the factor x_k by which it is multiplied, since $x_k \le x_j$ on $D(d)$.] For $x \in D(d)$ we have $u_1(x) \le 0$ when $x_1 = 1$, and $u_d(x) \ge 0$ when $x_d = 0$. Moreover, for $k = 1, \dots, d-1$, if $x_k = x_{k+1}$ then $p_k(x) = 0$ so

$$\begin{aligned} u_{k+1}(x) &= x_k^n\big(\mu(x,k) - \mu(x,k+1)\big) \le \mu(x,k) - \mu(x,k+1) \\ &\le x_{k-1}^n \mu(x, k-1) - x_k^n \mu(x,k) \le u_k(x). \end{aligned}$$

The conclusion now follows by standard arguments. $\square$

PROOF OF THEOREM 2.1. Suppose that $(w(t):t\geq 0)$ is a solution to $\dot w(t)=v(w(t))$ in D starting from $w(0)$, with $z(0)\leq w(0)$, that is to say $z_k(0)\leq w_k(0)$ for all k. Fix d and write $x(t)=z^{(d)}(t)$ for the solution to $\dot x(t)=u^{(d)}(x(t))$ in $D(d)$ starting from $(z_1(0),\dots,z_d(0),0,0,\dots)$. Set $y(t)=(w_1(t),\dots,w_d(t),0,0,\dots)$ and note that $x(0)\leq y(0)$ and $y(t)\in D(d)$ for all t. We shall show that $x(t)\leq y(t)$ for all t. Now consider $D(d)$ as a subset of $\mathbb{R}^d$. We have

$$\dot y(t)=u(y(t))+w^{d+1}(t)e_d,$$

where $e_d=(0,\dots,0,1)$. Note that $w^{d+1}(t)\geq 0$ for all t. Now u is Lipschitz on $D(d)$ and for $k=1,\dots,d$ we can show that[5]

$$x,y\in D(d),\qquad x\leq y,\qquad x_k=y_k\quad\Longrightarrow\quad u_k(x)\leq u_k(y).$$

Hence, by a standard argument $z^{(d)}(t)=x(t)\leq y(t)\leq w(t)$ for all t. The same argument shows that $z^{(d)}(t)\leq z^{(d+1)}(t)$ for all t, so the limit $z_k(t)=\lim_{d\to\infty}z_k^{(d)}(t)$ exists for all k and t, and $z(t)\leq w(t)$ for all t.

Fix k and take $d\geq k+1$. Then the following equation holds for all t:

$$\begin{aligned}z_k^{(d)}(t)&+\int_0^t\lambda z_k^{(d)}(s)^n\mu\big(z^{(d)}(s),k\big)\,ds+\int_0^t z_k^{(d)}(s)\,ds\\&=z_k(0)+\int_0^t\lambda z_{k-1}^{(d)}(s)^n\mu\big(z^{(d)}(s),k-1\big)\,ds+\int_0^t z_{k+1}^{(d)}(s)\,ds.\end{aligned}$$

On letting $d\to\infty$, we see by monotone convergence that

$$\begin{aligned}z_k(t)&+\int_0^t\lambda z_k(s)^n\mu(z(s),k)\,ds+\int_0^t z_k(s)\,ds\\&=z_k(0)+\int_0^t\lambda z_{k-1}(s)^n\mu\big(z(s),k-1\big)\,ds+\int_0^t z_{k+1}(s)\,ds.\end{aligned}$$

Since $z(t)\in D$ for all t, all integrands in this equation are bounded by 1. It is now straightforward to see that $(z(t):t\geq 0)$ is a solution.

Now

$$\begin{aligned}w_k(t)&+\int_0^t\lambda w_k(s)^n\mu(w(s),k)\,ds+\int_0^t w_k(s)\,ds\\&=w_k(0)+\int_0^t\lambda w_{k-1}(s)^n\mu\big(w(s),k-1\big)\,ds+\int_0^t w_{k+1}(s)\,ds.\end{aligned}$$

---

[5]An elementary calculation shows that $\mu(x,j)\leq\mu(y,j)$ for all j whenever $x\leq y$. This will also be shown by a soft probabilistic argument in Section 2.6. The further condition $x_k=y_k$ gives the inequality

$$y_{k-1}^n-x_{k-1}^n\geq n(y_{k-1}-x_{k-1})x_k^{n-1}=p_{k-1}(y)-p_{k-1}(x)\geq\frac{y_k^{2n}}{1-p_{k-1}(y)}-\frac{x_k^{2n}}{1-p_{k-1}(x)}.$$

By summing these equations over $k \in \{1, \ldots, d\}$ we see that the map $t \mapsto \sum_{k=1}^{d} w_k(t) - \lambda t$ is nonincreasing for all d. Hence, $m(w(t)) \le m(w(0)) + \lambda t < \infty$. The equations can then be summed over all k and rearranged to obtain

$$m(w(t)) = m(w(0)) + \lambda t - \int_0^t w_1(s)\,ds.$$

On the other hand,

$$m\big(z^{(d)}(t)\big) = m\big(z^{(d)}(0)\big) + \lambda t - \int_0^t z_1^{(d)}(s)\,ds - \lambda \int_0^t z_d^{(d)}(s)^n \mu\big(z^{(d)}(s), d\big)\,ds$$

so

$$\begin{aligned} m(w(t)) - m\big(z^{(d)}(t)\big) \le m(w(0)) &- m\big(z^{(d)}(0)\big) \\ &+ \lambda \int_0^t z_d(s)^n \mu(z(s), d)\,ds. \end{aligned} \tag{36}$$

If $w(0) = z(0)$ then the right-hand side tends to 0 as $d \to \infty$ so we must have $z(t) = w(t)$ for all t. $\square$

2.3. *Properties of the fixed point.* Recall the definition (29) of the fixed point a. Since $\mu(a, k) \le 1$ for all k, we have

$$a_k \le \lambda^{1+n+\cdots+n^{k-1}}$$

so $a_k \to 0$ as $k \to \infty$. Theorem 2.2 is then a straightforward corollary of the following estimate.

