STATE-DEPENDENT QUEUES: APPROXIMATIONS AND APPLICATIONS

AVI MANDELBAUM* AND GENNADY PATS*

Abstract. A state-dependent queue is an exponential service system, where arrival and service rates depend on queue length. For properly normalized queueing processes, we derive functional strong laws of large numbers and functional central limit theorems. The former support fluid approximations and the latter diffusion refinements. Our analysis is based on strong approximations, which provide a unified framework for most existing approximations of state-dependent queues.

Key words. State–Dependent Queues, Strong Approximations, Fluid and Diffusion Approximations, Many–Server and Finite–Populations Queues.

1. Introduction. State–dependent exponential $M_\xi/M_\xi/1$ queues are models in which arrival and service rates depend on the state ξ—the queue length. For properly normalized queueing processes, we derive functional strong law of large numbers (FSLLN, Theorem 4.1) and functional central limit theorems (FCLT, Theorems 4.2 and 4.3). The former support fluid approximations and the latter diffusion refinements. The current analysis is a first step in an ongoing effort to cover queueing networks.

The strong limit in FSLLN (henceforth called the *fluid limit*) is the unique solution to an autonomous first–order ordinary differential equation with reflection. In such an equation, the derivative depends explicitly only on state. Consequently, the fluid limit of a single queue is a monotone continuous function, which absorbs at zero if it ever reaches it.

The weak limit given in FCLT (henceforth called the *diffusion limit*) is the unique strong solution to a stochastic differential equation with a certain type of reflection. The diffusion limits are Markov processes with upper semi-continuous sample-paths. Weak convergence is with respect to Skorokhod's M_1-*topology* (see Appendix B and the discussion in Subsection 4.5).

Our technique for obtaining limit theorems is based on strong approximations. It is similar to Kurtz [31], who considers density-dependent population processes, for which limits do not involve the reflection phenomenon (see also Ethier and Kurtz [14, Chapter 11 §2,3]). It differs from Kurtz [32],[33] and [34], that relies on multiparameter time transformations.

In Section 5, derivations of many available fluid and diffusion approximations for state-dependent queues are unified. Examples covered are models with reneging, finite population and finite or infinite number of servers. We are not assuming boundedness of arrival rates, service rates or

* Technion Institute, Haifa, 32000 Israel. Avi Mandelbaum was partially supported by the Fund for the Promotion of Research at the Technion.

populations, at the expense of some additional technicalities in our proofs.

Pioneering works on fluid and diffusion approximations for queues are Oliver and Samuel [39], Newell (e.g., [38]), Kingman (e.g., [24],[25]), Borovkov [5],[6], and Iglehart and Whitt [20],[21]. For later advances, readers are referred to the following survey papers and references therein: Whitt [46], covering the period up to 1974, Lemoine [35], up to 1978, Coffman and Reiman [12], through 1984, Glynn [15], through 1990 and finally Chen and Mandelbaum [9],[10], up to 1992. Related recent research is Anulova [2] and Krichagina [30], who take a martingale approach to cover, as far as the single queue is concerned, special cases of our models. Additional representative examples of martingale-based fluid and diffusion approximations are Kogan *et al.* [28] and Kogan and Liptser [27], where certain types of closed exponential networks with state–dependent service are treated. Fluid approximations for state- and time-dependent queueing networks are described in [9]. Our analysis resembles Mandelbaum and Massey [37], who establish strong approximations, FSLLN and FCLT for the time-dependent $M_t/M_t/1$ system. The similarity is mainly a consequence of the fact that, in both models, diffusion approximations enjoy time-dependent drifts and variances (see (4.9)).

We use the model of an $M_\xi/M_\xi/1$ queue with the so-called autonomous server (Borovkov [5]). This means, roughly speaking, that the server is working permanently while actual departures are generated only when the system is not empty. (For an additional discussion see, e.g., Iglehart and Whitt [20]). Our mathematical formulation (equations (2.1)–(2.4)) are as in Prabhu [41] and Bremaud [7], but we focus on approximations rather than exact analysis. Fluid and diffusion approximations for state-*independent* systems have been commonly analyzed within the framework of "non-autonomous" server models (see the survey papers mentioned above). Difficulties, however, arise even in the mere interpretation of state-*dependent* non-autonomous queues, so this will not be pursued any further.

The $M_\xi/M_\xi/1$ queue is, in fact, a one-dimensional birth and death process (see Subsection 2.2). As such, it has been amply covered and a broad spectrum of (mainly elementary) tools is available and sufficient for its analysis. Here, however, we are concerned with transient evolution, and this seems challenging enough to deserve the analysis that follows. Also provided is a framework for most existing approximations of state-dependent queues, including stationary distributions when they exist. (Extensions to queueing networks, namely multi-dimensional birth and death processes, are currently being developed.)

The remainder of the paper is organized as follows. In Section 2 we present our model of the $M_\xi/M_\xi/1$ queue and discuss different representations of its queueing process. Section 3 deals with reflection maps, that characterize subsequent fluid and diffusion limits. In Section 4 we outline FSLLN and FCLT. Section 5 is devoted to applications of our results. Proofs of the main theorems are provided in Sections 6 and 7. In sec-

tion 8 we outline directions for future research. Technical background on Skorokhod's reflection problem and on M_1-convergence is presented in Appendices A and B, while the main notation is summarized in Appendix C.

2. The model of the $M_\xi/M_\xi/1$ queue. The subject of our study is the state-dependent $M_\xi/M_\xi/1$ queue. We analyze its *queueing process* $Q = \{Q_t\,, t \geq 0\}$, whose value at time t, Q_t, describes the total number of customers, waiting or being served at that time. A formal pathwise construction of Q is an outcome of the observation that there exists a unique stochastic process Q, satisfying the following relations at all $t \geq 0$:

$$Q_t = Q_0 + A_t - D_t\,, \tag{2.1}$$

$$A_t = N_+\left(\int_0^t \lambda(Q_u)\,du\right), \tag{2.2}$$

$$D_t = \int_0^t 1[Q_{u-}>0]\,dS_u\,, \tag{2.3}$$

$$S_t = N_-\left(\int_0^t \mu(Q_u)\,du\right). \tag{2.4}$$

Here Q is constructed in terms of the following primitives:

Q_0 is a nonnegative random variable,

$\lambda\,, \mu$ are nonnegative locally Lipschitz functions on $[0, \infty)$,

$N_+\,, N_-$ are standard (rate 1) Poisson processes.

(The path construction is straightforward and of no significance to later development, hence it is omitted). The random Q_0, $N_+\,, N_-$ are defined on a common probability space and are assumed to be independent. The entities involved in the construction have the following interpretation: Q_0 is an initial queue; $A = \{A_t\,, t \geq 0\}$ and $D = \{D_t\,, t \geq 0\}$ are RCLL point processes— A_t and D_t represent the cumulative number of arrivals and departures during $(0, t]$ respectively; finally, $\lambda(Q)$ and $\mu(Q)$ are, respectively, instantaneous arrival and service rates while at state Q. Equations (2.3) and (2.4) indicate that there are no departures when customers are absent. Thus, $S = \{S_t\,, t \geq 0\}$ represents a *potential* for departures, which is fully realized only when $Q > 0$.

Remark 2.1. The sample–paths of Q are piecewise constant RCLL functions. If Q is non–explosive, that is $\mathrm{P}\{Q_t < \infty, \forall t \geq 0\} = 1$, then $D[0, \infty)$ is a suitable space for sample–paths. Otherwise, a one–point compactification of $\mathbb{R}$ can be used, with λ and μ appropriately modified. A simple sufficient condition, that ensures non-explosion of Q, is a linear growth constraint on λ:

$$\lambda(\xi) \leq K(1 + \xi), \quad \xi \geq 0, \tag{2.5}$$

for some constant $K > 0$, The limit theorems in this paper are stated for non-explosive processes. (Generalizations are only commented on; see Proposition 3.1 and Remark 4.1.) □

2.1. Representation in terms of reflection. We recast equations (2.1)–(2.4) in a form that is convenient for analysis, namely the reflection problem described in Appendix A:

$$\text{(2.6)} \qquad \begin{cases} Q_t = X_t + Y_t \geq 0, \quad t \geq 0, \\ Y \text{ nondecreasing, } Y_0 = 0, \\ \int_0^\infty 1[Q_t>0]\, dY_t = 0, \end{cases}$$

where

$$\text{(2.7)} \qquad X_t = Q_0 + N_+\left(\int_0^t \lambda(Q_u)\, du\right) - N_-\left(\int_0^t \mu(Q_u)\, du\right),$$

$$\text{(2.8)} \qquad Y_t = \int_0^t 1[Q_{u-}=0]\, dS_u\,.$$

The process Y represents cumulative losses of potential departures, due to server idleness.

Substituting X and Y into (2.6) and comparing the result with definition (2.1) of Q reveals that only the last equation of (2.6) requires verification. This equation is a complementarity relation between Y and Q: Y_t increases at time t only if $Q_t = 0$. By (2.8), it is equivalent to

$$\text{(2.9)} \qquad \int_0^\infty 1[Q_t>0]\, 1[Q_{t-}=0]\, dS_t = 0,$$

whose verification we now outline. Assume, to the contrary, that (2.9) is not satisfied or, equivalently, that for some $t > 0$: $Q_{t-} = 0$, $Q_t > 0$, and $S_{t-} \neq S_t$, that is $S_t = S_{t-} + 1$. In words, the following two events occur at time t: first, a customer arrives to an *empty* system—A jumps; second, a potential service is completed—S jumps. However, as long as $Q = 0$, A and S evolve like independent Poisson processes with intensities $\lambda(0)$ and $\mu(0)$ respectively (see (2.2) and (2.4)). Such processes a.s. do not jump simultaneously, hence (2.9) a.s. prevails.

Remark 2.2. Equations (2.6) differ from the standard Skorokhod's reflection problem in that here, X itself depends on Q. Nevertheless, it turns out useful that

$$Q = \Phi(X), \quad Y = \Psi(X),$$

where Φ and Ψ are the Lipschitz operators in Appendix A, and X is given by (2.7). □

2.2. Representation as a birth and death process. The distribution of Q is the same as that of a birth and death process on the integers,

starting at Q_0 and evolving according to the following transition rates:

$$\begin{cases} q_{k,k+1} = \lambda(k), & k = 0, 1, \ldots, \\ q_{k,k-1} = \mu(k), & k = 1, 2, \ldots, \\ q_{0,-1} = 0; \end{cases}$$

(Theorem 4.1 of Chapter 6 in Ethier and Kurtz [14]). In particular, the effective service rate at a time $t \geq 0$ is $\mu_{eff}(Q_t) = 1[Q_t > 0]\mu(Q_t)$.

3. Reflection problems. In this section, we introduce two reflection problems that provide the mathematical framework for our main theorems.

3.1. A differential equation with reflection. Consider the following problem: given q_0—a nonnegative number, and θ—a locally Lipschitz function on $[0, \infty)$, find a pair (q, y) of absolutely continuous functions such that

$$\text{(3.1)} \qquad \begin{cases} q_t = q_0 + \int_0^t \theta(q_s)\, ds + y_t \geq 0, \quad t \geq 0, \\ y \text{ nondecreasing}, \quad y_0 = 0, \\ \int_0^\infty 1[q_t > 0]\, dy_t = 0. \end{cases}$$

Remark 3.1. Analogously to Remark 2.2, (3.1) can be rewritten as

$$q = \Phi(x), \quad y = \Psi(x),$$

where

$$x_\bullet = q_0 + \int_0^\bullet \theta(q_s)\, ds,$$

and Φ, Ψ are the reflection operators from Appendix A. □

Existence, uniqueness and some properties of the solution to (3.1) are given by

PROPOSITION 3.1. *If θ is locally Lipschitz then there exists a unique solution (q, y) to (3.1). For this solution, q is a monotone function and it is non–explosive if and only if at least one of the following two conditions is satisfied:*

$$\theta(\xi_1) \leq 0, \quad \textit{for some } \xi_1 \geq q_0\,,$$

or

$$\int_{q_0}^\infty \frac{1}{\theta(s)}\, ds = \infty.$$

Remark 3.2. If $\theta(\xi) > 0$ for all $\xi \geq q_0$, then a linear growth of θ over $[q_0\,, \infty)$ (as in (2.5)) suffices for the second (integral) condition. □

Outline of the Proof. Uniqueness follows from the Lipschitz properties of θ, Φ, and Ψ. To prove existence, associate with (3.1) the following ordinary differential equation:

$$\dot{z}_t = \theta(z_t), \quad z_0 = q_0 \,. \tag{3.2}$$

When θ is locally Lipschitz, such an equation has a unique solution up to (a possible) *explosion time* [16]. This solution must be either strictly monotone or a constant function [16, page 40]. It gives rise to the unique solution of (3.1) in the following manner: q coincides with z up to the time $t > 0$ when z intersects zero, past which q vanishes. (If z is non-negative on $(0, \infty)$ then $q \equiv z$.) The first (non-positivity) condition of the theorem ensures that q remains bounded, and the second (integral)—that q approaches infinity only at infinite time. □

To support later analysis, we now elaborate on the explicit forms that solutions to (3.1) can take. They are described in the following four Cases:

1. Strictly positive

1.1 *Strictly increasing.*
If $\theta(\xi) > 0$ for all $\xi \geq q_0$, then

$$\dot{q}_t = \theta(q_t),\ t > 0; \quad q_t \uparrow\uparrow \infty; \quad y \equiv 0.$$

1.2 *Strictly increasing with horizontal asymptote.*
If there exists $\xi_1 > q_0$ such that

$$\theta(\xi_1) = 0 \text{ and } \theta(\xi) > 0,\ \xi \in [q_0\,, \xi_1),$$

then

$$\dot{q}_t = \theta(q_t),\ t > 0; \quad q_t \uparrow\uparrow \xi_1\,; \quad y \equiv 0.$$

1.3 *Strictly decreasing with horizontal asymptote.*
If $q_0 > 0$, and there exists $\xi_1 \in [0, q_0)$, such that

$$\theta(\xi_1) = 0 \text{ and } \theta(\xi) < 0, \quad \xi \in (\xi_1\,, q_0],$$

then

$$\dot{q}_t = \theta(q_t),\ t > 0; \quad q_t \downarrow\downarrow \xi_1\,; \quad y \equiv 0.$$

1.4 *Non-zero constant.*
If $q_0 > 0$ and $\theta(q_0) = 0$ then

$$q \equiv q_0 \quad y \equiv 0.$$

2. Vanishing, without reflection
If $q_0 = 0$ and $\theta(0) = 0$ then

$$q \equiv 0, \quad y \equiv 0.$$

3. Vanishing, with reflection

If $q_0 = 0$ and $\theta(0) < 0$ then

$$q \equiv 0, \quad y_t = -\theta(0)t, \quad t \geq 0.$$

4. Strictly decreasing and absorbing at zero

If $q_0 > 0$ and $\theta(\xi) < 0$ for all $\xi \in [0, q_0]$ then

$$\begin{cases} \dot{q}_t = \theta(q_t), & t \in [0, t_0], \\ q_t = 0, & t \geq t_0\,; \end{cases}$$

$$\begin{cases} y_t = 0, & t \in [0, t_0], \\ y_t = -\theta(0)(t - t_0), & t \geq t_0\,, \end{cases}$$

where $t_0 = \inf\{t \geq 0 : q_t = 0\}$, $0 < t_0 < \infty$.

Remark. Explosion can occur only in Case 1.1, otherwise q is bounded. Furthermore, q does not leave zero after reaching it. □

3.2. Derivatives of reflection operators. For a background and references on M_1-convergence see Appendix B. Notations are summarized in Appendix C. All functions below are defined on $[0, \infty)$. The following lemma plays a key role in our later formulation and proof of FCLT.

LEMMA 3.2. *Let b and x be continuous functions. Assume that $x_0 \geq 0$ and that x is either strictly monotone or constant. Suppose further that $b_0 \geq 0$ if $x_0 = 0$. Then, the sequence of continuous functions, given by*

$$\Psi(nx + b) - \Psi(nx) = \overline{(nx + b)^-} - n\overline{x^-}, \quad n = 1, 2, \ldots,$$

decreases monotonically, as $n \uparrow \infty$, to an upper semi-continuous function $\widetilde{b}$. This convergence holds in the M_1-topology.

The proof is omitted as it resembles that of Lemma 4.2 in [37].

For each x satisfying the conditions of Lemma 3.2, denote by $C^x[0, \infty)$ the set of continuous functions

$$C^x[0, \infty) \triangleq \begin{cases} C[0, \infty), & x_0 > 0, \\ C_0[0, \infty), & x_0 = 0. \end{cases} \tag{3.3}$$

Introduce the operators Ψ^x and Φ^x, with domain $C^x[0, \infty)$, by

$$\Psi^x(b) \triangleq \widetilde{b}, \quad \Phi^x(b) \triangleq b + \Psi^x(b). \tag{3.4}$$

The notation $\Phi^x(b)$ is justified in view of the M_1-convergence

$$\Phi(nx + b) - \Phi(nx) = b + \Psi(nx + b) - \Psi(nx) \longrightarrow b + \Psi^x(b), \tag{3.5}$$

which prevails by the continuity of b and the continuity of addition in the M_1-topology (See Appendix B).

To justify the title of the current subsection, note that Lemma 3.2 can be stated as follows:

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} [\Psi(x + \varepsilon b) - \Psi(x)] = \Psi^x(b),$$

in the M_1-topology. Thus, $\Psi^x(b)$ can be interpreted as some form of a directional derivative of the operator $x \longrightarrow \Psi(x)$, at the point x in the direction b. Analogously, $\Phi^x(b)$ is a directional derivative of the operator $x \longrightarrow \Phi(x)$.

The transformation Φ^x is central for our results. We now elaborate on its explicit forms, recalling that its domain is $C^x[0, \infty)$ for those x that satisfy the conditions of Lemma 3.2. The following four Cases arise:

1. Identity operator

If x is strictly positive over $(0, \infty)$, then

$$\Phi^x(b) \equiv b.$$

2. Ordinary reflection operator

If x is identically zero, then

$$\Phi^x(b) = \Phi(b).$$

3. Delayed zero operator

If x is strictly decreasing with $x_0 = 0$, then

$$\Phi^x_t(b) = \begin{cases} b_0\,, & t = 0, \\ 0, & t > 0. \end{cases}$$

4. Restricted identity operator

If x is strictly decreasing with $x_0 > 0$, and if $x(t_0) = 0$ for some $t_0 \in (0, \infty)$, then

$$\Phi^x_t(b) = \begin{cases} b_t\,, & t < t_0\,, \\ 0, & t > t_0\,, \\ 0 \vee b_{t_0}\,, & t = t_0\,. \end{cases}$$

Remark 3.3. If $\Psi^x(b)$ is a continuous function at some point b, then the M_1-convergence in Lemma 3.2 reduces to U-convergence. A similar assertion holds with respect to Φ^x and the convergence in (3.5). Consequently, in Cases 1 and 2 the convergence in (3.5) is uniform on compact subsets of $[0, \infty)$. In Case 3, the convergence is uniform on compact subsets of $(0, \infty)$ if $b_0 \neq 0$, and of $[0, \infty)$ otherwise. In Case 4, the convergence is in $(\widetilde{D}[0, \infty), M_1)$, and the values of Φ^x are upper semi-continuous functions. Furthermore, in Case 4 when $b_{t_0} < 0$ (respectively $b_{t_0} > 0$), the convergence is uniform on compact subsets of $[0, t_0)$ and $[t_0\,, \infty)$ (respectively $[0, t_0]$ and $(t_0\,, \infty)$). If $b_{t_0} = 0$, the convergence is uniform on compact subsets of $[0, \infty)$. □

The following explicit expression applies to Φ^x (by analogy to (4.5),(4.6) in [37]):

$$\Phi^x(b) = \sup_{s \in \widehat{H}_t} (-\widehat{b}_s), \quad t \geq 0,$$

where

$$\widehat{H_t} \equiv \left\{ 0 \leq s \leq t \mid x_s^- = \sup_{0 \leq u \leq t} x_u^- \right\},$$

$$\widehat{b_t} = \begin{cases} b_t, & x_t < 0, \\ b_t \wedge 0, & x_t = 0, \\ 0, & x_t > 0. \end{cases}$$

Such a representation is expected to be useful for the analysis of queues that are both time and state dependent.

4. Main theorems. FSLLN and FCLT are presented in Subsection 4.1 and 4.2 respectively. A refinement of FCLT, useful in applications, is formulated in Subsection 4.3. In Subsection 4.4 we analyze the rescaling procedure that lead to our limit theorems. The subject of Section 4.5 is an interpretation of discontinuous diffusion limits. We conclude the section with alternative types of rescaling. This motivates a later discussion, in Subsection 4.6, of models that are not covered in the current paper.

4.1. Fluid approximations (FSLLN). Consider a sequence $M_\xi^n/M_\xi^n/1$, $n = 1, 2, \ldots$, of queueing systems, each as in (2.6)–(2.8). The n-th system is described in terms of the following primitives: a random variable Q_0^n representing the initial queue, and non-negative locally Lipschitz functions λ^n and μ^n defining, respectively, the dependence of the arrival and service rates on the queue length Q^n. The queueing process Q^n can be realized as the unique solution to the following reflection problem (see Remark 2.2):

$$(4.1) \quad \begin{cases} Q^n = \Phi(X^n), \\ X_\bullet^n = Q_0^n + N_+\left(\int_0^\bullet \lambda^n\left(Q_s^n\right) ds\right) - N_-\left(\int_0^\bullet \mu^n\left(Q_s^n\right) ds\right). \end{cases}$$

Introduce the rescaled processes $q^n = \{q_t^n, t \geq 0\}$ given by

$$(4.2) \qquad q_t^n = \frac{1}{n} Q_t^n.$$

Then, due to the homogeneity of Φ and Ψ (Appendix A),

$$(4.3) \quad \begin{cases} q^n = \Phi(x^n), \\ x_\bullet^n = q_0^n + \frac{1}{n} N_+\left(\int_0^\bullet \lambda^n\left(nq_s^n\right) ds\right) - \frac{1}{n} N_-\left(\int_0^\bullet \mu^n\left(nq_s^n\right) ds\right). \end{cases}$$

The asymptotic behavior of $\{q^n\}$ emerges from the following theorem, the proof of which is postponed to Section 6.

THEOREM 4.1 (FSLLN). *Suppose that*

$$\frac{1}{n}\lambda^n(n\xi) \longrightarrow \lambda(\xi) \quad \text{and} \quad \frac{1}{n}\mu^n(n\xi) \longrightarrow \mu(\xi), \quad \text{u.o.c.}, \tag{4.4}$$

as $n \uparrow \infty$, where λ and μ are given locally Lipschitz functions, as well as

- $\frac{1}{n}\lambda^n(n\xi) \leq K(1+\xi)$, $\xi \geq 0$, *where K is a given positive constant;*
- $\lim_{n\uparrow\infty} q_0^n = q_0$ a.s., *where q_0 is a given non–negative scalar, and the sequence $\{Eq_0^n\}$ of expectations is uniformly bounded.*

Then, as $n \uparrow \infty$, the sequence $\{q^n\}$ of solutions to (4.3) converges u.o.c. over $[0,\infty)$, a.s., to a deterministic function q, given by

$$\begin{cases} q = \Phi(x), \\ x_\bullet = q_0 + \int_0^\bullet (\lambda(q_s) - \mu(q_s))\, ds. \end{cases} \tag{4.5}$$

That is, q is the unique solution to the differential equation with reflection (3.1), with

$$\theta(\xi) = \lambda(\xi) - \mu(\xi), \quad \xi \geq 0. \tag{4.6}$$

In what follows, q will be referred to as the *fluid limit* associated with the queueing sequence under consideration. An analogous result holds for the sequence $\{y^n\}$, that is associated with losses of potential departures due to idleness. Specifically, for

$$y^n = \frac{1}{n}\Psi(X^n),$$

with X^n as in (4.1), we have $y^n \longrightarrow y = \Psi(x)$, a.s., u.o.c., where x is as in (4.5).

Remark 4.1. The growth condition imposed on λ^n ensures non–explosion of q^n and q. We believe, however, that FSLLN can be generalized to cover cases when q^n and/or q are explosive (see Remarks 2.1, 3.2). The theorem ought then to remain valid over the domain of existence of q. In particular, Theorem 4.1 ought to hold over $[0,\infty)$ when the linear growth constraint on λ^n is replaced by any condition that ensures non-explosion of q. Necessary and sufficient conditions for q to be non-explosive are given by Proposition 3.1. An example of a limit theorem that gives rise to explosive processes is Barbour [3]. □

The forms of the solutions to (3.1), listed at the end of Subsection 3.1, characterize possible fluid limits which, in turn, identify four modes of operation for the $M_\xi/M_\xi/1$ queue. They are depicted in Figure 1 and described by the following four Cases (based on (4.6)):

1. Permanent large queues

1.1 *Overloaded:* $\lambda(\xi) > \mu(\xi)$ for all $\xi \geq q_0$.

1.2 *Overloaded, with asymptotic transition to critically loaded:* there exists $\xi_1 > q_0$ such that

$$\lambda(\xi_1) = \mu(\xi_1); \quad \lambda(\xi) > \mu(\xi), \quad \xi \in [q_0, \xi_1).$$

1.3 *Underloaded, with large initial queue and asymptotic transition to critically loaded:* $q_0 > 0$, and there exists $\xi_1 \in [0, q_0)$, such that

$$\lambda(\xi_1) = \mu(\xi_1); \quad \lambda(\xi) < \mu(\xi), \quad \xi \in (\xi_1, q_0].$$

1.4 *Critically loaded with large initial queue:* $q_0 > 0$, $\lambda(q_0) = \mu(q_0)$.

2. Critically loaded: $q_0 = 0$ and $\lambda(0) = \mu(0)$.

3. Underloaded: $q_0 = 0$ and $\lambda(0) < \mu(0)$.

4. Underloaded with large initial queue: $q_0 > 0$ and $\lambda(\xi) < \mu(\xi)$, $\xi \in [0, q_0]$.

4.2. Diffusion approximations (FCLT). Introduce the sequences of stochastic processes $V^n = \{V_t^n, t \geq 0\}$, $n = 1, 2, \ldots$, given by

$$V_t^n = \sqrt{n}\,(q_t^n - q_t), \quad t \geq 0. \tag{4.7}$$

This sequence amplifies deviations of the rescaled queueing processes q^n from their fluid limit q. The asymptotic behavior of $\{V^n\}$ is the subject of the next theorem, the proof of which is presented in Section 7.

THEOREM 4.2 (FCLT). *Let the conditions of Theorem 4.1 (FSLLN) be satisfied. Assume further that λ, μ in (4.4) are continuously differentiable with locally Lipschitz derivatives,*

$$\begin{cases} \sqrt{n}\left[\frac{\lambda^n(n\xi)}{n} - \lambda(\xi)\right] \longrightarrow f_\lambda(\xi), & \text{u.o.c.}, \\ \sqrt{n}\left[\frac{\mu^n(n\xi)}{n} - \mu(\xi)\right] \longrightarrow f_\mu(\xi), & \text{u.o.c.}, \end{cases} \tag{4.8}$$

where f_λ, f_μ are locally Lipschitz functions, and that $V_0^n \xrightarrow{\mathrm{d}} V_0$, as $n \uparrow \infty$, where V_0 is a given random variable.
Then the sequence $\{V^n\}$ converges weakly in $(\widetilde{D}[0,\infty), M_1)$ to a Markov process V with upper semi-continuous sample-paths. The process V is the unique (strong) solution to the following stochastic differential equation with reflection:

$$\begin{cases} V = \Phi^x(X), \\ X_\bullet = V_0 + \int_0^\bullet (f_\lambda(q_s) - f_\mu(q_s))\,ds + \int_0^\bullet (\lambda'(q_s) - \mu'(q_s))V_s\,ds \\ \qquad + \int_0^\bullet \sqrt{\lambda(q_s) + \mu(q_s)}\,dW_s\,. \end{cases} \tag{4.9}$$

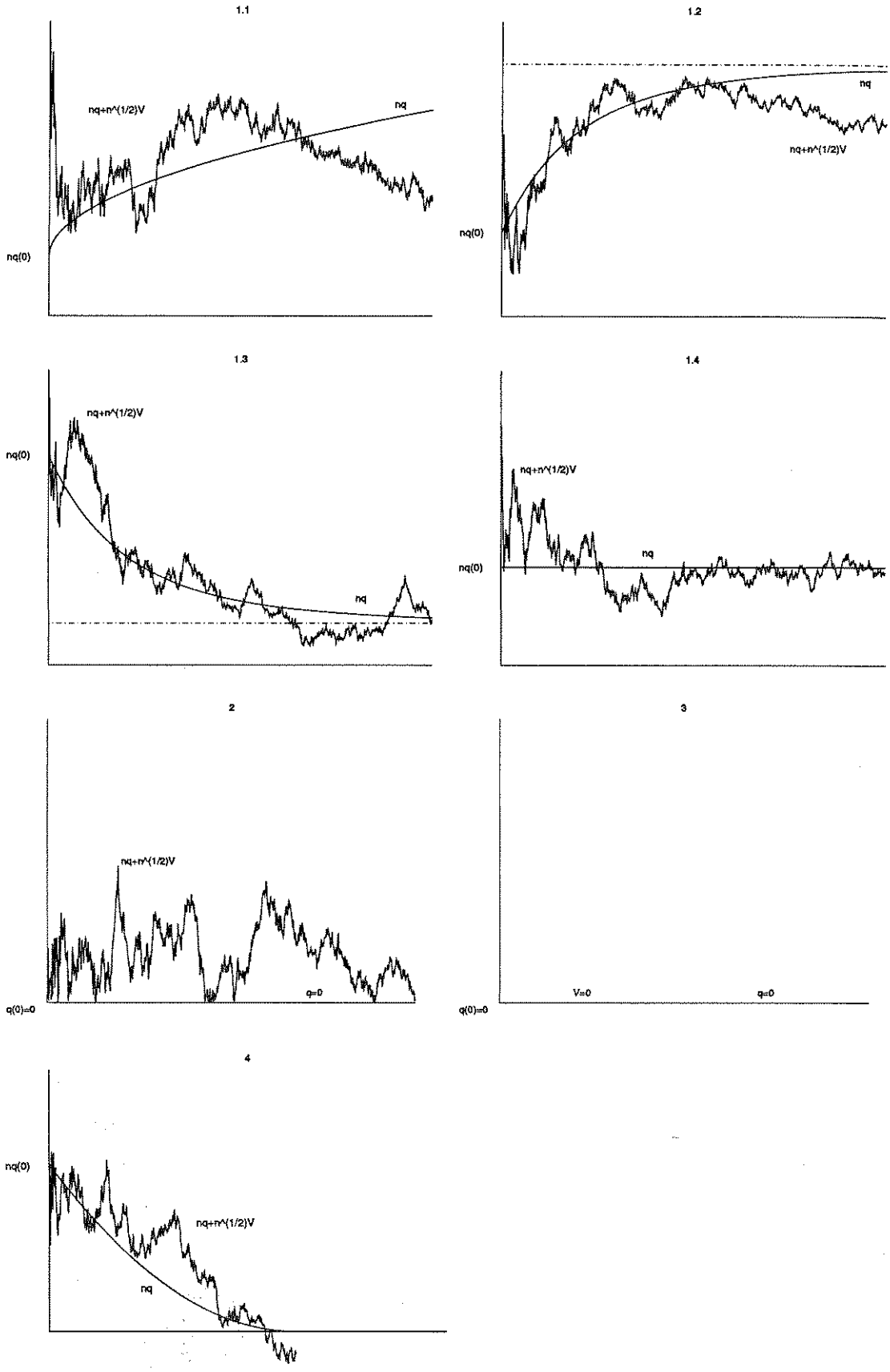

FIG. 1. *Fluid and Diffusion Limits*

Here x and q are given by (4.5), Φ^x is the operator defined by (3.4), and W is a standard Brownian motion.