PROPOSITION 2.5. *There is a constant $C(\lambda, n) < \infty$ such that, for all $k \ge 0$,*

$$C^{-1} a_k^{\alpha} \le a_{k+1} \le C a_k^{\alpha}, \tag{37}$$

where α is given by (30).

PROOF. Note that, since $a_1 = \lambda$, we have $p_{k-1}(a) = a_{k-1} - a_k \le \lambda \vee (1-\lambda) < 1$ for all k when $n = 1$. On the other hand, equation (27) gives $p_{k-1}(z) \le e^{-1/2}$ for all k when $n \ge 2$. Then from $\sum_{k=1}^{\infty} p_{k-1}(a) \le 1$, we obtain a constant $c < \infty$, which may depend on λ when $n = 1$, such that

$$\prod_{k=1}^{\infty} \frac{1}{1 - p_{k-1}(a)} \le c.$$

Then for $k \ge 0$

$$\lambda a_k^{2n} \prod_{j=1}^{k-1} a_j^n \le a_{k+1} \le c\lambda a_k^{2n} \prod_{j=1}^{k-1} a_j^n,$$

so for $k \geq 1$,

$$c^{-1}a_k^{2n+1}a_{k-1}^{-n} \leq a_{k+1} \leq ca_k^{2n+1}a_{k-1}^{-n}. \tag{38}$$

Note that $\lambda a_0^\alpha = \lambda = a_1 \leq \lambda^{-1}a_0^\alpha$. Fix $A \geq 1/\lambda$ and suppose inductively that

$$A^{-1}a_{k-1}^\alpha \leq a_k \leq Aa_{k-1}^\alpha.$$

On using these inequalities to estimate a_{k-1} in (38), we obtain

$$(cA^{n/\alpha})^{-1}a_k^\alpha \leq a_{k+1} \leq cA^{n/\alpha}a_k^\alpha,$$

where we have used the fact that $1 - n/\alpha = \alpha - 2n$. Hence, the induction proceeds provided we take $A \geq c^{\alpha/(\alpha-n)}$. $\square$

2.4. *Choice of fluid coordinates and fast variable*. In the remaining subsections we apply Theorem 1.6 to deduce Theorem 2.3. Define d as in Theorem 2.3 and take as auxiliary space $I = \mathbb{N}$ when $n = 1$ and $I = \mathbb{Z}^+$ when $n \geq 2$. Make the following choice of fluid and auxiliary coordinates: for $\xi = (z, y) \in S$ with $z = (z_k : k \in \mathbb{N})$, set

$$x^k(\xi) = z_k, \qquad k = 1, \dots, d, \qquad y(\xi) = y.$$

Thus our fluid variable is $\mathbf{X}_t = \mathbf{x}(X_t) = (Z_t^1, \dots, Z_t^d)$ and our fast variable is $Y_t = y(X_t)$. Note that when $n = 1$, if $Y_0 \geq 1$ then $Y_t \geq 1$ for all t, so Y takes values in $I = \mathbb{N}$.

Let us compute the drift vector $\beta(\xi)$ for $\mathbf{X}$ when X is in state $\xi = (z, y) \in S$. Note that X^k makes a jump of size $1/N$ when a customer arrives at a queue of length $k - 1$, and makes a jump of size $-1/N$ when a customer departs from a queue of length k, otherwise X^k is constant. The length of the queue which an arriving customer joins depends on the length of the memory queue y and on the lengths of the sampled queues. Denote the vector of sampled queue lengths by $V = V(z) = (V_1, \dots, V_n)$ and write $V^{(1)} \leq V^{(2)} \leq \cdots \leq V^{(n)}$ for the ordered queue lengths. Define $\min(v) = v_1 \wedge \cdots \wedge v_n$ and set $M = \min(V)$. Then $M = V^{(1)}$ and

$$\mathbb{P}(M \geq k) = z_k^n.$$

A new customer will go to a queue of length at least k if and only if $M \geq k$ and $y \geq k$. So the rate for an arrival to a queue of length exactly $k - 1$ is

$$N\lambda\mathbb{P}(M \geq k-1)1_{\{y \geq k-1\}} - N\lambda\mathbb{P}(M \geq k)1_{\{y \geq k\}}.$$

The rate for a departure from a queue of length k is $N(z_k - z_{k+1})$. Hence, setting $z_0 = 1$, we have

$$\beta_k(\xi) = \lambda z_{k-1}^n 1_{\{y \geq k-1\}} - \lambda z_k^n 1_{\{y \geq k\}} - (z_k - z_{k+1}).$$

We now compute (an approximation to) the jump rates $\gamma(\xi, y')$ for Y when X is in state $\xi = (z, y) \in S$. The rate of departures from the memory queue is at most 1.

Arrivals to the system occur at rate $N\lambda$. Occasionally, the memory queue falls in the sample, an event of probability no greater than n/N and hence, of rate no greater than λn. Assuming that the the memory queue does not fall in the sample, the length of the memory queue after an arrival is given by

$$F(y, V) = (y+1)1_{\{y \le M-1\}} + y1_{\{M \le y \le P\}} + P1_{\{y \ge P+1\}}, \tag{39}$$

where $P = p(V)$ is given by $P = M + 1$ when $n = 1$ and $P = (M + 1) \wedge V^{(2)}$ otherwise. Hence, we have

$$\sum_{y' \neq y(\xi)} \bigl|\gamma(\xi, y') - N\lambda\mathbb{P}\bigl(F(y, V(z)) = y'\bigr)\bigr| \le 1 + \lambda n. \tag{40}$$

2.5. *Choice of limit characteristics and coupling mechanism.* Define

$$U = \{x \in \mathbb{R}^d : 0 \le x_d \le \cdots \le x_1 \le 1 \text{ and } x_1 \le (\lambda+1)/2 \text{ and } x_k \le 2a_k \text{ for all } k\}.$$