In what follows, V will be referred to as the *diffusion limit* associated with the queueing sequence under consideration.

The possible forms of x in (4.5) (Cases 1–4 at the end of Subsection 4.1) reveal that x adheres to the conditions imposed in Lemma 3.2, and thus Φ^x is well-defined. In correspondence with the specific forms of Φ^x (Cases 1–4 at the end of Subsection 3.2), the relations (4.9) reduce to the following four Cases (see Figure 1 for suggestive sample paths of V):

1. Permanent large queues

(4.9) is the linear stochastic differential equation:

$$\begin{aligned} dV_t &= [f_\lambda(q_t) - f_\mu(q_t) + (\lambda'(q_t) - \mu'(q_t))\, V_t]\, dt \\ &\quad + \sqrt{\lambda(q_t) + \mu(q_t)}\, dW_t \end{aligned}$$

In particular, if V_0 is a Gaussian variable then V is a Gaussian process. With the notation

$$m_t \triangleq \mathrm{E}V_t, \quad h_t \triangleq \mathrm{E}[V_t - m_t]^2,$$

we have [22]

$$\begin{cases} \dot{m}_t = f_\lambda(q_t) - f_\mu(q_t) + (\lambda'(q_t) - \mu'(q_t))m_t\,, \\ \dot{h}_t = 2(\lambda'(q_t) - \mu'(q_t))h_t + \lambda(q_t) + \mu(q_t), \\ m_0 = \mathrm{E}V_0, \\ h_0 = \mathrm{Var}V_0. \end{cases}$$

2. Critically loaded

(4.9) is a stochastic differential equation with reflection (see Remark 2.2 and Subsection 3.1):

$$V = \Phi(X),$$

where

$$X_\bullet = V_0 + (\lambda'(0) - \mu'(0)) \int_0^\bullet V_s\, ds + \sqrt{2\lambda(0)}\, W_\bullet\,.$$

Equivalently,

$$dV_t = (\lambda'(0) - \mu'(0))V_t\, dt + \sqrt{2\lambda(0)}\, dW_t + dY_t\,, \quad Y = \Psi(X).$$

3. Underloaded

(4.9) degenerates to

$$V_t = 0, \quad t > 0.$$

4. Underloaded with large initial queue

For $t < t_0$, (4.9) coincides with the linear stochastic differential equation of Case 1, and for $t \geq t_0$, it reduces to

$$V_t = \begin{cases} 0 \vee V_{t_0-}, & t = t_0, \\ 0, & t > t_0. \end{cases}$$

Further applications of (4.9) to specific λ, μ, q_0 and V_0, are the subject of Section 5.

4.3. Generalizations. We now present an extension of FCLT, covering λ and μ with piecewise continuous derivatives. This version will be used in Subsection 5.6, in the analysis of finite server queues.

THEOREM 4.3. *Assume that all the conditions of Theorem 4.2 are satisfied, but allow the derivatives λ' and μ' to be piecewise continuous functions with a finite number of discontinuities in each compact subinterval of $(0, \infty)$.*
If, in addition,

$$\text{(4.10)} \qquad \begin{cases} q_t \equiv q_0 > 0, \\ \lambda'(q_0-) \neq \lambda'(q_0+) \ \text{ or } \ \mu'(q_0+) \neq \mu'(q_0+), \end{cases}$$

then the sequence $\{V^n\}$ converges weakly in $(D[0,\infty), J_1)$ to the unique (strong) solution of the following stochastic differential equation (without reflection):

$$dV_t = [f_\lambda(q_0) - f_\mu(q_0) + f(V_t)]\, dt + \sqrt{\lambda(q_0) + \mu(q_0)}\, dW_t,$$

where

$$f(v) = \begin{cases} (\lambda'(q_0+) - \mu'(q_0+))v, & v \geq 0, \\ (\lambda'(q_0-) - \mu'(q_0-))v, & v < 0. \end{cases}$$

If (4.10) does not prevail, then Theorem 4.2 applies without any changes.

Comments on the proof of the theorem will be given in Subsection 7.7. Note that under (4.10), the diffusion limit has continuous sample-paths. Furthermore, (4.10) describes the only case that renders in doubt the existence of the second integral on the right-hand side of (4.9).

4.4. Time acceleration. The FSLLN rescaling (4.2) and (4.4) is equivalent to a procedure of accelerating time and aggregating space units, both by a factor of n. Indeed, consider the simplest, yet illuminating, example of $\{\lambda^n\}$ and $\{\mu^n\}$ that satisfy (4.4):

$$\text{(4.11)} \qquad \lambda^n(\xi) = n\lambda\left(\frac{\xi}{n}\right), \quad \mu^n(\xi) = n\mu\left(\frac{\xi}{n}\right),$$

for some given λ and μ. Equations (4.11) arise naturally for systems with linear, or piecewise–linear, dependence of arrival and service rates on the queue length. (See Subsections 5.4–5.8.) Alternatively to (4.11), consider a sequence $\bar{M}_\xi^n/\bar{M}_\xi^n/1$, $n = 1, 2, \ldots$, of queueing systems, with rates

$$\bar{\lambda}^n(\xi) = \lambda\left(\frac{\xi}{n}\right), \quad \bar{\mu}^n(\xi) = \mu\left(\frac{\xi}{n}\right), \tag{4.12}$$

and queueing processes $\bar{Q}^n$. Introduce the processes $\bar{q}^n = \{\bar{q}_t^n, t \geq 0\}$, given by

$$\bar{q}_t^n = \frac{1}{n}\bar{Q}_{nt}^n, \tag{4.13}$$

Then, rewriting equations (4.1) in terms of $\bar{\lambda}^n$, $\bar{\mu}^n$, and changing variables yields:

$$\bar{q}^n = q^n, \quad n = 1, 2, \ldots. \tag{4.14}$$

4.5. An interpretation of discontinuous diffusion limits. In this subsection we attempt a qualitative explanation of some phenomena that are amplified by our analysis.

As apparent from Subsection 4.2 and Subsection 4.1 (see Case 4), the diffusion limit has a discontinuity in light traffic (underloaded) with large initial queues: $\lambda(\xi) < \mu(\xi)$, for all $\xi \in [0, q_0]$, and

$$\frac{1}{n}Q_0^n \longrightarrow q_0 > 0, \quad \text{a.s.}, \quad n \uparrow \infty.$$

Discontinuity arises at time $t_0 > 0$, given by $t_0 = \inf\{t \geq 0 : q_t = 0\}$.

For simplicity of presentation, let us assume that $Q_0^n = nq_0$, for some $q_0 \in \mathcal{Z}^+$, and the rates in the n-th system are given by (4.11). We suppose also that the V^n converges a.s. to a process V with upper semi-continuous sample paths.

Consider first the case $V_{t_0-} > 0$ and thus $Q_{t_0}^n/\sqrt{n} \longrightarrow V_{t_0}$. To expose the causes of discontinuity, consider Q^n, which is a birth and death process on the integers with the following transitions rates:

$$\begin{cases} q_{k,k+1}^n = n\lambda(k/n), & k = 0, 1, \ldots, \\ q_{k,k-1}^n = n\mu(k/n), & k = 1, 2, \ldots, \\ q_{0,-1}^n = 0; \end{cases} \tag{4.15}$$

One distinguishes three phases in the evolution of Q^n:

1. First relaxation phase of duration t_0: at the beginning of this phase, the queue length is nq_0, reducing to $\sim \sqrt{n}V_{t_0}$ at the end.

2. Second relaxation phase of duration $\sim 1/\sqrt{n}$, starting at t_0. This phase arises from the fact that the queue length at the outset is $\sim \sqrt{n}V_{t_0}$, while the rates in (4.15) are $\sim n\lambda(0)$ and $\sim n\mu(0)$ over this phase ($\lambda(0) < \mu(0)$). This phase shrinks, as $n \uparrow \infty$ ultimately resulting in a discontinuity of V at t_0. At the end of the phase, the queue is $o(\sqrt{n})$.
3. Light traffic phase. Fluid and diffusion limits vanish, as for the underloaded state-independent queues with constant rates $\lambda(0) < \mu(0)$ and with small, $o(\sqrt{n})$, initial queues. (See Subsection 5.1 for the explicit expressions of fluid and diffusion limits in state-independent systems.)

When $V_{t_0} < 0$, similar conclusions apply. The only difference is that the first phase ends $\sim 1/\sqrt{n}$ prior to t_0, the second phase terminates and the third phase starts at time t_0.

When V has a discontinuity at time zero (Case 3, Subsection 4.2 and Subsection 4.1), the first phase is skipped in the evolution of Q^n.

Note that the fluid and diffusion limits both vanish beyond t_0. These are simple examples of state-space collapse, when the limiting process is of a lower dimension than the process it approximates (see Reiman [42], Mandelbaum and Chen [8]). State-space collapse occurs here, when systems operate under light-traffic conditions. To obtain nonzero, more informative limits, one must formulate other limit theorems. Light traffic limit theorems usually involve various cumulative processes such as sums and integrals of the original process (see the survey by Glynn [15]). An alternative limit theorem for the distribution of Q^n in light traffic (in a particular closed network) is presented by Kogan and Liptser [27]. It is possible to obtain a non-degenerate diffusion limit for the second phase via a slower rescaling, namely, considering locally a process $Q^n(t_0 \pm \tau/\sqrt{n})$, for some $\tau > 0$.

Remark. An explanation for a discontinuity of V can be given also in terms of the transient behavior of $\bar{Q}^n$, introduced in the preceding subsection. Again, three successive phases arise in the evolution of $\bar{Q}^n$: relaxation of duration $\sim nt_0$, relaxation of duration $\sim \sqrt{n}$ and, finally, light traffic phase. The rescaling used (acceleration of time by a factor n) shrinks the duration of the second phase, resulting in a discontinuity of V. □

4.6. Alternative rescaling. We now describe rescaling procedures other than (4.4),(4.8), which lead to different approximations of state-dependent queues. Specifically, assume that

$$\frac{1}{n}\lambda^n(n^\alpha\xi) \longrightarrow \lambda(\xi), \quad \sqrt{n}\left[\frac{\lambda^n(n^\alpha\xi)}{n} - \lambda(\xi)\right] \longrightarrow f_\lambda(\xi), \ \ u.o.c., \tag{4.16}$$

and

$$\frac{1}{n}\mu^n(n^\alpha\xi) \longrightarrow \mu(\xi), \quad \sqrt{n}\left[\frac{\mu^n(n^\alpha\xi)}{n} - \mu(\xi)\right] \longrightarrow f_\mu(\xi), \ \ u.o.c., \tag{4.17}$$

as $n \uparrow \infty$, for some specific $\alpha \geq 0$. Evidently, our limits correspond to the case $\alpha = 1$ (see (4.4) and (4.8)). Alternative rescaling procedures were considered by Yamada in [49] and [50]: the case $\alpha = 0$ is treated in [49], where the diffusion limit is of a Bessel type with a negative drift; the case $\alpha = 1/2$ is considered in [50], where the diffusion limit is a solution to a stochastic differential equations with *state*-dependent coefficients (while in our case the coefficients are *time*-dependent). The fluid limits vanish in both [49] and [50].

A comparison of the different approaches is summarized in Table 1 (with preference to clarity of presentation over precision). The expressions for the rates on the first upper part of Table 1 are based on (4.16). Combining these expressions with the fluid and diffusions limits from the third and forth parts of the table yields the last part.

Remark. The reflection $\widetilde{Y}$ in the expression for V, when $\alpha = 0$, is characterized by the condition:

$$\int_0^t V_s \, d\widetilde{Y}_s = \gamma \cdot t, \quad t \geq 0,$$

for some $\gamma > 0$. Such reflection gives rise to a Bessel–type distribution for V. □

To recapitulate, our approach leads to a second order approximations for queueing processes: fluid limits provide approximations for actual values of queues, while diffusions limits—for their fluctuations. When fluid limits vanish, the three approaches provide approximations for systems in which arrival and service rates are sensitive to small ($\alpha = 1$, $\mathcal{O}\left(n^{-1/2}V\right)$), medium ($\alpha = 1/2$, $\mathcal{O}\left(V\right)$) and large ($\alpha = 0$, $\mathcal{O}\left(\sqrt{n}V\right)$) fluctuations of queues.

In Subsection 5.9, we compare the three types of rescaling, $\alpha = 0, 1/2, 1$, by applying them to a single queueing system.

5. Examples and applications. This section is devoted to some applications of the limit theorems presented in Section 4. In Subsections 5.1 and 5.3 we characterize the conditions under which diffusion limits are Brownian or Ornstein-Uhlenbeck processes. In Subsection 5.2 we consider asymptotically small initial queues.

A unified approach is offered to obtain fluid and diffusion limits for state dependent queueing systems. We demonstrate this through application and simplification of some completely or partially known results (see Subsections 5.4–5.8). Different types of rescaling are applied in Subsection 5.9 to a single model, thus highlighting their differences. In Subsection 5.10 we outline guidelines for implementing some of our approximations.