The condition $x_1 \le (\lambda + 1)/2$ ensures that $1 - x_1$ is uniformly positive on U. Define $b : U \times \mathbb{Z}^+ \to \mathbb{R}^d$ by

$$b_k(x, y) = \lambda x_{k-1}^n 1_{\{y \ge k-1\}} - \lambda x_k^n 1_{\{y \ge k\}} - (x_k - x_{k+1}), \tag{41}$$

where we set $x_0 = 1$ and $x_{d+1} = 0$. Then, for $\xi \in S$ with $\mathbf{x}(\xi) \in U$, we have

$$\beta(\xi) = b(\mathbf{x}(\xi), y(\xi)) + (0, \ldots, 0, z_{d+1}). \tag{42}$$

It is convenient to specify our choice of the generator matrices $(G_x : x \in U)$ and our choice of coupling mechanism at the same time. Set $\nu = N\lambda$ and take as auxiliary space $E = (\mathbb{Z}^+)^n$. Define a family of probability distributions $\mu = (\mu_x : x \in U)$ on E, taking μ_x to be the law of a random sample $V = V(x) = (V_1, \ldots, V_n)$ with

$$\mathbb{P}(V_1 \ge k) = \cdots = \mathbb{P}(V_n \ge k) = x_k$$

for $k = 0, 1, \ldots, d + 1$. Note that

$$\|\mu_x - \mu_{x'}\| \le 2n \sum_{k=1}^{d} |x_k - x'_k|. \tag{43}$$

Then define for distinct $y, y' \in \mathbb{Z}^+$,

$$g(x, y, y') = N\lambda\mathbb{P}\bigl(F(y, V(x)) = y'\bigr),$$

where F is given by (39). We take as coupling mechanism the triple (ν, μ, F).

Note that $F(y, v) = F(\bar{y}, v)$ for all $y, \bar{y} \in I$ whenever $p(v) = \min I$. For $x \in U$ we have

$$\mathbb{P}\bigl(p(V(x)) = 1\bigr) \ge 1 - x_1 > \frac{1-\lambda}{2},$$

when $n = 1$, whereas for $n \ge 2$ we have

$$\mathbb{P}\bigl(p(V(x)) = 0\bigr) \ge (1 - x_1)^2 > \left(\frac{1-\lambda}{2}\right)^2.$$

Hence, we obtain, in all cases, $m(x, y, \bar{y}) \le \tau$, where we set

$$\tau = \frac{4}{N\lambda(1-\lambda)^2}.$$

For $\xi \in S$ with $x = \mathbf{x}(\xi) \in U$ we can realize a sample $V(z)$ (from the distribution of queue lengths) and the sample $V(x)$ on the same probability space by setting $V_i(x) = V_i(z) \wedge d$. Write $M(x) = \min(V(x))$ and $P(x) = p(V(x))$. Then $M(x) = M(z) \wedge d$ and $P(x) = P(z) \wedge (d+1)$ when $n = 1$ and $P(x) = P(z) \wedge d$ when $n \ge 2$. The difference between the two cases is that there is no second shortest queue in the sample when $n = 1$. We have, for $n = 1$,

$$\mathbb{P}\big(P(z) \ne P(x)\big) \le \mathbb{P}\big(M(z) \ge d+1\big) = z_{d+1}$$

and, for $n \ge 2$,

$$\mathbb{P}\big(P(z) \ne P(x)\big) \le \mathbb{P}\big(P(z) \ge d+1\big) \le n z_d z_{d+1}^{n-1}.$$

Now $P(x) = P(z)$ implies $M(x) = M(z)$ and hence, $F(y, V(x)) = F(y, V(z))$ for all y. Hence,

$$\mathbb{P}\big(F(y, V(z)) \ne F(y, V(x))\big) \le \mathbb{P}\big(P(x) \ne P(z)\big).$$

On combining this with (40) we obtain

$$\sum_{y' \ne y(\xi)} |\gamma(\xi, y') - g(\mathbf{x}(\xi), y(\xi), y')| \le 1 + \lambda n + N\lambda n z_{d+1}. \tag{44}$$

2.6. *Local equilibrium distribution.* The Markov chain determined by the generator G_x has a unique closed communicating class, which is contained in $\{0, 1, \dots, d\}$. Hence, G_x has a unique equilibrium distribution π_x which is supported on $\{0, 1, \dots, d\}$. Consider a continuous-time Markov chain[6] $Y = (Y_t)_{t \ge 0}$ with generator G_x, and initial distribution π_x. Set $\mu(x, k) = \mathbb{P}(Y_0 \ge k)$. Then Y jumps into $\{0, 1, \dots, k\}$ from j at rate $\alpha = N\lambda(1 - x_{k+1}^n - p_k(x))$, for all $j \ge k+1$. On the other hand, Y jumps out of $\{0, 1, \dots, k\}$ only from k, and that at rate $\beta = N\lambda x_{k+1}^n$. Since the long run rates of such jumps must agree, we deduce that $\alpha\mu(x, k+1) = \beta\pi_x(k)$. Hence, we obtain

$$\mu(x, k+1)\big(1 - p_k(x)\big) = x_{k+1}^n \mu(x, k)$$

and so

$$\mu(x, k) = \prod_{j=1}^{k} \frac{x_j^n}{1 - p_{j-1}(x)}, \qquad k = 1, \dots, d. \tag{45}$$

[6] See footnote 3.

Hence, our present notation is consistent with the definition (26). Note also that, for $z \in D$, $\mu(z,k)$ depends only on $z_1,\dots,z_k$; in particular, if $x = (z_1,\dots,z_d)$ then $\mu(z,k) = \mu(x,k)$ for all $k \le d$. Note that $\bar{b}$ is given by

$$\bar{b}_k(x) = \lambda x_{k-1}^n \mu(x,k-1) - \lambda x_k^n \mu(x,k) - (x_k - x_{k+1}), \qquad k = 1,\dots,d, \tag{46}$$

where $x_0 = 1$ and $x_{d+1} = 0$. Hence, $\bar{b} = u^{(d)}$ as defined in Section 2.2.