TABLE 1. *Rescaling State-Dependent Queues*

$\lambda^n(Q^n) \approx n\,\lambda\left(\frac{1}{n^\alpha}Q^n\right)+\sqrt{n}\,f_\lambda\left(\frac{1}{n^\alpha}Q^n\right)$ $\mu^n(Q^n) \approx n\,\mu\left(\frac{1}{n^\alpha}Q^n\right)+\sqrt{n}\,f_\mu\left(\frac{1}{n^\alpha}Q^n\right)$			
Approaches	$\alpha = 1$	$\alpha = 1/2$ Yamada 93	$\alpha = 0$ Yamada 86
Fluid FSLLN	$\frac{Q^n}{n} \xrightarrow{a.s.} q$ $\dot{q}=\lambda(q)-\mu(q)+dy,\ q \perp dy$	$\frac{Q^n}{n} \xrightarrow{P} 0$ Critical Loading	
Diffusion	$\sqrt{n}\left(\frac{Q^n}{n}-q\right) \xrightarrow{d} V,\ M_1$	$\frac{Q^n}{\sqrt{n}} \xrightarrow{d} V,\ J_1$	
FCLT	$dV_t=b(q_t)V_t\,dt+c(q_t)\,dW_t+dY_t$	$dV_t=\beta(V_t)\,dt+\gamma(V_t)\,dW_t+dY_t$	$V_t=-bt+cW_t+\widetilde{Y}_t$
Effective Rates	$\lambda^n(Q^n) \approx n\,\lambda(q)+\sqrt{n}\left[\lambda'(q)V+f_\lambda(q)\right]$ $\mu^n(Q^n) \approx -''-$	$\lambda^n(Q^n) \approx n\,\lambda(V)+\sqrt{n}\,f_\lambda(V)$ $\mu^n(Q^n) \approx -''-$	$\lambda^n(Q^n) \approx n\,\lambda\left(\sqrt{n}\,V\right)+\sqrt{n}\,f_\lambda\left(\sqrt{n}\,V\right)$ $\mu^n(Q^n) \approx -''-$

5.1. State–independent models: linear fluid and Brownian diffusion limits. Consider a sequence of state-independent queues $M^n/M^n/1$ with constant rates $\lambda^n = n\lambda$ and $\mu^n = n\mu$ ($f_\lambda = f_\mu \equiv 0$ has been chosen in (4.8) for simplicity).

Theorems 4.1 and 4.2 yield the following expressions for the fluid and diffusion limits:

$$x_t = q_0 + (\lambda - \mu)t, \quad q = x + y; \quad V = \Phi^x\left(V_0 + \sqrt{\lambda + \mu}\, W_\bullet\right).$$

Three modes of evolution arise

1. *Overloaded:* $\lambda > \mu$. Here $q_t = q_0 + (\lambda - \mu)t$ and

$$V_\bullet = V_0 + W((\lambda + \mu)\bullet) \overset{d}{\sim} BM_{V_0}(0, \lambda + \mu).$$

2. *Critically loaded:* $\lambda = \mu$. Here $q \equiv q_0$;
   If $q_0 > 0$, then $V_\bullet = V_0 + W((\lambda + \mu)\bullet) \overset{d}{\sim} BM_{V_0}(0, \lambda + \mu)$;
   If $q_0 = 0$, then $V = \Phi[V_0 + W((\lambda + \mu)\bullet)] \overset{d}{\sim} RBM_{V_0}(0, \lambda + \mu)$.

3. *Underloaded:* $\lambda < \mu$.
   If $q_0 > 0$ (large initial queues):

$$q_t = \begin{cases} q_0 + (\lambda - \mu)t, & t \le t_0 = \dfrac{q_0}{\mu - \lambda}, \\ 0, & t > t_0 \end{cases};$$

$$V_t = \begin{cases} V_0 + W((\lambda + \mu)t), & t < t_0, \\ 0, & t > t_0, \\ \max[V_0 + W((\lambda + \mu)t), 0], & t = t_0. \end{cases}$$

   If $q_0 = 0, V_0 \neq 0$ (moderate initial queues): $q \equiv 0$ and $V_t = 0$, $t > 0$.
   If $q_0 = V_0 = 0$ (small initial queues): $q \equiv 0$ and $V \equiv 0$.

5.2. Small initial queues. In this subsection, our limit theorems are applied with asymptotically small initial queues, that is $q_0 = V_0 = 0$. This case is highlighted for its simplicity: diffusion limits are continuous on $[0, \infty)$ and the behavior of the fluid (positive, vanishing) and diffusion limits (with/without reflection, vanishing) depends solely on $\lambda(0)$, $\mu(0)$. Specifically:

1. *Overloaded:* $\lambda(0) > \mu(0)$. Here q is strictly positive over $(0, \infty)$ (Cases 1.1 or 1.2 of Subsection 4.1) and V is a diffusion as in Case 1 at the end of Subsection 4.2.
2. *Critically loaded:* $\lambda(0) = \mu(0)$. Here $q \equiv 0$, and V is a reflected diffusion (Cases 2 in Subsections 4.1 and 4.2).
3. *Underloaded:* $\lambda(0) < \mu(0)$. Here $q \equiv 0$ and $V \equiv 0$ (Cases 3 in Subsections 4.1 and 4.2).

5.3. Constant fluid and Ornstein-Uhlenbeck diffusion limits. We continue with queues whose diffusion limits are Ornstein-Uhlenbeck or reflected Ornstein-Uhlenbeck processes. As previously, $f_\lambda = f_\mu \equiv 0$ in (4.8) is chosen for simplicity.

Let λ, μ and $\xi_1 \geq 0$ satisfy:

$$\lambda(\xi_1) = \mu(\xi_1) \quad \text{and} \quad \lambda'(\xi_1) < \mu'(\xi_1). \tag{5.1}$$

Ornstein-Uhlenbeck diffusion limit. Add to (5.1) the assumption $\xi_1 > 0$. Two cases arise:

1. $q_0 = \xi_1$
   Theorems 4.1 and 4.2 yield $q \equiv \xi_1$ and
   $$dV_t = (\lambda'(\xi_1) - \mu'(\xi_1))V_t\,dt + \sqrt{2\lambda(\xi_1)}\,dW_t\,.$$
   Thus V is an Ornstein-Uhlenbeck process with
   $$m_t \stackrel{\Delta}{=} \mathrm{E}V_t = m_0 e^{-2\alpha t}\,,$$
   $$h_t \stackrel{\Delta}{=} \mathrm{Var}V_t = \frac{\sigma^2}{2\alpha} + \left(h_0 - \frac{\sigma^2}{2\alpha}\right) e^{-2\alpha t}\,,$$
   $$h_{s,t} \stackrel{\Delta}{=} \mathrm{Cov}(V_s\,,V_t) = \left[h_0 + \frac{\sigma^2}{2\alpha}(e^{2\alpha(t\wedge s)} - 1)\right]e^{-\alpha(t+s)}\,,$$
   where $\sigma = \sqrt{2\lambda(\xi_1)}$, and $\alpha = \mu'(\xi_1) - \lambda'(\xi_1)$. Taking $V_0 \stackrel{d}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{2\alpha}\right)$, V is the stationary Ornstein-Uhlenbeck process with
   $$h_{s,t} = \frac{\sigma^2}{2\alpha} e^{-\alpha|t-s|}\,.$$
2. $q_0 \neq \xi_1$
   Assume, in addition, that
   $$\lambda(\xi) > \mu(\xi), \quad \xi < \xi_1\,,$$
   $$\lambda(\xi) < \mu(\xi), \quad \xi > \xi_1\,.$$
   Then, it follows from Subsection 4.1 (see Cases 1.2, 1.3) that, $q_t \downarrow\downarrow \xi_1$ or $q_t \uparrow\uparrow \xi_1$ as $t \uparrow \infty$. For V we have
   $$dV_t = (\lambda'(q_t) - \mu'(q_t))V_t\,dt + \sqrt{\lambda(q_t) + \mu(q_t)}\,dW_t\,.$$
   The random variable V_t converges weakly, as $t \uparrow \infty$, to $V(\infty) \stackrel{d}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{2\alpha}\right)$, where σ and α are as defined above.

Reflected Ornstein-Uhlenbeck diffusion limit. Assume $\xi_1 = 0$ in (5.1). The diffusion limit V is then

$$dV_t = (\lambda'(0) - \mu'(0))V_t\,dt + \sqrt{2\lambda(0)}\,dW_t + dY_t\,,$$

which is a reflected Ornstein-Uhlenbeck process. This example was described by Liptser and Shiryayev [36, pages 753,754].

5.4. Finite population and general service (Liptser, Shiryayev [36]). Consider a sequence of $M/M^n_\xi/1/\infty/n$ systems, as in [36, pages 638–636]. The parameters of the n-th system are given by

$$\lambda^n(Q^n) = \lambda \cdot (n - Q^n), \quad \mu^n(Q^n) = n\mu\left(\frac{Q^n}{n}\right),$$

for some $\lambda \geq 0$ and a function μ. We identify the parameters of the fluid and diffusion limits via (4.4),(4.8): $\lambda(\xi) = \lambda \cdot (1-\xi)$, $\mu(\xi)$, and $f_\lambda = f_\mu \equiv 0$. Let Q^n_0 and μ satisfy the conditions of the FSLLN and FCLT (Theorems 4.1 and 4.2). Then the fluid limit is given by

$$\dot{q}_t = \lambda(1 - q_t) - \mu(q_t),$$

and if, for example, $q_t > 0$ for all $t > 0$, the diffusion is

$$V_\bullet = V_0 - \int_0^\bullet (\lambda + \mu'(q_s))\, V_s\, ds + \int_0^\bullet \sqrt{\lambda(1-q_s) + \mu(q_s)}\, dW_s\,.$$

Assume, in addition, that $\lambda(1-\xi_1) = \mu(\xi_1)$ and $\mu'(\xi_1) > -\lambda$, for some $\xi_1 \in (0,1]$, and consider separately two cases:

1. $q_0 = \xi_1$. The expressions above for q and V yield $q_t \equiv \xi_1$, and

$$dV_t = -(\lambda + \mu'(\xi_1))V_t\, dt + \sqrt{2\lambda(1-\xi_1)}\, dW_t\,.$$

With $V_0 \overset{d}{\sim} \mathcal{N}(0, \sigma^2)$, $\sigma^2 = [\lambda(1-\xi_1)]/[\lambda + \mu'(\xi_1)]$, V becomes a stationary Ornstein-Uhlenbeck process.

2. $q_0 \neq \xi_1$. Stipulating,

$$\lambda(\xi) > \mu(\xi),\ \ \xi < \xi_1\,; \quad \lambda(\xi) < \mu(\xi),\ \ \xi > \xi_1\,,$$

we obtain that $V_t \xrightarrow{\mathrm{d}} V(\infty)$, where $V(\infty) \overset{d}{\sim} \mathcal{N}(0, \sigma^2)$, with σ^2 as in case 1. The random variable $V(\infty)$ can be used to approximate the long-run behavior of Q^n, for sufficiently large n (see [36, pages 653–656] and Remark 5.1).

5.5. Infinite number of servers (Whitt [48]). Consider a sequence of $M^n/M/\infty$ systems, namely

$$\lambda^n(Q^n) \equiv n\lambda, \quad \mu^n(Q^n) = \mu Q^n,$$

for some $\lambda, \mu > 0$. This corresponds to an infinite–server queue under heavy traffic. By (4.4), $\lambda(\xi) \equiv \lambda$, $\mu(\xi) = \mu\xi$.

Assume that $q_0 = \rho \overset{\Delta}{=} \lambda/\mu$. Since $\mu(q_0) = \lambda$, we have $q_t \equiv \rho$, and

$$dV_t = -\mu V_t\, dt + \sqrt{2\lambda}\, dW_t\,,$$

(see Case 1.4 in Subsection 4.1 and Case 1 in Subsection 4.2). That is, if $V_0 \overset{d}{\sim} \mathcal{N}(0,\rho)$, then V is a stationary Ornstein-Uhlenbeck process.

For the general case $q_0 \neq \rho$, we obtain

$$q_t = \rho + (q_0 - \rho)\, e^{-\mu t}, \tag{5.2}$$

and

$$dV_t = -\mu V_t\, dt + \sqrt{2\lambda + (\mu q_0 - \lambda)\, e^{-\mu t}}\, dW_t \tag{5.3}$$

(Cases 1.2,1.3 in Subsection 4.1 and Case 1 in Subsection 4.2). The process V has a steady-state distribution $\mathcal{N}(0,\rho)$, which can be used to approximate the distribution of $Q^n(\infty)$. (See [15], [23] and [36, pages 653–656].)

5.6. Finite number of servers (Iglehart [18], Halfin and Whitt [17]).

The limit procedure of Borovkov [6] and Iglehart [18]. Consider a sequence of $M^n/M/n$ systems such that

$$\lambda^n(Q^n) \equiv n\lambda, \quad \mu^n(Q^n) = \mu \cdot (Q^n \wedge n),$$

for some $\lambda, \mu > 0$. By (4.4), $\lambda(\xi) \equiv \lambda$, $\mu(\xi) = \mu \cdot (\xi \wedge 1)$. The traffic intensity for the n-th system is given by

$$\rho^n = \frac{n\lambda}{n\mu} = \frac{\lambda}{\mu} \overset{\Delta}{=} \rho.$$

Three cases arise: $\rho < 1$, $\rho > 1$, $\rho = 1$.

1. $\rho < 1$

   Assume first that $q_0 \in [0,1]$. Then q and V are the same as in the case of an infinite number of servers (see (5.2) and (5.3)). In this sense, the sequence of $M^n/M/n$ systems with $\rho < 1$ is asymptotically $(n \uparrow \infty)$ indistinguishable from the sequence $M^n/M/\infty$, $n = 1, 2, \ldots$. In other words, due to (5.2),

   $$\frac{1}{n} Q^n \longrightarrow q, \text{ u.o.c, a.s., and } q_t < 1, \text{ for all } t > 0,$$

   and thus the probability $\mathrm{P}[Q_t^n > n]$, that all servers are busy, converges to zero, as $n \uparrow \infty$.

   If $q_0 \in (1,\infty)$, then for $t \leq (q_0 - 1)/(\mu - \lambda) \overset{\Delta}{=} t_1$ the limits coincide with those of the state-independent system (Subsection 5.1, underloaded regime with large initial queues). For $t \geq t_1$ the limits are the same as for an infinite number of servers, given $q_0 = 1$ (see (5.2) and (5.3)).

2. $\rho > 1$
   For the fluid limit,

$$q_t = \begin{cases} \rho + (q_0 - \rho)e^{-\mu t}, & t < t_2\,, \\ (\lambda - \mu)t, \quad t \geq t_2\,; & t_2 = \inf\{t : q_t \geq 1\}. \end{cases}$$

   If $q_0 \geq 1$, then $t_2 = 0$ and only the second equation in the above expression is relevant. Therefore q is a combination of two limits: a model with an infinite number of servers, $t < t_2$ (Subsection 5.5) and a state-independent system, $t \geq t_2$ (Subsection 5.1, overloaded regime). The diffusion limit enjoys a similar structure.
3. $\rho = 1$
   For $q_0 \in [0, 1)$, the limits are analogous to those of case $\rho < 1$. For $q_0 \in (1, \infty)$, the limits are the same as for the state-independent critically loaded case (Subsection 5.1).
   For $q_0 = 1$, the fluid limit is trivial: $q_t \equiv q_0$. While μ is non-differentiable at $q_0 = 1$, the generalized FCLT (Theorem 4.3) is applicable:

$$dV_t = f\,(V_t)\,dt + \sqrt{2\mu}\,dW_t\,,$$

   where

$$f(v) = \begin{cases} 0, & v \geq 0, \\ -\mu v, & v < 0. \end{cases}$$

   Thus, V is a combination of a Brownian motion and an Ornstein-Uhlenbeck process. Such limits were proposed by Halfin and Whitt [17] and are described in what now follows.

The limit procedure of Halfin and Whitt [17]. Reconsider the limit procedure of Borovkov and Iglehart described above. For that case, if $\rho < 1$, the probability that all servers are busy converges to zero, as $n \uparrow \infty$. Halfin and Whitt [17] proposed another limit procedure for a sequence $M^n/M/\infty$ (with traffic intensity in the n-th system $\rho^n < 1$), such that the probability of delay converges to a non–degenerate limit. It was shown in [17] that

$$\lim_{n\uparrow\infty} \mathrm{P}\,\{Q^n(\infty) \geq n\} = \alpha, \quad 0 < \alpha < 1,$$

if and only if

$$\lim_{n\uparrow\infty} (1 - \rho^n)\sqrt{n} = \beta, \quad 0 < \beta < \infty,$$

in which case $\alpha = \left[1 + \sqrt{2\pi}\beta\Phi(\beta)\exp(\beta^2/2)\right]^{-1}$. Here Φ is the standard normal distribution function. Note that $Q^n(\infty)$ exists since $\rho^n < 1$.

The limits described in [17] can also be deduced from our theorems. Indeed, let

$$\lambda^n(Q^n) \equiv \mu \cdot (n - \beta\sqrt{n}), \quad \mu^n(Q^n) = \mu \cdot (Q^n \wedge n), \ \ \mu > 0.$$

The parameters of the fluid and diffusion limits are identified via (4.4) and (4.8): $\lambda(\xi) \equiv \mu$, $\mu(\xi) = \mu \cdot (\xi \wedge 1)$, and $f_\lambda \equiv 0$, $f_\mu \equiv -\beta\mu$. Moreover,

$$\rho^n = \frac{\mu \cdot (n - \beta\sqrt{n})}{\mu n} = 1 - \frac{\beta}{\sqrt{n}}, \tag{5.4}$$

that is $\rho^n \uparrow\uparrow 1$, and $(1 - \rho^n)\sqrt{n} \longrightarrow \beta$, as $n \uparrow \infty$.

Assuming $q_0 = 1$ and noting that μ is non-differentiable at this point, we obtain that $q \equiv 1$, and by the generalized FCLT (Theorem 4.3),

$$dV_t = f(V_t)\, dt + \sqrt{2\mu}\, dW_t,$$

where

$$f(v) = \begin{cases} -\beta\mu, & v \geq 0, \\ -\mu(\beta + v), & v \leq 0. \end{cases}$$

Here the diffusion limit is a combination of an Ornstein-Uhlenbeck process and a Brownian Motion, both with negative drifts.

Note that the drift $(-\beta\mu)$ appears in V due to the specific rates of convergence in (5.4), while the special choice of $q_0 = 1$ (at this point μ is non-differentiable) gives rise to the compound structure of the diffusion limit. The stationary distribution of V can be used in approximating the distribution of $Q^n(\infty)$, as shown in [17] (see also Remark 5.1).

5.7. The repairman problems (Iglehart and Lemoine [19], Iglehart [18]). Consider a sequence of $M/M/k^n/\infty/n$ systems. The n-th system is interpreted as follows [19]. There are n operating units subject to breakdowns and k^n repair facilities. The rates of breakdown and repair are λ and μ respectively. Thus Q_t^n here is the number of operating units which are being repaired or are awaiting repair at time t. Introduce a process $Y^n = n - Q^n$, which describes the number of units operating at time t. The fluid and diffusion limits for $\{Y^n\}$ can be obtained immediately from those for $\{Q^n\}$, therefore we focus on the latter only.

The arrival and service rates in the n-th system are given by

$$\lambda^n(Q^n) = \lambda \cdot (n - Q^n), \quad \mu^n(Q^n) = \mu \cdot (Q^n \wedge k^n).$$

Assume that

$$\frac{k^n}{n} \longrightarrow k, \quad \sqrt{n}\left(\frac{k^n}{n} - k\right) \longrightarrow 0; \quad 0 < k < 1,$$

as $n \uparrow \infty$. Then, by (4.4) and (4.8), $\lambda(\xi) = \lambda \cdot (1 - \xi)$, $\mu(\xi) = \mu \cdot (\xi \wedge k)$, and $f_\lambda = f_\mu \equiv 0$. There are three combinations of the parameters λ, μ and

k, each corresponding to an essentially different fluid and diffusion limits. For simplicity, we pursue the cases where the fluid limit $q \equiv q_0$, which are sufficient to demonstrate the main modes of behavior.

1. $\dfrac{\lambda}{\lambda+\mu} < k$, $q_0 = \dfrac{\lambda}{\lambda+\mu}$. Here $dV_t = -(\lambda+\mu)V_t\,dt + \sqrt{2\frac{\lambda\mu}{\lambda+\mu}}\,dW_t$.
2. $\dfrac{\lambda}{\lambda+\mu} > k$, $q_0 = 1 - \dfrac{\mu k}{\lambda}$. Here $dV_t = -\lambda V_t\,dt + \sqrt{2\mu k}\,dW_t$.
3. $\dfrac{\lambda}{\lambda+\mu} = k$, $q_0 = \dfrac{\lambda}{\lambda+\mu}$.
   Note that μ is not differentiable at q_0. Hence we have (by Theorem 4.3)

$$dV_t = f(V_t)\,dt + \sqrt{2\frac{\lambda\mu}{\lambda+\mu}}\,dW_t,$$

where

$$f(v) = \begin{cases} -\lambda v, & v \geq 0, \\ -(\mu+\lambda)v, & v < 0. \end{cases}$$

Remark. The FSLLN reveals that, for all t, the probability $\mathrm{P}[Q_t^n > k^n]$ of all repair facilities being busy, converges to zero in Case 1 and to unity in Case 2. Case 1 is therefore preferable over Case 2 for real repair systems. □

Another repairman model is proposed in [18], which generalizes the one described above (see, also, Kurtz [33]). In this model, the n-th system has m^n spares, in addition to the elements described previously. These spares can immediately replace those operating units that have failed. In our terms, one can write

$$\lambda^n(Q^n) = \lambda n - \lambda \cdot (Q^n - m^n)^+, \quad \mu^n(Q^n) = \mu \cdot (Q^n \wedge k^n).$$

Following [18], assume that $m^n = nm$ and $k^n = nk$. In this case,

$$\lambda(\xi) = \lambda \cdot \left(1 - (\xi - m)^+\right), \quad \mu(\xi) = \mu \cdot (\xi \wedge k)$$

and $f_\lambda = f_\mu \equiv 0$. If, for example,

$$\mu k < \lambda, \quad k < m, \quad \text{and} \quad q_0 = 1 + m - \frac{\mu k}{\lambda},$$

then $q \equiv q_0$, and $dV_t = -\lambda V_t\,dt + \sqrt{2\mu k}\,dW_t$.

Note that our theorems apply to the more general case

$$\sqrt{n}\left(\frac{k^n}{n} - k\right) \longrightarrow \widetilde{k}, \quad \sqrt{n}\left(\frac{m^n}{n} - m\right) \longrightarrow \widetilde{m}, \quad \text{as } n \uparrow \infty; \quad q_0 \geq 0.$$

in which f_λ and f_μ need not vanish.

5.8. Queues with reneging (Coffman *et al.* [11]). Following [11], consider a sequence of queues with processor-shared service and reneging. Reneging means that a customer is lost when its sojourn time reaches an individual random deadline. Namely, we assume that, in the n-th system, the arrival and service rates are given by:

$$\lambda^n(Q^n) = n\lambda, \quad \mu^n(Q^n) = \frac{1}{Q^n} \cdot (n\lambda + \alpha\sqrt{n\lambda}) \cdot Q^n + \nu Q^n, \tag{5.5}$$

for arbitrary positive λ, α and ν. The quantity $n\lambda + \alpha\sqrt{\lambda n}$ is the service rate, shared among all customers in the system. The parameter ν is interpreted as the reneging rate. Assume that $q_0 = 0$. Then the fluid limit vanishes and the diffusion limit is the reflected Ornstein-Uhlenbeck process

$$dV_t = -\alpha\sqrt{\lambda}\, dt - \nu V_t\, dt + \sqrt{2\lambda}\, dW_t + dY_t,$$

which are limits for the critically loaded mode (Cases 2 in Subsections 4.1 and 4.2). With appropriate parameters in (5.5), one obtains limit theorems for other regimes, beyond [11].

5.9. A comparison of different rescaling procedures. The three types of rescaling, described in Section 4.6, are now applied to a single queueing system, operating in different modes.

Consider a sequence $M^n_\xi/M^n_\xi/1$, $n = 1, 2, \ldots$, with arrival and service rates given by

$$\lambda^n(Q^n) = b^n + c^n \cdot (Q^n \wedge \delta^n), \quad \mu^n(Q^n) = \beta^n + \gamma^n \cdot (Q^n \wedge \delta^n), \tag{5.6}$$

where b^n, c^n, δ^n, β^n, γ^n are positive constants, $c^n \le \gamma^n$. Reviewing the examples from the previous subsections, one can offer various interpretations for the n-th system:

1. Service is provided simultaneously by δ^n servers (each at a rate γ^n) and by a processor-shared server (at a rate β^n). The arrival process consists of exogenous arrivals (rate b^n) and served customers that leave for a while, then return for rework with probability $c^n/\gamma^n \le 1$. (The time till their return is assumed short enough that the queue does not change much, and long enough that they are independent of exogenous arrivals.) This is a possible model to some human-service systems.
2. Service is provided by a single server, at a rate that increases with queue length, but only up to an exhaustion level $\beta^n + \gamma^n \cdot \delta^n$. Arrival rates, which increase with queue length, describe a possible scenario where a long queue attracts customers being a source of information on service value.

Assume that $Q^n_0 = 0$. The three examples, presented below, exhibit different diffusion limits V, according the choice of parameters in (5.6).

These diffusion limits are obtained through the three types of rescaling discussed in Section 4.6.

1. $\alpha = 1$. Let $b^n = \beta^n = nb$, $c^n = \gamma^n = c$ and $\delta^n = n\delta$; then

$$V = \sqrt{2b}W + Y;$$

2. $\alpha = 1/2$. Let $b^n = nb$, $\beta^n = nb + \sqrt{n}$, $c^n = \sqrt{n}\,c$, $\gamma^n = \sqrt{n}\,c + 1$ and $\delta^n = \sqrt{n}\,\delta$; then

$$\begin{aligned} dV_t &= -[1 + (V_t \wedge \delta)]\,dt + \sqrt{b + c(V_t \wedge \delta)}\,dW_t^1 \\ &\quad + \sqrt{b + c(V_t \wedge \delta)}\,dW_t^2 + dY_t; \end{aligned}$$

3. $\alpha = 0$. Let $b^n = \beta^n = nb$, $c^n = nc$, $\gamma^n = nc + \sqrt{n}$ and $\delta^n = \delta$; then

$$V_t = -\delta \cdot t + \sqrt{2c\delta + 2b}\,W_t + Y_t.$$

Here $b, c, \delta > 0$, W, W^1, W^2 are standard Brownian Motions (W^1 and W^2 are independent) and Y is a normal reflection term. For all three examples, the fluid limit $q \equiv 0$ and $Q^n/\sqrt{n} \xrightarrow{\mathrm{d}} V$. The examples mainly differ by the number of servers relative to the queue size, which are $n : \sqrt{n}$, $\sqrt{n} : \sqrt{n}$, $1 : \sqrt{n}$ in examples 1,2,3 respectively.

5.10. Approximating queueing systems. Our approximations typically apply when some natural parameters of the systems are taken to an extreme. For example, large number of servers, population size, initial queue or traffic intensity. The sequence $M_\xi^n/M_\xi^n/1$, $n = 1, 2, \ldots$ is used to formalize the approximation, which always takes the form

$$Q^n \overset{\mathrm{d}}{\approx} nq + \sqrt{n}\,V, \tag{5.7}$$

where q and V are the fluid and diffusion limits respectively.

Remark 5.1. A relation analogous to (5.7) can be written, at least formally, for the stationary distributions $Q(\infty)$ and $V(\infty)$, when they exist. Examples of theorems that support such approximations are Halfin and Whitt [17], Kaspi and Mandelbaum [23], Ethier and Kurtz [14, Chapter 4,§9], Liptser and Shiryayev [36]. □

6. Proof of FSLLN. This section is devoted to the proof of Theorem 4.1. To simplify the presentation, we consider (4.11) only. The general case requires minor notational changes.

The linear growth constraint on the function λ implies that both q^n, $n = 1, 2, \ldots$, and q are non-explosive (see Remarks 2.1 and 3.2).

Let T be an arbitrary positive constant. Subtracting the equation for q in (4.5) from the equation for q^n in (4.3) and using the Lipschitz property

of Φ and Ψ (see Appendix A), we obtain

$$
\begin{aligned}
\| q^n - q \|_t \;\leq\; & C \mid q_0^n - q_0 \mid \\
& + C \left\| \frac{1}{n} N_+ \left(n \int_0^\bullet \lambda(q_s^n)\, ds \right) - \int_0^\bullet \lambda(q_s^n)\, ds \right\|_t \\
& + C \left\| \frac{1}{n} N_- \left(n \int_0^\bullet \mu(q_s^n)\, ds \right) - \int_0^\bullet \mu(q_s^n)\, ds \right\|_t \\
& + C \left\| \int_0^\bullet (\lambda(q_s^n) - \lambda(q_s))\, ds \right\|_t \\
& + C \left\| \int_0^\bullet (\mu(q_s^n) - \mu(q_s))\, ds \right\|_t ,
\end{aligned}
\tag{6.1}
$$

for all $t \leq T$, where C is the Lipschitz constant for Φ and Ψ. Note that the first term on the right-hand side of (6.1) converges to zero, by the conditions of the theorem.