A comparison of (35) and (41) now shows that $\bar{b} = u^{(d)}$.

Recall that $\rho = 4/(1-\lambda)$ when $n = 1$ and that $\rho = 2^n/(1-e^{-1/2})$ when $n \ge 2$. Then for $x \in U$ and $k \ge 1$ we have

$$\begin{aligned}\mu(x,k) &= x_k^n \mu(x,k-1)/\bigl(1 - p_{k-1}(x)\bigr) \le \rho a_k^n \mu(x,k-1) \\ &\le \rho a_k^n \mu(x,k-1)/\bigl(1 - p_{k-1}(a)\bigr)\end{aligned}$$

so, for all $x \in U$ and inductively for all $k \ge 1$, we obtain

$$\mu(x,k) \le \rho^k \mu(a,k). \tag{47}$$

The following argument shows that $\mu(z,k) \le \mu(z',k)$ for all k whenever $z \le z'$. Fix $d \ge k$ and set $x' = (z'_1,\dots,z'_d)$. Assume that $x \le x'$. By a standard construction we can realize samples $V = V(x)$ and $V' = V(x')$ on a common probability space such that $V_i \le V'_i$ for all i. Then we can construct Markov chains Y and Y', having generators G_x and $G_{x'}$, respectively, on the canonical space of a marked Poisson process of rate $N\lambda$, where the marks are independent copies of (V, V'), as follows. Set $Y_0 = Y'_0 = 1$ and define recursively at each jump time T of the Poisson process $Y_T = F(Y_{T-}, V_T)$ and $Y'_T = F(Y'_{T-}, V'_T)$, where (V_T, V'_T) is the mark at time T. Then, since F is nondecreasing in both arguments, we see by induction that $Y_t \le Y'_t$ for all t. Hence, by convergence to equilibrium,

$$\mu(z,k) = \mu(x,k) = \lim_{t\to\infty} \mathbb{P}(Y_t \ge k) \le \lim_{t\to\infty} \mathbb{P}(Y'_t \ge k) = \mu(x',k) = \mu(z',k).$$

2.7. *Corrector upper bound.* We take as our reference state $\bar{y} = \min I$ and note that, under the coupling mechanism, we have $\bar{Y}_t \le Y_t$ for all t. Then the kth component of the corrector for b is given by

$$\chi_k(x,y) = \lambda \mathbb{E}_y \int_0^{T_c} \bigl(x_{k-1}^n 1_{\{\bar{Y}_s < k-1 \le Y_s\}} - x_k^n 1_{\{\bar{Y}_s < k \le Y_s\}}\bigr)\, ds,$$

so for $x \in U$ and all $y \in I$ we have

$$|\chi_k(x,y)| \le \tau x_{k-1}^n \le C a_{k-1}^n / N.$$

Now fix $y \le k-2$ and consider the stopping time $T = \inf\{t \ge 0 : Y_t = k-1\}$. Note that Y can enter state $k-1$ only from $k-2$ and does so at rate $N\lambda x_{k-1}^n$, whereas T_c occurs in state $k-2$ at rate at least $N\lambda(1-\lambda)^2/4$. Hence, $\mathbb{P}_y(T \le T_c) \le C a_{k-1}^n$ and so, by the Markov property,

$$\mathbb{E}_y \int_0^{T_c} 1_{\{\bar{Y}_s < k-1 \le Y_s\}}\, ds \le \mathbb{E}_y\bigl(1_{\{T \le T_c\}} m(x, k-1, \bar{Y}_T)\bigr) \le C a_{k-1}^n \tau \le C a_{k-1}^n / N.$$

Hence, we obtain, for $x \in U$ and all $y \in I$,

$$|\chi_k(x, y)| \le C\big(a_{k-1}^n 1_{\{y \ge k-1\}} + a_{k-1}^{2n}\big)/N. \tag{48}$$

2.8. *Quadratic variation upper bound.* The growth rate at ξ of the quadratic variation of the corrected kth coordinate is given by

$$\alpha_k(\xi) = \sum_{\xi' \ne \xi} \{\bar{\mathbf{x}}_k(\xi') - \bar{\mathbf{x}}_k(\xi)\}^2 q(\xi, \xi').$$

Recall that $\bar{x}_k = x_k - \chi_k(\mathbf{x}, y)$. We estimate separately, writing $x = \mathbf{x}(\xi)$ and $y = y(\xi)$,

$$\sum_{\xi' \ne \xi} \{\mathbf{x}_k(\xi') - x_k\}^2 q(\xi, \xi') \le N^{-2}\big(N x_k + N\lambda x_{k-1}^n 1_{\{y \ge k-1\}}\big)$$

and

$$\sum_{\xi' \ne \xi} \{\chi_k(\mathbf{x}(\xi'), y(\xi')) - \chi_k(x, y)\}^2 q(\xi, \xi') \le C N^{-1} x_{k-1}^{2n}.$$

When $y \le k-2$, we can improve the last estimate by splitting the sum in two and using

$$\sum_{\xi' \ne \xi, y(\xi') \le k-2} \{\chi_k(\mathbf{x}(\xi'), y(\xi')) - \chi_k(x, y)\}^2 q(\xi, \xi') \le C N^{-1} x_{k-1}^{4n}$$

and

$$\sum_{\xi' \ne \xi, y(\xi') \ge k-1} \{\chi_k(\mathbf{x}(\xi'), y(\xi')) - \chi_k(x, y)\}^2 q(\xi, \xi') \le C N^{-1} x_{k-1}^{3n}.$$

We used (48) for the first inequality and for the second used

$$\sum_{\xi' \ne \xi, y(\xi') \ge k-1} q(\xi, \xi') \le C N x_{k-1}^n.$$

On combining these estimates we obtain

$$\alpha_k(\xi) \le C\big(x_k + x_{k-1}^n 1_{\{y \ge k-1\}} + x_{k-1}^{3n}\big) \le a_k(y(\xi))/N,$$

where

$$a_k(y(\xi)) = C\big(a_k + a_{k-1}^n 1_{\{y \ge k-1\}}\big)/N.$$

Then, using the estimate (47) and the limit (31), we have

$$\bar{a}_k(x) = C\big(a_k + a_{k-1}^n \mu(x, k-1)\big) \le C\rho^{d-1} a_k/N \le C(\log N)^C a_k/N,$$

so we have for all $x \in U$ and $y \in I$

$$a_k(y) \le A a_k, \bar{a}_k(x) \le \bar{A} a_k$$

with $A = C/(N a_d^{1-n/\alpha})$ and $\bar{A} = C(\log N)^C/N$. It is straightforward to check that, for N sufficiently large, we have $\bar{A} \le A \le \Lambda \bar{J}^2$.

2.9. *Truncation estimates.* A specific feature of the problem we consider is that the limit dynamics is infinite dimensional, while the general fluid limit estimate applies in a finite-dimensional context. In this subsection we establish some truncation estimates which will allow us to reduce to finitely many dimensions.

Let $(z(t):t \geq 0)$ be the solution in D to $\dot{z}(t) = v(z(t))$ starting from 0, as in Theorem 2.3. Let $(x(t):t \geq 0)$ be the solution to $\dot{x}(t) = \bar{b}(x(t))$ starting from 0.

LEMMA 2.6. *We have*

$$\sum_{k=1}^{d} |z_k(t) - x_k(t)| \leq t a_{d+1}.$$

PROOF. Since $\bar{b} = u^{(d)}$, we have $x(t) = z^{(d)}(t)$ for all t, and so, from (36), we obtain

$$\begin{aligned}\sum_{k=1}^{d} |z_k(t) - x_k(t)| &\leq \sum_{k=1}^{d} \big(z_k(t) - z_k^{(d)}(t)\big) \leq \lambda \int_0^t z_d(s)^n \mu(z(t), d)\, ds \\ &\leq t\lambda a_d^n \mu(a, d) = t a_{d+1}.\end{aligned}$$

□

Denote by $A_k(t)$ the number of arrivals to queues of length at least k by time t. Note that $N Z_t^{k+1} \leq A_k(t)$ for all $k \geq 1$ and all t. Recall that

$$\tilde{a}_{d+1} = N^{-1} a_d^n + \rho^d a_{d+1}.$$

LEMMA 2.7. *There is a constant $C(\lambda, n) < \infty$ such that, for all $t \geq 0$ and all N, we have*

$$\mathbb{E}\big(A_d(T \wedge t)\big) \leq C e^{Ct} N \tilde{a}_{d+1}.$$

PROOF. Consider the function f on $U \times I$ given by $f(x, y) = 1_{\{y \geq d\}}$ and note that $\bar{f}(x) = \mu(x, d)$. Let χ be the corrector for f given by (11). Then, for all $x \in U$ and all $y \in I$,

$$|\chi(x, y)| \leq 2\tau \|f\|_\infty = 2\tau = C N^{-1}$$

and, whenever $x = \mathbf{x}(\xi)$ and $x' = \mathbf{x}(\xi')$ with $q(\xi, \xi') > 0$, by the estimates (13) and (43),

$$|\chi(x, y) - \chi(x', y)| \leq 2\nu\tau^2 \|f\|_\infty \|\mu_x - \mu_{x'}\| \leq C N^{-2}. \tag{49}$$

By optional stopping,

$$\left| \int_0^{T \wedge t} Q(\chi(\mathbf{x}, y))(X_s)\, ds \right| = \big|\mathbb{E}\big(\chi(\mathbf{X}_{T \wedge t}, Y_{T \wedge t}) - \chi(\mathbf{X}_0, Y_0)\big)\big| \leq C N^{-1}.$$

Now

$$Q(\chi(\mathbf{x}, y))(\xi) = 1_{\{y \geq d\}} - \mu(\mathbf{x}(\xi), d) - \Delta_1(\xi) - \Delta_2(\xi),$$

where Δ_1, Δ_2 are given by (23), (24). We use (44) to obtain the estimate

$$|\Delta_1(\xi)| \le 2\tau(1 + \lambda n + N\lambda n z_{d+1}) = C(z_{d+1} + N^{-1})$$

and from (49) deduce that

$$|\Delta_2(\xi)| \le N(1+\lambda)CN^{-2} = CN^{-1}.$$

So

$$\mathbb{E}\int_0^{T\wedge t} 1_{\{Y_s \ge d\}}\,ds \le CN^{-1} + \mathbb{E}\int_0^{T\wedge t}\bigl(\mu(\mathbf{X}_s, d) + CZ_s^{d+1} + CN^{-1}\bigr)\,ds.$$

Set $g(t) = \mathbb{E}(A_d(T\wedge t))$, then

$$\begin{aligned} g(t) &= N\lambda\mathbb{E}\int_0^{T\wedge t}(X_s^d)^n 1_{\{Y_s\ge d\}}\,ds \le N\lambda 2^n a_d^n \mathbb{E}\int_0^{T\wedge t} 1_{\{Y_s \ge d\}}\,ds \\ &\le Ca_d^n\biggl(1 + \int_0^t \bigl(N\rho^d\mu(a,d) + 1 + g(s)\bigr)\,ds\biggr) \\ &\le CN\tilde{a}_{d+1}(1+t) + C\int_0^t g(s)\,ds. \end{aligned}$$

Here we have used the estimate $\mu(x,d) \le \rho^d \mu(a,d)$ for $x \in U$. The claimed estimate now follows by Gronwall's lemma. $\square$