It will be proved in Lemma 6.1 below that

$$
\forall T > 0 \; \exists A_T < \infty \;\ni\; \overline{\lim_n} \| q^n \|_T \leq A_T \;\text{ a.s.} \tag{6.2}
$$

(A_T is a non-random scalar.) Consequently, the second and the third terms on the right-hand side of (6.1) converge to zero, by the continuity of λ and μ, combined with the FSLLN for any Poisson process N:

$$
\lim_{n \uparrow \infty} \left\| \frac{1}{n} N(nt) - t \right\|_T = 0, \quad \forall T \geq 0, \;\text{ a.s.}
$$

From (6.2) and the Lipschitz property of λ and μ, the last two terms in (6.1) satisfy a.s. the following inequality:

$$
\left\| \int_0^\bullet (\lambda(q_s^n) - \lambda(q_s))\, ds \right\|_t + \left\| \int_0^\bullet (\mu(q_s^n) - \mu(q_s))\, ds \right\|_t \leq C_T \int_0^t \| q^n - q \|_s \, ds,
$$

for all but finitely many values (in general, random) of n. Here

$$
C_T = C_T^\lambda + C_T^\mu ,
$$

where C_T^λ and C_T^μ denote the Lipschitz constants for λ and μ respectively in $[0, (A_T \vee \|q\|_T) + 1]$. Now combining all of the above, we obtain

$$
\| q^n - q \|_t \leq \epsilon^n(T) + B \int_0^t \| q^n - q \|_s \, ds, \quad 0 \leq t \leq T, \tag{6.3}
$$

where $\epsilon^n(T)$, which is the sum of the first three terms on the right-hand side of (6.1) (with $t = T$), converges to zero; and $B = C \cdot C_T$. Finally, applying Gronwall's inequality (e.g. [14, page 428]) to (6.3) completes the proof of the theorem. □

It remains to show that

LEMMA 6.1. *Assertion (6.2) holds under the conditions of Theorem 4.1.*

Proof. We prove the lemma by bounding $\{q^n\}$ from above, with a sequence of processes $\{g^n\}$, for which the assertion holds. To this end, consider the sequence of processes $G^n = \{G^n_t, t \geq 0\}$, $n = 1, 2, \ldots$, which are solutions to:

$$G^n_t = Q^n_0 + N_+ \left(nK \int_0^t \left(1 + \frac{G^n_s}{n} \right) ds \right),$$

where K is the constant from the linear growth condition of the theorem. The process G^n is pure birth with parameters

$$G^n_0 = Q^n_0, \quad q^n_{k,k+1} = nK \left(1 + \frac{k}{n} \right), \quad k \in \mathcal{Z}^+,$$

as apparent from the interpretation discussed in Subsection 2.2. A pathwise analysis can now be used to deduce that

$$Q^n_t \leq G^n_t, \quad t \geq 0, \quad n = 1, 2, \ldots, \quad \text{a.s.} \tag{6.4}$$

In order to prove the lemma, it is sufficient to show that (6.2) holds with q^n_t replaced by

$$g^n_t = \frac{G^n_t}{n} = q^n_0 + \frac{1}{n} N_+ \left(n \int_0^t (1 + g^n_s)\, ds \right).$$

However, by Theorem 2.2 of Kurtz [31], $g^n \longrightarrow g$, u.o.c., a.s., as $n \uparrow \infty$, where g is the unique solution to

$$g_\bullet = q_0 + K \int_0^\bullet (1 + g_s)\, ds, \quad t \geq 0.$$

Hence, assertion (6.2) for g^n is established. The proof is now complete. □

Remark. The domination argument was required, since the general theorems of [31] do not treat the reflection phenomenon. Note that reflection does not arise for g^n. □

7. Proof of FCLT. This section is devoted to the proof of Theorem 4.2. As previously, we restrict the proof to the case (4.11) only (consequently, $f_\lambda = f_\mu \equiv 0$). Commentary on the general case is provided in Remark 7.1 at the end of Subsection 7.3.

7.1. Existence and uniqueness. First we confirm that (4.9) is well-defined and enjoys a unique strong solution.

The right-hand side of (4.9) is well-defined. To see that, according to the definition of Φ^x in (3.4) we must show, *first*, that x given by (4.5) satisfies the conditions of Lemma 3.2 and *second*, that the argument of Φ^x in (4.9) is always in $C^x[0, \infty)$ (see (3.3)). First, as explained

in Subsection 3.1, x given by (4.5) is continuous with $x_0 = q_0 \geq 0$ and strictly monotone or constant, which is precisely what was imposed on x in Lemma 3.2. Second, it follows from (4.5) and (4.7) that if $x_0 = q_0 = 0$, then V_0^n, $n = 1, 2, \ldots$, are non-negative, as well as V_0. Observing that the argument of Φ^x is a continuous function therefore establishes that (4.9) is well-defined.

We now appeal to known results that support existence and uniqueness of the solution to (4.9). Review the four explicit forms of (4.9) listed at the end of Subsection 4.2. In Cases 1,3 and 4, a strong unique solution exists by the arguments, given in Section 5.6 of the book by Karatzas and Shreve [22]. In Case 2, $q \equiv 0$ and existence and uniqueness of (4.9) follows from Theorem 4.1 of Tanaka [44].

7.2. The set-up of strong approximations. We prove FCLT within the framework of strong approximations. For this, recall the strong approximation result presented in Ethier and Kurtz [14, Chapter 7, Corollary 5.5]. Adapted to our context, it guarantees the existence of a probability space on which a Poison process N and a Brownian motion B can jointly be realized so that

$$\sup_{t \geq 0} \frac{|\, N(t) - t - B(t) \,|}{\log(2 \vee t)} < \infty, \quad \text{a.s.}$$

Thus, we may start with two *independent* Brownian motions W_+ and W_- such that, for all $t \geq 0$, the following inequalities hold a.s.:

$$\left| N_+ \left(n \int_0^t \lambda(q_s^n)\, ds \right) - n \int_0^t \lambda(q_s^n)\, ds - W_+ \left(n \int_0^t \lambda(q_s^n)\, ds \right) \right| \leq K_+ \log \left(2 \vee n \int_0^t \lambda(q_s^n)\, ds \right), \tag{7.1}$$

$$\left| N_- \left(n \int_0^t \mu(q_s^n)\, ds \right) - n \int_0^t \mu(q_s^n)\, ds - W_- \left(n \int_0^t \mu(q_s^n)\, ds \right) \right| \leq K_- \log \left(2 \vee n \int_0^t \lambda(q_s^n)\, ds \right), \tag{7.2}$$

for some random variables K_+ and K_-. Assume further that

$$\lim_{n \uparrow \infty} V_0^n = V_0, \quad \text{a.s.}, \tag{7.3}$$

and V_0 is independent of W_+ and W_-. (See the assumptions on the primitives in Section 2.)

7.3. The main steps. As explained in Subsection 3.2, the limit process V is sometimes continuous over $[0, \infty)$ or $(0, \infty)$, in which case we have U-convergence. (Note that this depends only on x.) To prove the

theorem for U-convergence, we construct a sequence $\{\widetilde{V}^n\}$ of continuous path stochastic processes such that, for all $T > 0$,

$$\text{(7.4)} \qquad \begin{cases} \mathrm{P} - \lim\limits_{n\uparrow\infty} \left\| V^n - \widetilde{V}^n \right\|_T = 0, \\ \widetilde{V}^n \stackrel{\mathrm{d}}{=} V. \end{cases}$$

The assertion of the theorem now follows from Theorem 4.1 of Billingsley [4], preparing the ground for M_1-convergence. We now outline the main steps. Technical details are given subsequently.

Start by rewriting the expression for V^n (defined by (4.7)) in a form that is amenable to further calculations. To this end, introduce the sequence of processes $\widetilde{X}^n = \{\widetilde{X}^n_t, t \geq 0\}$, by

$$\text{(7.5)} \qquad \begin{aligned} \widetilde{X}^n_\bullet &= nq^n_0 + n\int_0^\bullet (\lambda(q^n_s) - \mu(q^n_s))\, ds \\ &\quad + W_+\left(n\int_0^\bullet \lambda(q^n_s)\, ds\right) - W_-\left(n\int_0^\bullet \mu(q^n_s)\, ds\right). \end{aligned}$$

Then, according to (4.3) and (4.5),

$$\text{(7.6)} \qquad \begin{aligned} V^n &= \sqrt{n}\,[\Phi(x^n) - \Phi(x)] \\ &= \left[\Phi\left(\sqrt{n}\,x^n\right) - \Phi\left(\frac{1}{\sqrt{n}}\widetilde{X}^n\right)\right] + \left[\Phi\left(\frac{1}{\sqrt{n}}\widetilde{X}^n\right) - \Phi\left(\sqrt{n}\,x\right)\right]. \end{aligned}$$

Equations (7.5) and (7.6) imply that

$$\text{(7.7)} \qquad V^n = \Delta^n + \left[\Phi(\sqrt{n}\,x + H^n + \epsilon^n) - \Phi(\sqrt{n}\,x)\right].$$

Here the processes $H^n = \{H^n_t, t \geq 0\}$, $n = 1, 2, \ldots$, are given by

$$\text{(7.8)} \qquad \begin{aligned} H^n_\bullet &= V^n_0 + \int_0^\bullet (\lambda'(q_s) - \mu'(q_s))V^n_s\, ds \\ &\quad + \frac{1}{\sqrt{n}}W_+\left(n\int_0^\bullet \lambda(q_s)\, ds\right) - \frac{1}{\sqrt{n}}W_-\left(n\int_0^\bullet \mu(q_s)\, ds\right), \end{aligned}$$

and the processes Δ^n and ϵ^n, by

$$\text{(7.9)} \qquad \Delta^n = \Phi\left(\sqrt{n}\,x^n\right) - \Phi\left(\frac{1}{\sqrt{n}}\widetilde{X}^n\right),$$

$$\text{(7.10)} \qquad \epsilon^n = \bar{\epsilon}^n + \epsilon^n_+ - \epsilon^n_-,$$

$$\text{(7.11)} \qquad \begin{aligned} \bar{\epsilon}^n(\bullet) &= \sqrt{n}\int_0^\bullet (\lambda(q^n_s) - \lambda(q_s))\, ds - \sqrt{n}\int_0^\bullet (\mu(q^n_s) - \mu(q_s))\, ds \\ &\quad - \int_0^\bullet (\lambda'(q_s) - \mu'(q_s))\, V^n_s\, ds, \end{aligned}$$

$$\text{(7.12)} \qquad \epsilon^n_+(\bullet) = \frac{1}{\sqrt{n}}W_+\left(n\int_0^\bullet \lambda(q^n_s)\, ds\right) - \frac{1}{\sqrt{n}}W_+\left(n\int_0^\bullet \lambda(q_s)\, ds\right),$$

$$\text{(7.13)} \qquad \epsilon^n_-(\bullet) = \frac{1}{\sqrt{n}}W_-\left(n\int_0^\bullet \mu(q^n_s)\, ds\right) - \frac{1}{\sqrt{n}}W_-\left(n\int_0^\bullet \mu(q_s)\, ds\right).$$

We claim that there exist processes $\widetilde{V}^n = \{\widetilde{V}^n_t, t \geq 0\}$ and $\widetilde{H}^n = \{\widetilde{H}^n_t, t \geq 0\}$, such that

$$(7.14)\quad \begin{cases} \widetilde{V}^n = \Phi^x(\widetilde{H}^n), \\ \widetilde{H}^n_\bullet = V_0 + \displaystyle\int_0^\bullet (\lambda'(q_s) - \mu'(q_s))\widetilde{V}^n_s \, ds + \frac{1}{\sqrt{n}} W_+ \left(n \int_0^\bullet \lambda(q_s)\, ds \right) \\ \qquad\qquad - \dfrac{1}{\sqrt{n}} W_- \left(n \displaystyle\int_0^\bullet \mu(q_s)\, ds \right). \end{cases}$$

Note that the arguments, used to establish that (4.9) is well–defined and possesses a unique strong solution, apply equally to (7.14). For convenience, rewrite equation (4.9) in a form similar to (7.14):

$$(7.15)\quad \begin{cases} V &= \Phi^x(H), \\ H_\bullet &= V_0 + \displaystyle\int_0^\bullet (\lambda'(q_s) - \mu'(q_s)) V_s \, ds \\ &\quad + \displaystyle\int_0^\bullet \sqrt{\lambda(q_s) + \mu(q_s)}\, dW_s \,. \end{cases}$$

In view of the scale invariance property of any Brownian motion B, $B(\bullet) \stackrel{\mathrm{d}}{=} B(n\bullet)/\sqrt{n}$, and because

$$W_+ \left(\int_0^\bullet \lambda(q_s)\, ds \right) - W_- \left(\int_0^\bullet \mu(q_s)\, ds \right) \stackrel{\mathrm{d}}{=} \int_0^\bullet \sqrt{\lambda(q_s) + \mu(q_s)}\, dW_s,$$

the relations (7.14) and (7.15) yield that for all n:

$$(7.16)\quad \begin{cases} \widetilde{V}^n \stackrel{\mathrm{d}}{=} V, \\ \widetilde{H}^n \stackrel{\mathrm{d}}{=} H. \end{cases}$$

Comparing now (7.16) with (7.4) reveals that only the first assertion of (7.4) remains to be proved. For this, rewrite (7.7) in the form

$$(7.17)\quad \begin{aligned} V^n &= \Delta^n + \left[\Phi(\sqrt{n}\, x + H^n + \epsilon^n) - \Phi(\sqrt{n}\, x + \widetilde{H}^n) \right] \\ &\quad + \left[\Phi(\sqrt{n}\, x + \widetilde{H}^n) - \Phi(\sqrt{n}\, x) \right], \end{aligned}$$

that enables us to sketch the general idea behind the rest of the convergence proof. It will be shown that the first (Δ^n) and second terms in (7.17) (the expression within the first pair of brackets on the right-hand side) converge in probability to zero, as $n \uparrow \infty$, with respect to the U-topology. The last term in (7.17) is then shown to converge weakly in the M_1-topology to $\Phi^x(H) = V$. The proof of the theorem is thus complete, by the continuity property of addition (see Appendix B) and because the limits of the first and second terms in (7.17) are continuous and non-random. (See

Theorem 4.4 by Billingsley [4] and the paper by Whitt [47] for more details.) However, the convergence proof for the second term in (7.17) is not straightforward, since this term itself depends on V^n (see (7.8)). As a standard tool in such situations, Gronwall's inequality will be used.

The proof will be carried out in two steps—first, U-convergence, followed by M_1-convergence.

Remark 7.1. To prove the general case, given by (4.8) with nonzero f_λ , f_μ , one can replicate all the considerations, but with

$$\begin{aligned}\widetilde{X}^n_\bullet &= nq^n_0 + n\int_0^\bullet (\lambda(q^n_s) - \mu(q^n_s))\,ds + \sqrt{n}\int_0^\bullet (f_\lambda(q^n_s) - f_\mu(q^n_s))\,ds \\ &\quad + W_+\left(n\int_0^\bullet \lambda(q^n_s)\,ds\right) - W_-\left(n\int_0^\bullet \mu(q^n_s)\,ds\right),\end{aligned}$$

$$\begin{aligned}H^n_\bullet &= V^n_0 + \int_0^\bullet (\lambda'(q_s) - \mu'(q_s))V^n_s\,ds + \int_0^\bullet (f_\lambda(q_s) - f_\mu(q_s))\,ds \\ &\quad + \frac{1}{\sqrt{n}}W_+\left(n\int_0^\bullet \lambda(q_s)\,ds\right) - \frac{1}{\sqrt{n}}W_-\left(n\int_0^\bullet \mu(q_s)\,ds\right),\end{aligned}$$

instead of (7.5) and (7.8).

7.4. U-convergence. In this subsection we prove the theorem for those cases where, as $n \uparrow \infty$,

$$\Phi(\sqrt{n}\,x + b) - \Phi(\sqrt{n}\,x) \longrightarrow \Phi^x(b), \quad \text{u.o.c.}, \tag{7.18}$$

for all $b \in C^x[0,\infty)$. (See Cases 1,2 and 3 in Subsection 3.2.)