Fix $R \in (0,\infty)$ and define

$$\tilde{T} = \inf\{t \ge 0 : A_d(t) \ge RN\tilde{a}_{d+1}\} \wedge T.$$

LEMMA 2.8. *There is a constant $C(\lambda, n) < \infty$ such that, for all $t \ge 0$ and all N, we have*

$$\mathbb{E}\bigl(A_{d+1}(\tilde{T}\wedge t)\bigr) \le C(1+t)R^n\tilde{a}_{d+1}^n + CtR^{n+1}N\tilde{a}_{d+1}^{n+1}.$$

PROOF. We argue as in the preceding proof, except now taking $f(x,y) = 1_{\{y \ge d+1\}}$, for which $\bar{f}(x) = 0$. We obtain

$$\mathbb{E}\int_0^{\tilde{T}\wedge t} 1_{\{Y_s \ge d+1\}}\,ds \le CN^{-1} + C\mathbb{E}\int_0^{\tilde{T}\wedge t}(Z_s^{d+1} + CN^{-1})\,ds$$

and hence,

$$\begin{aligned} \mathbb{E}\bigl(A_{d+1}(\tilde{T}\wedge t)\bigr) &= N\lambda\mathbb{E}\int_0^{\tilde{T}\wedge t}(Z_s^{d+1})^n 1_{\{Y_s \ge d+1\}}\,ds \\ &\le N\lambda(R\tilde{a}_{d+1})^n\mathbb{E}\int_0^{\tilde{T}\wedge t} 1_{\{Y_s\ge d+1\}}\,ds \\ &\le C(1+t)R^n\tilde{a}_{d+1}^n + CtR^{n+1}N\tilde{a}_{d+1}^{n+1}. \end{aligned}$$

$\square$

2.10. *Proof of Theorem* 2.3. Recall that α is defined by (30) and that $\alpha \in (2n, 2n+1)$. Note that $\alpha^2 - (2n+1)\alpha + n = 0$. Recall that $\kappa = (2\alpha)^{-1}$ and

$$d = d(N) = \sup\{k \in \mathbb{N} : Na_k > N^\kappa\}.$$

The asymptotic growth rate (31) follows from Theorem 2.2. We shall use without further comment below the inequalities

$$C^{-1} a_k^{1/\alpha} \le a_{k+1} \le C a_k^\alpha, \qquad k \ge 0,$$

proved in Proposition 2.5 and the inequalities

$$a_{d+1} \le N^{-(1-\kappa)} \le a_d, \qquad d \le \log\log N,$$

the last being valid for all sufficiently large N.

By the truncation estimate, Lemma 2.6, we have

$$\sup_{t \le t_0} \sup_{k \le d} \frac{|z_k(t) - x_k(t)|}{\sqrt{a_k}} \le t_0 \frac{a_{d+1}}{\sqrt{a_d}} \le C t_0 a_{d+1}^{1-1/(2\alpha)} \le C t_0 N^{-1/2}.$$

Since $\phi(N) \to \infty$ as $N \to \infty$ it will therefore suffice to show (32) with $(z(t) : t \ge 0)$ replaced by $(x(t) : t \ge 0)$.

We apply the general procedure of Section 1.3. Take as norm scales $\sigma_k = \sqrt{a_k}$ so that

$$\|x\| = \max_k |x_k| / \sqrt{a_k}, \qquad x \in \mathbb{R}^d.$$

We now identify suitable regularity constants Λ, B, τ, J, $J_1(b)$, $J(\mu)$, K. We write C for a finite positive constant which may depend on λ and n and whose value may vary from line to line. We shall see that, as $N \to \infty$, the inequalities between these regularity constants required in Theorem 1.6 become valid. The maximum jump rate is bounded above by

$$\Lambda = N(1+\lambda) = CN.$$

We refer to the form of $b(x, y)$ given at (41) and note that, for $x \in U$ and $y \in I$,

$$\|b(x, y)\| \le B = 2^n a_d^{-1/2+n/\alpha} = C a_d^{-1/2+n/\alpha}.$$

We showed in Section 2.5 the following upper bound on the mean coupling time of our coupling mechanism:

$$m(x, y, \bar{y}) \le \tau = \frac{4}{N\lambda(1-\lambda)^2} = CN^{-1}.$$

We refer to Section 1.3 for the definitions of the jump bounds J, $J_1(b)$, $J(\mu)$ and leave the reader to check the validity of the following inequalities:

$$J \le N^{-1} a_d^{-1/2}, \qquad J_1(b) \le CN^{-1} a_d^{-1/2+(n-1)/\alpha}, \qquad J(\mu) \le 2nN^{-1}.$$

Recall from (46) the form of $\bar{b}$. In estimating the Lipschitz constant K for $\bar{b}$ on U, first note that, for $x \in U$ and for $j = 1, \dots, k-1$,

$$\left|\frac{\partial}{\partial x_j} x_{k-1}^n \mu(x, k-1)\right| \le C x_{k-1}^n \mu(x, k-1)(x_j^{-1} + 1).$$

Here we have used the explicit form (26) of $\mu(x, k-1)$ and the fact that $(1 - p_{j-1}(x))^{-1} \le C$ on U. Also note the inequalities

$$x_{k-1}^{2n-1}\sqrt{\frac{a_{k-1}}{a_k}} \le 2^{2n-1} a_{k-1}^{2n-1/2-\alpha/2} \le C, \qquad \sum_{j=1}^{\infty} \sqrt{a_j} \le C.$$

We find, after some further straightforward estimation, that we can take $K = C$.

Recall the choice of function ϕ in the statement of Theorem 2.3. Set

$$\varepsilon = \sqrt{\frac{\phi(N)}{N}}, \qquad \delta = \varepsilon e^{-Kt_0}/7, \qquad \delta(\beta, b) = \delta, \qquad \delta(\gamma, g) = \delta/(2\tau B).$$

Recall that $\mathbf{X}_0 = (1/N, 0, \dots, 0)$ and $x_0 = 0$ and that the driving rate ν for the coupling mechanism is equal to $N\lambda$. It is now straightforward to check that all the inequalities required in the statement of Theorem 1.5 are valid, for all sufficiently large N.

Now we check the tube condition of Theorem 1.5. The inequalities $0 \le x_d(\xi) \le \cdots \le x_1(\xi) \le 1$ hold for all $\xi \in S$. By a monotonicity property established in the proof of Theorem 2.1, we have $x_k(t) \le a_k$ for all $t \ge 0$ and for $k = 1, \dots, d$. Hence, for N sufficiently large, if $\|\mathbf{x}(\xi) - x(t)\| \le 2\varepsilon$ for some $t \ge 0$, then $x^k(\xi) \le a_k + 2\varepsilon\sqrt{a_k} \le 2a_k$ and $x^1(\xi) \le a_1 + 2\varepsilon\sqrt{a_1} \le \lambda + (1-\lambda)/2 \le (1+\lambda)/2$, so $\mathbf{x}(\xi) \in U$ and the tube condition is satisfied.

Now we turn to the extra conditions needed to apply Theorem 1.6. We noted in Section 2.8 the quadratic variation bounds

$$a_k(y) \le A\sigma_k^2, \qquad \bar{a}_k(x) \le \bar{A}\sigma_k^2,$$

valid for all $x \in U$ and $y \in I$, where

$$A = C/(N a_d^{1-n/\alpha}), \qquad \bar{A} = C(\log N)^C/N$$

and where $\bar{A} \le A \le \Lambda \bar{J}^2$ for sufficiently large N. It is now straightforward to check, also for N sufficiently large, that the remaining inequalities required in the statement of Theorem 1.6 hold. Theorem 1.6 therefore applies to give

$$\begin{aligned}\mathbb{P}\Big(\sup_{t \le t_0} \|\mathbf{X}_t - x_t\| > \varepsilon\Big) &\le 2d e^{-\delta^2/(4\bar{A}t_0)} + 2d e^{-(\bar{A}/A)^2 t_0/(6400\Lambda\tau^2)} \\ &\quad + \mathbb{P}\big(\Omega(\beta, b)^c \cup \Omega(\gamma, g)^c\big).\end{aligned} \tag{50}$$

Now, for N sufficiently large, we have $d \le \log\log N$ and, by our choice of ϕ and κ,

$$\delta^2/(4\bar{A}t_0) \ge \phi(N)/((\log N)^C t_0) \ge \log N$$

and

$$(\bar{A}/A)^2 t_0/(6400\Lambda\tau^2) \geq \log N.$$

Hence, the first and second terms on the right-hand side of (50) tend to 0 as $N \to \infty$.

Recall from (16) and (17) the form of the events $\Omega(\beta, b)$ and $\Omega(\gamma, g)$. In the present example, the complementary exceptional events arise either as a result of truncation or because of finite N effects in the fast variable dynamics, as shown by (42) and estimate (44). Recall that $\delta(\beta, b) = \delta$ and $\delta(\gamma, g) = \delta/(2\tau B)$. Then

$$\Omega(\beta, b)^c \subseteq \left\{ \int_0^{T\wedge t_0} \frac{Z_t^{d+1}}{\sqrt{a_d}}\, dt \geq \delta(\beta, b) \right\} \subseteq \left\{ A_d(T \wedge t_0) \geq \frac{N\delta\sqrt{a_d}}{t_0} \right\}. \tag{51}$$

It is straightforward to check that, for all sufficiently large N, $\delta(\gamma, g) \geq 2t_0(1 + \lambda n)$, which implies that

$$\begin{aligned} \Omega(\gamma, g)^c &\subseteq \left\{ \int_0^{T\wedge t_0} (1 + \lambda n + N\lambda n Z_t^{d+1})\, dt \geq \delta(\gamma, g) \right\} \\ &\subseteq \left\{ A_d(T \wedge t_0) \geq \frac{\delta}{4\lambda n t_0 \tau B} \right\}. \end{aligned} \tag{52}$$

To see that $\mathbb{P}(\Omega(\beta, b)^c \cup \Omega(\gamma, g)^c) \to 0$ as $N \to \infty$, we use the bound on $\mathbb{E}(A_d(T \wedge t_0))$ proved in Lemma 2.7 and Markov's inequality. It then suffices to show that in the limit $N \to \infty$,

$$C(a_d^n + \rho^d N a_{d+1}) e^{Ct_0} \ll \frac{\sqrt{\phi(N) N a_d}}{4nt_0}.$$

For the term involving a_d^n this is easy. For the other term, involving Na_{d+1}, we can check that, in fact,

$$Na_{d+1} \ll \sqrt{Na_d}, \qquad \rho^d \leq (\log N)^C \ll \sqrt{\phi(N)}.$$

This completes the proof of (32). Limit (33) follows immediately from Lemma 2.7 using Markov's inequality. Finally, note that, as $N \to \infty$,

$$\tilde{a}_{d+1} \leq CN^{-1} a_d^n + (\log N)^C a_{d+1} \leq C\bigl(N^{-1} + (\log N)^C N^{-1+\kappa}\bigr) \to 0$$

and

$$N\tilde{a}_{d+1}^{n+1} \leq C\bigl(N^{-(1-\kappa)(n+1)/\alpha} + (\log N)^C N^{1-(1-\kappa)(n+1)}\bigr) \to 0.$$

Then the limit (34) follows from (33) and Lemma 2.8 using Markov's inequality.

2.11. *Monotonicity of the queueing model.* Here we prove a natural monotonicity property of the supermarket model with memory which is a microscopic counterpart of the monotonicity of solutions to the differential equation (28) shown in Theorem 2.1. We do not rely on this result in the rest of the paper.