We fix $T > 0$, restrict attention to the interval $[0,T]$ and verify the first assertion in (7.4). Subtracting the expression for $\widetilde{V}^n$, given by (7.14), from (7.17) and using the Lipschitz property of Φ (C being the Lipschitz constant), one can write for $t \le T$:

$$\begin{aligned}\left|\left|V^n - \widetilde{V}^n\right|\right|_t &\le ||\Delta^n||_T + C\,||\epsilon^n||_T + C\left|\left|H^n - \widetilde{H}^n\right|\right|_t \\ &\quad + \left|\left|\Phi(\sqrt{n}\,x + \widetilde{H}^n) - \Phi(\sqrt{n}\,x) - \Phi^x(\widetilde{H}^n)\right|\right|_T .\end{aligned} \tag{7.19}$$

For the third term on the right-hand side of (7.19) we have ($t \le T$):

$$\left|\left|H^n - \widetilde{H}^n\right|\right|_t \le |V^n_0 - V_0| + C_T\int_0^t \left|\left|V^n - \widetilde{V}^n\right|\right|_s ds, \tag{7.20}$$

where $C_T = ||\lambda'(q_s) - \mu'(q_s)||_T$ is finite by the continuity of λ', μ' and q. Combining (7.19) with (7.20) and applying Gronwall's inequality yields

$$\begin{aligned}&\left|\left|V^n - \widetilde{V}^n\right|\right|_t \\ &\quad \le (C\,|V^n_0 - V_0| + ||\Delta^n||_T + C\,||\epsilon^n||_T + ||\epsilon^n_\Phi||_T)e^{T\cdot C\cdot C_T},\end{aligned} \tag{7.21}$$

where the process ϵ_Φ^n is given by

$$\epsilon_\Phi^n = \Phi(\sqrt{n}\,x + \widetilde{H}^n) - \Phi(\sqrt{n}\,x) - \Phi^x(\widetilde{H}^n).$$

In view of (7.3), U-convergence will be established once it is shown that the second, third and fourth terms in the parentheses on the right-hand side of (7.21) converge in probability to zero. Each of these terms will now be analyzed separately.

Review the definition (7.9) of Δ^n. Denote the quantities between the absolute value signs on the left-hand side in (7.1) and in (7.2) by $\Delta_\lambda^n(t)$ and $\Delta_\mu^n(t)$ respectively. It follows, by the Lipschitz property of Φ (C being the Lipschitz constant), that

$$(7.22)\quad \|\Delta^n\|_T \le C\left\|\sqrt{n}\,x^n - \frac{1}{\sqrt{n}}\widetilde{X}^n\right\|_T \le C\frac{1}{\sqrt{n}}(\|\Delta_\lambda^n\|_T + \|\Delta_\mu^n\|_T).$$

Due to the FSLLN (or by Lemma 6.1) and by the locally Lipschitz property of λ and μ, relations (7.1), (7.2) and (7.22) imply the convergence, as $n \uparrow \infty$:

$$(7.23)\qquad \frac{\Delta_\lambda^n}{\sqrt{n}}, \frac{\Delta_\mu^n}{\sqrt{n}}, \Delta^n \longrightarrow 0 \ \ u.o.c., a.s.$$

Review now the definition (7.10) of ϵ^n. We show that each term in the sum on the right-hand side of (7.10) converges in probability to zero with respect to the U-topology, and, hence, ϵ^n does so too.

It will be shown in Lemma 7.1 (see Subsection 7.6) that

$$(7.24)\qquad \lim_{n\uparrow\infty} \|\bar{\epsilon}^n\|_T = 0, \ \text{a.s.}$$

We now apply Lemma 7.2 presented in Subsection 7.6, with

$$g_\bullet^n = \int_0^\bullet \lambda(q_s^n)\,ds, \ \ g_\bullet = \int_0^\bullet \lambda(q_s)\,ds,$$

to get

$$(7.25)\qquad \mathrm{P} - \lim_{n\uparrow\infty} \|\epsilon_+^n\|_T = 0.$$

(The conditions of Lemma 7.2 are satisfied by the FSLLN and because of the auxiliary assertion (7.30) obtained in Lemma 7.1.) Similarly, we obtain

$$(7.26)\qquad \mathrm{P} - \lim_{n\uparrow\infty} \|\epsilon_-^n\|_T = 0.$$

Finally, (7.16) and (7.18) imply

$$(7.27)\qquad \mathrm{P} - \lim_{n\uparrow\infty} \|\epsilon_\Phi^n\|_T = 0.$$

This completes the proof of U-convergence.

7.5. M_1-convergence. We now consider the case

$$\Phi(\sqrt{n}\,x+b)-\Phi(\sqrt{n}\,x)\longrightarrow\Phi^x(b),\quad \text{as } n\uparrow\infty,$$

in the M_1-topology, but not in the U-topology (see Cases 4 in Subsection 3.2). Then, combining (7.17) with (7.16) reveals that the third term in (7.17), namely the expression within the last pair of brackets, converges weakly in the M_1-topology to $\Phi^x(H)=V$. It will be shown further that the second term in (7.17) converges in probability to zero with respect to the U-topology. In view of (7.23), the proof is then complete. (See Theorem 4.4 by Billingsley [4] and the paper by Whitt [47] and recall the arguments at the end of Subsection 7.3.)

In order to show that the second term in (7.17) converges to zero in the U-topology, we apply the Lipschitz property of Φ^x to this term and obtain ($t\le T$):

$$\begin{aligned}&\left|\left|\Phi(\sqrt{n}\,x+H^n+\epsilon^n)-\Phi(\sqrt{n}\,x+\widetilde{H}^n)\right|\right|_t\\ &\qquad\le C\,||\epsilon^n||_T+C\left|\left|H^n-\widetilde{H}^n\right|\right|_T.\end{aligned}\tag{7.28}$$

It will be shown further that the last term on the right–hand side of (7.28) converges to zero in probability. Then, combining (7.28) with definition (7.10) of ϵ^n and using (7.24)–(7.26) proves the desired assertion.

In order to deduce that the last term in (7.28) converges in probability to zero, review inequality (7.20). Since the first term on the right-hand side of (7.20) converges to zero a.s. (by (7.3)), it is sufficient to check that the second term in (7.20), for $t=T$, converges in probability to zero. But this follows from Remark 3.3 and (7.30), and the fact that theorem is already proved for the case of U-convergence.

7.6. Lemmata.

LEMMA 7.1. *The sequence $\{\bar{\epsilon}^n\}$ given by (7.11) satisfies (7.24).*

Proof. Our calculations resemble those in Chapter 8,§3 of the book by Liptser and Shiryayev [36, pages 635,636], where they are presented in the context of martingale theory.

From the definition (7.11) of $\bar{\epsilon}^n$ and by the locally Lipschitz continuity of λ' and μ' we obtain, in view of the definition (4.7) of V^n:

$$\begin{aligned}|\bar{\epsilon}^n(t)| &\le \int_0^t \left(\left|\lambda'(q_s+\varphi_s^1(q_s^n-q_s))\sqrt{n}(q_s^n-q_s)-\lambda'(q_s)V_s^n\right|\right.\\ &\qquad\left.+\left|\mu'(q_s+\varphi_s^2(q_s^n-q_s))\sqrt{n}(q_s^n-q_s)-\mu'(q_s)V_s^n\right|\right)ds\\ &\le C_T\,||V^n||_t\,||q^n-q||_t\,t,\quad t\le T,\end{aligned}\tag{7.29}$$

where $\varphi_s^1,\varphi_s^2\in[0,1]$, and C_T is a constant. By the FSLLN (Theorem 4.1), it follows from the last equation that to prove the lemma it is sufficient to

show that the following holds a.s.:

$$\overline{\lim_n} ||V^n||_t < \infty, \quad t \le T. \tag{7.30}$$

In order to prove (7.30) we continue as follows. From the definition of V^n (see (4.7)) one obtains

$$\begin{aligned} ||V^n||_t &\le C\sqrt{n}\,||x^n - x||_t \\ &\le C\,|V_0^n| + C\left|\left|\sqrt{n}\left(\frac{1}{n}N_+\left(n\int_0^\bullet \lambda(q_s^n)\,ds\right) - \int_0^\bullet \lambda(q_s^n)\,ds\right)\right|\right|_t \\ &\quad + C\left|\left|\sqrt{n}\left(\frac{1}{n}N_-\left(n\int_0^\bullet \mu(q_s^n)\,ds\right) - \int_0^\bullet \mu(q_s^n)\,ds\right)\right|\right|_t \\ &\quad + C\left|\left|\sqrt{n}\int_0^t (\lambda(q_s^n) - \lambda(q_s))\,ds\right|\right|_t \\ &\quad + C\left|\left|\sqrt{n}\int_0^t (\mu(q_s^n) - \mu(q_s))\,ds\right|\right|_t, \quad t \le T. \end{aligned}$$

Using the FSLLN and the local Lipschitz continuity of λ and μ, we obtain the existence of a (possibly) random M, and positive non-random scalars F_T, L_T, such that for all $n \ge M$ the following inequality holds a.s. ($t \le T$):

$$\begin{aligned} ||V^n||_t &\le C\,|V_0^n| + C\left|\left|\sqrt{n}\left(\frac{1}{n}N_+(n\bullet) - \bullet\right)\right|\right|_{F_T} \\ &\quad + C\left|\left|\sqrt{n}\left(\frac{1}{n}N_-(n\bullet) - \bullet\right)\right|\right|_{F_T} + L_T\int_0^t ||V^n||_s\,ds. \end{aligned} \tag{7.31}$$

Note that, as $n \uparrow \infty$,

$$\sqrt{n}\left(\frac{1}{n}N_+(n\bullet) - \bullet\right) \longrightarrow W_+, \quad \text{a.s.}$$

and analogously for N_-. This fact, the convergence (7.3) of $\{V_0^n\}$ and Gronwall's inequality applied to (7.31) complete the proof of the lemma. □

Condition (7.30) implies the so-called compact containment condition (see Ethier and Kurtz [14, page 129]):

$$\lim_{\ell\uparrow\infty} \overline{\lim_n}\, \mathrm{P}\{||V^n||_t > \ell\} = 0, \quad t \le T.$$

This condition is often involved in proving weak limit theorems and is used in the following

LEMMA 7.2. *Let $\{g^n\}$ be a sequence of stochastic processes with monotone increasing sample paths. Let g be a monotone increasing deterministic function, and let B denote a Brownian motion. Further, for all n, let $g_0^n = g_0 = 0$. Assume, in addition, that for all $T > 0$,*

$$\lim_{n\uparrow\infty} ||g^n - g||_T = 0, \quad \text{a.s.}, \tag{7.32}$$

and

$$\lim_{\ell\uparrow\infty}\overline{\lim_{n}}\,\mathrm{P}\left\{\sqrt{n}\,\|g^n-g\|_T>\ell\right\}=0. \tag{7.33}$$

Then

$$\mathrm{P}-\lim_{n\uparrow\infty}\frac{1}{\sqrt{n}}\,\|B(ng^n)-B(ng)\|_T=0, \tag{7.34}$$

for all $T>0$.

Proof. Introduce the random variables T^n and $\widetilde{T}^n$ by

$$T^n=n\,\|g^n-g\|_T\ ,$$

$$\widetilde{T}^n=n\,(g^n(T)\vee g(T))\ .$$

Evidently, $0\le T^n\le\widetilde{T}^n$. Without loss of generality, consider the case $0<T^n<\widetilde{T}^n$.

Fix $\varepsilon>0$. By (7.33),

$$\lim_{\ell\uparrow\infty}\overline{\lim_{n}}\mathrm{P}\left\{\frac{T^n}{\sqrt{n}}>\ell\right\}=0.$$

In view of (7.32), we can choose $\ell_\varepsilon>0$, a natural number N_ε and a set B_ε such that for all $n>N_\varepsilon$

$$\mathrm{P}\left\{\frac{T^n}{\sqrt{n}}>\ell_\varepsilon\right\}<\varepsilon, \tag{7.35}$$

and

$$\begin{cases}\widetilde{T}^n\le F^n\triangleq 2n(g(T)\vee 1)\ \text{ on }\ B_\varepsilon,\\ \mathrm{P}\,\{B_\varepsilon^c\}\le\varepsilon.\end{cases}$$

Denoting

$$A^n=\frac{1}{\sqrt{n}}\,\|B(ng^n)-B(ng)\|_T\ ,$$

$$S(\alpha,\beta,\gamma)=\{u,v:0\le u,v\le\alpha,\ \beta<|\,u-v\,|\le\gamma\}\ ,$$

we obtain for $n > N_\varepsilon$:

$$\begin{aligned}
\mathrm{P}\{A^n > \varepsilon\} &\leq \mathrm{P}\left\{(A^n > \varepsilon) \bigcap B_\varepsilon\right\} + \mathrm{P}\{B_\varepsilon^c\} \\
&\leq \mathrm{P}\left\{\sup_{S(F^n,0,T^n)} |B(u) - B(v)| > \varepsilon\sqrt{n}\right\} + \varepsilon \\
&\leq \mathrm{P}\left\{\sup_{S(F^n,0,\ell_\varepsilon\sqrt{n})} |B(u) - B(v)| > \varepsilon\sqrt{n}\right\} \\
&\quad + \mathrm{P}\left\{\sup_{S(F^n,\ell_\varepsilon\sqrt{n},T^n)} |B(u) - B(v)| > \varepsilon\sqrt{n}\right\} + \varepsilon.
\end{aligned}$$

By (7.35), the second term on the right-hand side of the last inequality is less than ε, and therefore we restrict our attention to the first term only.

Recall Lemma 1.2.1 in the book by Csörgő and Révész [13], which asserts the following. For any positive δ, there exists a constant $C = C(\delta)$ such that the inequality

$$\mathrm{P}\left\{\sup_{S(F,0,\ell)} |B(u) - B(v)| > p\sqrt{\ell}\right\} \leq C(1 + \frac{F}{\ell})e^{-\frac{p^2}{2+\delta}}$$

holds for every positive p and $0 < \ell < F$. (This form of Lemma 1.2.1 in [13] is taken from [10].) Using this assertion and continuing our calculation, we obtain

$$\begin{aligned}
&\mathrm{P}\{A^n > \varepsilon\} \\
&\quad \leq 2\varepsilon + \mathrm{P}\left\{\sup_{S(F^n,0,\ell_\varepsilon\sqrt{n})} |B(u) - B(v)| > \frac{\varepsilon\sqrt{n}}{(\ell_\varepsilon\sqrt{n})^{1/2}}(\ell_\varepsilon\sqrt{n})^{1/2}\right\} \\
&\quad \leq 2\varepsilon + C\left(1 + \frac{2(g(T)\vee 1)\sqrt{n}}{\ell_\varepsilon}\right)e^{-\frac{\varepsilon^2\sqrt{n}}{\ell_\varepsilon(2+\delta)}},
\end{aligned}$$

which implies the assertion of the lemma.

7.7. Proof of Theorem 4.3. The proof of Theorem 4.3 is omitted, being similar to that of Theorem 4.2, except for the following comments. Recall that fluid limits q are strictly monotone or constant and reconsider (7.29) (the step in the proof where the Lipschitz properties of λ', μ' are used). By a simple modification of the arguments, one concludes that Theorem 4.2 holds without any changes, with the exception of the special situation (4.10). In that case, one must separate the analysis of (7.29) to the right and the left neighborhood of q_0.

8. Directions for future research. Of interest are extensions to the current model that cover time- and state-dependent rates, other performance measures such as waiting time and work-loads, and random or discontinuous λ and μ. The latter would enable, among other things, analysis of models with finite buffers, breakdowns and batch service.

Other possible extensions are to non-exponential models. The approach taken here should carry over, but the details would naturally depend on the particular model at hand. (See, for example, a steady-state analysis of state-dependent $M_\xi/G_\xi/1$ queues in Knessl *et al.* [26]; diffusion approximations of phase-type models in Whitt [48] and Krichagina [29]; fluid and diffusion approximations of various semi-Markovian models in Anisimov [1]).

Work is currently ongoing on approximating state–dependent networks, that includes state–dependent routing. Fluid limits for such networks are solutions to autonomous ordinary differential equations with *state–dependent* oblique reflection. Diffusion limits are solutions to stochastic differential equations with *time–dependent* oblique reflection. The diffusion limits are Markov processes with possibly discontinuous sample-paths. Weak convergence is with respect to Skorokhod's M_1*-topology.*

Fluid limits of networks (as solutions to a multi-dimensional differential equation) need not be monotone functions and can leave a boundary, after having reached it. As a consequence, the diffusion limits could have multiple points of discontinuity. Furthermore, the characterization of fluid and diffusion limits involves reflection problems with non-constant directions of reflections, varying with time and state. Such mappings are less well-behaved than the usual multi-dimensional Skorokhod maps (in particular, they need not be Lipschitz). All this suggests that new tools must be developed in order to establish convergence, existence and uniqueness of the limits.

A. Skorokhod's reflection problem. We use the following version of the one-dimensional Skorokhod's reflection problem (taken from [9]):

THEOREM. *For any* $x \in D_0[0,\infty)$, *there exist a unique pair* $(q,y) \in D_0[0,\infty) \times D_0[0,\infty)$ *satisfying*

$$\begin{cases} q_t = x_t + y_t \geq 0, \quad t \geq 0, \\ y \text{ } \textit{nondecreasing, with } y_0 = 0, \\ \displaystyle\int_0^\infty 1[q_t>0]\, dy_t = 0. \end{cases}$$

The operators Φ and Ψ with domain $D_0[0,\infty)$, given by

$$q = \Phi(x), \quad y = \Psi(x),$$

are both Lipschitz continuous with respect to the uniform norm. Namely, there exists a constant $C > 0$, such that

$$\begin{aligned} \left|\left|\Phi(x^1) - \Phi(x^2)\right|\right|_T &\leq C \left|\left|x^1 - x^2\right|\right|_T , \\ \left|\left|\Psi(x^1) - \Psi(x^2)\right|\right|_T &\leq C \left|\left|x^1 - x^2\right|\right|_T , \end{aligned}$$

for all $x^1, x^2 \in D_0[0,\infty)$ and $T > 0$. Furthermore, Φ and Ψ are both homogeneous of degree 1:

$$\begin{aligned} \Phi(\gamma x) &= \gamma \Phi(x), \\ \Psi(\gamma x) &= \gamma \Psi(x), \end{aligned}$$

for all $x \in D_0[0,\infty)$ and $\gamma \geq 0$.

Note that both the theorem cited above and all properties of Φ and Ψ hold when we use, instead of $D_0[0,\infty)$, the space of the $\mathbb{R}^d$-valued RCLL functions with non-negative values at zero. However, only in the one-dimensional case do Φ and Ψ have the explicit forms:

$$\begin{aligned} \Psi_t(x) &= \sup_{0 \leq s \leq t} (x_s^-), \quad t \geq 0, \\ \Phi(x) &= x + \Psi(x) = x + \overline{x^-}. \end{aligned}$$

B. Weak convergence. We use in this paper the set $\widetilde{D}[0,\infty)$ of all real-valued functions on $[0,\infty)$ with right and left limits at each point. Values of functions are assumed to be equal to either the left or the right limit. Note that discontinuities at zero are admissible.

Our weak convergence results are proved for the space $(\widetilde{D}[0,\infty), M_1)$, that is $\widetilde{D}[0,\infty)$ endowed with Skorokhod's M_1-topology, see [43]. The appropriate definitions of the M_1-topology and, respectively M_1-convergence, for $\widetilde{D}[0,\infty)$ (which slightly differs from the space used in [43]) can be given within the unified graph approach of Pomarede [40]. For the extension of Pomarede's definitions to the non-compact interval $[0,\infty)$, see, e.g., Whitt [45] and [47].

We use the following properties of the M_1-topology:

1. Let $\{x^n\}$ converge to x in the M_1-topology. If x is an element of $C[0,\infty)$, then the M_1-topology reduces to the topology of uniform convergence on compact sets (U-topology). Uniform convergence is referred to as U-convergence.

2. Theorem 3.1 by Pomarede [40], on M_1-convergence: Let $x^n \longrightarrow x$, $y^n \longrightarrow y$, as $n \uparrow \infty$. Then, $x^n + y^n \longrightarrow x + y$, if x and y have no common points of discontinuity.

In Theorem 4.3 we use the ordinary Skorokhod space $(D[0,\infty), J_1)$.

C. Notation.

RCLL	right-continuous with left limits
u.o.c	uniformly on compact
$1[S]$	indicator function of a set S
$f_t \uparrow\uparrow a$	f is strictly increasing and $\lim_{t\uparrow\infty} f_t = a$
$f_t \downarrow\downarrow a$	f is strictly decreasing and $\lim_{t\uparrow\infty} f_t = a$
$\vee$ and $\wedge$	maximum and minimum
$a^- = -(a \wedge 0)$	the negative part of a
$\overline{f}(t) = \sup_{0\le s\le t} f_s$	the upper envelope of f
$\|\|f\|\|_T = \sup_{0\le s\le T} \mid f \mid$	the uniform norm of f on the interval $[0,T]$
$\mathcal{Z}^+$ and $\mathbb{R}^+$	the sets of non-negative integer and real numbers
$C[0,\infty)$	the set of continuous real-valued functions on $[0,\infty)$
$C_0[0,\infty)$	$\{f \in C[0,\infty) \mid f_0 \ge 0\}$
$D[0,\infty)$	the set of RCLL real-valued functions
$D_0[0,\infty)$	$\{f \in D[0,\infty) \mid f_0 \ge 0\}$
$D_E[0,\infty)$	the set of RCLL E-valued functions
$\widetilde{D}[0,\infty)$	see Appendix B
$(\widetilde{D}[0,\infty), J_1)$	the space $\widetilde{D}[0,\infty)$ endowed with Skorokhod's J_1-topology
$(\widetilde{D}[0,\infty), M_1)$	the space $\widetilde{D}[0,\infty)$ endowed with Skorokhod's M_1-topology
$\overset{d}{\sim}$	is distributed as
$\xrightarrow{\mathrm{d}}$	convergence in distribution
$\mathrm{P}-\lim$	limit in probability
$\mathcal{N}(\delta,\sigma^2)$	the normal distribution with mean δ and variance σ^2
$BM(\delta,\sigma^2)$	Brownian motion with drift δ and variance σ^2, starting at 0
$BM_x(\delta,\sigma^2)$	Brownian motion with drift δ and variance σ^2, starting at x
$RBM(\delta,\sigma^2)$	Reflected Brownian motion with drift δ and variance σ^2, starting at 0
$RBM_x(\delta,\sigma^2)$	Reflected Brownian motion with drift δ and variance σ^2, starting at x

REFERENCES

[1] V. V. Anisimov. Switching processes: Asymptotic theory and applications. In A. N. Shiryayev *et. al.*, editors, *New Trends in Probability and Statistics.* To appear.

[2] S. V. Anulova. Functional limit theorems for network of queues. In *IFAC Congress*, Tallinn, 1990. Abstract.

[3] A. D. Barbour. On a functional central limit theorems for Markov population processes. *Advances in Applied Probability*, 6:21–39, 1974.

[4] P. Billingsley. *Convergence of Probability Measures.* John Wiley and Sons, New York, 1968.

[5] A. Borovkov. Some limit theorems in the theory of mass service, II. *Theory of Probability and Its Applications*, 10:375–400, 1965.

[6] A. Borovkov. On limit laws for service processes in multi-channel systems. *Siberian Mathematical Journal*, 8:746–763, 1967.

[7] P. Bremaud. *Point Processes and Queues: Martingale Dynamics.* Springer-Verlag, Berlin, 1981.

[8] H. Chen and A. Mandelbaum. Discrete flow networks: Diffusion approximations and bottlenecks. *The Annals of Probability*, 19:1463–1519, 1991.

[9] H. Chen and A. Mandelbaum. Hierarchical modelling of stochastic networks, part I: Fluid models. In D. Yao, editor, *Stochastic Modelling and Analysis of Manufacturing Systems*, pages 47–105. Springer-Verlag, New York, 1994.

[10] H. Chen and A. Mandelbaum. Hierarchical modelling of stochastic networks, part II: Strong Approximations. In D. Yao, editor, *Stochastic Modelling and Analysis of Manufacturing Systems*, pages 107–131. Springer-Verlag, New York, 1994.

[11] E. G. Coffman, Jr., A. A. Puhalskii, M. I. Reiman, and P. Wright. Processor shared buffers with reneging. *Performance Evaluation*, 19:25–46, 1994.

[12] E. G. Coffman, Jr. and M. I. Reiman. Diffusion approximation for computer communication systems. In G. Iazeolla, P. J. Courtois, and A. Hordijk, editors, *Mathematical Computer Performance and Reliability*, pages 33–53. North-Holland, Amsterdam, 1984.

[13] M. Csörgő and P. Révész. *Strong Approximations in Probability and Statistics.* Academic Press, New York, 1981.

[14] S. N. Ethier and T. G. Kurtz. *Markov Process: Characterization and Convergence.* John Wiley and Sons, New York, 1986.

[15] P. W. Glynn. Diffusion approximations. In D. P. Heyman and M. J. Sobel, editors, *Handbooks in Operations Research and Management Science, Vol. 2*, pages 145–198. North-Holland, Amsterdam, 1990.

[16] J. Hale. *Ordinary Differential Equations.* J. Wiley & Sons/Interscience, New York, 1969.

[17] S. Halfin and W. Whitt. Heavy-traffic limits theorem for queues with many exponential servers. *Operations Research*, 29:567–588, 1981.

[18] D. L. Iglehart. Limit diffusion approximations for the many-server queue and repairman problem. *Journal of Applied Probability*, 2:429–441, 1965.

[19] D. L. Iglehart and A. J. Lemoine. Approximations for the repairman problem with two repair facilities, I: No spares. *Advances in Applied Probability*, 5:595–613, 1973.

[20] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic, I. *Advances in Applied Probability*, 2:150–177, 1970.

[21] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic, II: Sequences, networks, and batches. *Advances in Applied Probability*, 2:355–364, 1970.

[22] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus.* Springer-Verlag, New York, 1988.

[23] H. Kaspi and A. Mandelbaum. Regenerative closed queueing networks. *Stochastics*

and Stochastics Reports, 39:239–258, 1992.

[24] J. F. C. Kingman. The single server queue in heavy traffic. *Proceedings of the Cambridge Philosophical Society*, 57:902–904, 1961.

[25] J. F. C. Kingman. The heavy traffic approximations in the theory of queues. In W. Smith and W. Wilkinson, editors, *Proceedings of the Symposium on Congestion Theory*, pages 137–159. The University of North California Press, Chapel Hill, 1965.

[26] C. Knessl, B. J. Matkowsky, Z. Schuss, and C. Tier. On the performance of state-dependent single server queues. *SIAM Journal on Applied Mathematics*, 46(4):657–697, August 1986.

[27] Y. A. Kogan and R. S. Liptser. Limit non-stationary behavior of large closed queueing networks with bottlenecks. *Queueing Systems*, 14:33–55, 1993.

[28] Y. A. Kogan, R. S. Liptser, and A. V. Smorodinskii. Gaussian diffusion approximation of closed Markov models of computer networks. *Problems of Information Transmission*, 22(1):38–51, 1986.

[29] E. V. Krichagina. Diffusion approximation for a queue in a multiserver system with multistage service. *Automation and Remote Control*, 50(3):346–354, 1989.

[30] E. V. Krichagina. Asymptotic analysis of queueing networks (martingale approach). *Stochastics and Stochastics Report*, 40:43–76, 1992.

[31] T. G. Kurtz. Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and Their Applications*, 6:223–240, 1978.

[32] T. G. Kurtz. Representation of markov processes as multiparameter time changes. *The Annals of Probability*, 8(4):682–715, 1980.

[33] T. G. Kurtz. Representation and approximation of counting processes. In W. H. Fleming and L. G. Gorostiza, editors, *Advances in Filtering and Optimal Stochastic Control, Lecture Notes in Control and Information Sci. 42*, pages 177–191. Springer-Verlag, Berlin, 1982.

[34] T. G. Kurtz. Gaussian approximations for markov chains and counting processes. In *44th Session of the International Statistical Institute*, Madrid, Spain, September 1983.

[35] A. J. Lemoine. Networks of queues—a survey of weak convergence results. *Management Science*, 24:1175–1193, 1978.

[36] R. S. Liptser and A. N. Shiryayev. *Theory of Martingales.* Kluwer Academic Publishers, 1989.

[37] A. Mandelbaum and W. A. Massey. Strong approximations for time–dependent queues. To be published in Mathematics of Operations Research, 1994.

[38] G. F. Newell. *Applications of Queueing Theory.* Chapman and Hall, 1982.

[39] R. M. Oliver and A. H. Samuel. Reducing letter delays in post offices. *Operations Research*, 10:839–892, 1962.

[40] J. L. Pomarede. *A Unified Approach via Graphs to Skorokhod's Topologies on the Function Space.* PhD thesis, Department of Statistics, Yale University, 1976.

[41] N. Prabhu. *Stochastic Storage Processes, Queues, Insurance Risk and Dams.* Springer-Verlag, New York, 1980.

[42] M. I. Reiman. Some diffusion approximations with state-space collapse. In F. Baccelli and G. Fayolle, editors, *Modelling and Performance Evaluation Methodology.* Springer-Verlag, 1984.

[43] A. V. Skorokhod. Limit theorems for stochastic processes. *Theory of Probability and Its Applications*, 1:261–290, 1956.

[44] H. Tanaka. Stochastic differential equations with reflected boundary conditions in convex region. *Hiroshima Mathematical Journal*, 9:163–174, 1974.

[45] W. Whitt. Weak convergence of first passage time processes. *Journal of Applied Probability*, 8:417–422, 1971.

[46] W. Whitt. Heavy traffic theorems for queues: A survey. In A. B. Clarke, editor, *Mathematical Methods in Queueing Theory*, pages 307–350. Springer-Verlag, Berlin, 1974.

[47] W. Whitt. Some useful functions for functional limit theorems. *Mathematics of*

Operations Research, 5(1):67–85, February 1980.

[48] W. Whitt. On the heavy traffic limit theorem for $GI/G/\infty$ queues. *Advances in Applied Probability*, 14:171–190, 1982.

[49] K. Yamada. Multi-dimensional Bessel processes as heavy traffic limits of certain tandem queues. *Stochastic Processes and Their Applications*, 23:35–56, 1986.

[50] K. Yamada. Diffusion approximations for open state-dependent queueing networks under heavy traffic situation. Technical report, Institute of Information Science and Electronics, University of Tsukuba, Tsukuba, Ibaraki 305, Japan, 1993.