First we construct, on a single probability space, for all $\xi = (z, y) \in S$, a version $X = X(\xi)$ of the supermarket model with memory starting from ξ. Set $y_1 = y_1(\xi) = y$ and determine $y_i = y_i(\xi) \in \mathbb{Z}^+$ for $i = 2, \dots, N$ by the conditions

$$y_2 \le \cdots \le y_N, \qquad z_k = \bigl|\{i \in \{1, \dots, N\} : y_i \ge k\}\bigr|/N, \qquad k \in \mathbb{N}.$$

We work on the canonical space of a marked Poisson process of rate $N(1 + \lambda)$, where the marks are either, with probability $1/(1 + \lambda)$, independent copies of a uniform random variable J in $\{1, \dots, N\}$ or, with probability $\lambda/(1 + \lambda)$, independent copies of a uniform random sample $(J_1, \dots, J_n)$ from $\{1, \dots, N\}$. Fix $\xi = (z, y) \in S$ and define a process $X = X(\xi) = (X_t : t \ge 0)$ in S as follows. Set $X_t = \xi$ for all $t < T$, where T is the first jump time of the Poisson process. If the first mark is a random variable, J say, take the sequence $y_1, \dots, y_N$ and replace y_J by $(y_J - 1)^+$ to obtain a sequence $u_1, \dots, u_N$ say; set $\tilde{y}_1 = u_1$ and write $u_2, \dots, u_N$ in nondecreasing order to obtain $\tilde{y}_2 \le \cdots \le \tilde{y}_N$. If the first mark is a random sample, $(J_1, \dots, J_n)$ say, select components $(y_i : i \in \{1, J_1, \dots, J_n\})$ and write these in nondecreasing order, $w_1 \le \cdots \le w_m$ say; replace w_1 by $w_1 + 1$ and write the resulting sequence, again in nondecreasing order, $v_1 \le \cdots \le v_m$ say; set $\tilde{y}_1 = v_1$ and write $v_2, \dots, v_m$ combined with the unselected components $(y_i : i \notin \{1, J_1, \dots, J_n\})$ in nondecreasing order to obtain $\tilde{y}_2 \le \cdots \le \tilde{y}_N$. Set $X_T = ((Z_T^k : k \in \mathbb{N}), Y_T)$, where

$$Z_T^k = \bigl|\{i \in \{1, \dots, N\} : \tilde{y}_i \ge k\}\bigr|/N, \qquad k \in \mathbb{N}, \qquad Y_T = \tilde{y}_1,$$

and repeat the construction from X_T in the usual way.

For $\xi, \xi' \in S$ write $\xi \le \xi'$ if $y_i(\xi) \le y_i(\xi')$ for $i = 1, \dots, N$.

THEOREM 2.9. *Let $\xi, \xi' \in S$ with $\xi \le \xi'$. Then $X_t(\xi) \le X_t(\xi')$ for all $t \ge 0$.*

PROOF. It will suffice to check that the desired inequality holds at the first jump time T, that is to say, with obvious notation, that $\tilde{y}_i \le \tilde{y}_i'$ for all i. Note that if $a_i \le b_i$ for all i for two sequences $(a_1, \dots, a_n)$ and $(b_1, \dots, b_n)$, then the same is true for their nondecreasing rearrangements. In the case where the first mark is a random variable J, since $y_i \le y_i'$ for all i, we have $u_i \le u_i'$ for all i and so $\tilde{y}_i \le \tilde{y}_i'$ for all i. On the other hand, when the first mark is a random sample $(J_1, \dots, J_n)$, we have $w_j \le w_j'$ for all j, so $v_j \le v_j'$ for all j, and so $\tilde{y}_i \le \tilde{y}_i'$ for all i. □

REFERENCES

[1] BALL, K., KURTZ, T. G., POPOVIC, L. and REMPALA, G. (2006). Asymptotic analysis of multiscale approximations to reaction networks. *Ann. Appl. Probab.* **16** 1925–1961. MR2288709

[2] DARLING, R. W. R. and NORRIS, J. R. (2008). Differential equation approximations for Markov chains. *Probab. Surv.* **5** 37–79. MR2395153

[3] GRAHAM, C. (2000). Chaoticity on path space for a queueing network with selection of the shortest queue among several. *J. Appl. Probab.* **37** 198–211. MR1761670

[4] GRAHAM, C. (2005). Functional central limit theorems for a large network in which customers join the shortest of several queues. *Probab. Theory Related Fields* **131** 97–120. MR2105045

[5] LUCZAK, M. J. and MCDIARMID, C. (2006). On the maximum queue length in the supermarket model. *Ann. Probab.* **34** 493–527. MR2223949

[6] LUCZAK, M. J. and MCDIARMID, C. (2007). Asymptotic distributions and chaos for the supermarket model. *Electron. J. Probab.* **12** 75–99. MR2280259

[7] LUCZAK, M. J. and NORRIS, J. (2005). Strong approximation for the supermarket model. *Ann. Appl. Probab.* **15** 2038–2061. MR2152252

[8] MITZENMACHER, M., PRABHAKAR, B. and SHAH, D. (2002). Load balancing with memory. In *Proc. 43rd Ann. Symp. Found. Comp. Sci.* 799–808. IEEE, Los Alamitos, CA.

[9] VVEDENSKAYA, N., DOBRUSHIN, R. L. and KARPELEVICH, F. I. (1996). A queueing system with a choice of the shorter of two queuesŮan asymptotic approach. (Russian) *Problemy Peredachi Informatsii* **32** 20–34; translation in *Probl. Inf. Transm.* **32** 15–27. MR1384927

SCHOOL OF MATHEMATICAL SCIENCES
QUEEN MARY, UNIVERSITY OF LONDON
MILE END ROAD
LONDON E1 4NS
UNITED KINGDOM
E-MAIL: m.luczak@qmul.ac.uk

STATISTICAL LABORATORY
CENTRE FOR MATHEMATICAL SCIENCES
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD
CAMBRIDGE, CB3 0WB
UNITED KINGDOM
E-MAIL: j.r.norris@statslab.cam.ac.uk