A diffusion regime with non-degenerate slowdown: Appendix

Rami Atar

June 15, 2011

Abstract

In this appendix to "A diffusion regime with non-degenerate slowdown", *Op. Res. (2012)*, we provide the proofs of the main results. Theorem 2.1 is proved in Subsection A.1 and Theorem 2.2 in Subsection A.2.

For the notation used in this appendix, see the end of Section 1 of the main body of the paper. Equation and theorem numbers, such as (1), (2),..., Theorem 2.1, etc. refer to the main body of the paper. Literature citation such as [1], [2],... refers to the bibliographic list at the end of this appendix.

A.1 Convergence of queue-length

In this subsection we prove Theorem 2.1. Let us introduce some notation. Let

$$T_n(t) = \sum_{k\in K_n} T_{kn}(t), \qquad \bar{T}_n(t) = n^{-1}T_n(t), \tag{A.1}$$

$$\widehat{A}_n(t) = n^{-1/2}(A_n(t) - \lambda_n t),$$

$$V_n(t) = n^{-1/2}\sum_{k\in K_n}(D_{kn}(t) - T_{kn}(t)) = n^{-1/2}(D_n(t) - T_n(t)), \tag{A.2}$$

$$Z^*_n(t) = n^{-1/2}Z(n^{1/2}t) - t, \quad t \ge 0, \tag{A.3}$$

$$F_n(t) = Z^*_n\Big(\gamma_n\int_0^t \widehat{X}_n(s)^+ds\Big) + (\gamma_n - \gamma)\int_0^t \widehat{X}_n(s)^+ds, \tag{A.4}$$

$$W_n(t) = \widehat{A}_n(t) - V_n(t), \qquad \widetilde{W}_n(t) = (\widehat{\lambda}_n - \widehat{\mu}_n)t + W_n(t) - F_n(t). \tag{A.5}$$

Lemma A.1. *The following identities hold*

$$\widehat{X}_n(t) = \widehat{X}_n(0) + \widetilde{W}_n(t) + L_n(t) - \gamma\int_0^t \widehat{X}_n(s)^+ds, \tag{A.6}$$

$$T_n(t) + \sqrt{n}L_n(t) = \mu_n t. \tag{A.7}$$

Proof. It follows from (11), (16), (18) and (20) that

$$\widehat{X}_n(t) = \widehat{X}_n(0) + W_n(t) + \widehat{\lambda}_n t + n^{1/2}\lambda t$$
$$- n^{-1/2}(T_n(t) - \mu_n t) - \widehat{\mu}_n t - n^{1/2}\mu t - \gamma \int_0^t \widehat{X}_n(s)^+ ds - F_n(t).$$

Noting by (10) that the fourth and seventh terms on the r.h.s. of the above display cancel out, and using the notation just introduced, we obtain (A.6). Next, (A.7) follows from the identity $B_{kn} + I_{kn} = 1$, along with definitions (6), (13), (22) and (A.1). □

Note the resemblance between (A.6) and the Skorohod equation (24). The main ingredient of the proof will be to show that $\widetilde{W}_n$ converge to a Brownian motion, for which we use martingale convergence methods (see Lemma A.3). We then argue by continuity of the solution map to the Skorohod equation (see the proof of Theorem 2.1).

It is well known that the scaled processes $\widehat{A}_n$ converge in distribution, uniformly on compacts, to a Brownian motion starting from zero, with zero mean and diffusion coefficient $\lambda^{1/2}C_{IA}$ [5, Section 17]. We denote this limit process by w_A. It easily follows that Z_n^* converges to zero in distribution, uniformly on compacts.

In what follows fix $u \in [1, \infty)$. Fix also $\varrho \in (0, \frac{1}{2})$. Let

$$\tau_n = \inf\{t \geq 0 : L_n(t) \geq n^{\varrho}\} \wedge u. \tag{A.8}$$

We begin by proving some elementary estimates.

Lemma A.2. *With $\bar{T}(t) := \mu t$, $t \in [0, u]$, one has*

$$\bar{T}_n \to \bar{T} \text{ in probability, uniformly on } [0, u], \tag{A.9}$$

$$\sup_n \mathbb{E}[(|V_n|_u^*)^2] < \infty, \tag{A.10}$$

$$\text{the random variables } \quad |\widehat{X}_n^+|_u^* \quad \text{ are tight}, \tag{A.11}$$

$$F_n \to 0 \text{ in probability, uniformly on } [0, u], \tag{A.12}$$

and

$$\mathbb{P}(\tau_n < u) \to 0, \quad \text{as } n \to \infty. \tag{A.13}$$

Proof. First, let us show

$$\sup\{|\bar{T}_n(t) - \mu t| : t \leq \tau_n\} \to 0 \quad \text{in probability, as } n \to \infty. \tag{A.14}$$

Note by (13), and since B_{kn} take values in $\{0, 1\}$, that $\bar{T}_n(t) \leq \bar{\mu}_n t$, for every t (recall (7)). By (A.8), for $t \leq \tau_n$, $L_n(t) \leq n^{\varrho}$. Hence by (13), (A.1) and (22),

$$\bar{T}_n(t) = n^{-1} \sum_{k=1}^{N_n} \mu_{kn} \int_0^t (1 - I_{kn}(s))ds \geq \bar{\mu}_n t - n^{\varrho - 1/2}, \quad t \leq \tau_n.$$

As a result, the l.h.s. of (A.14) is bounded by $|\bar{\mu}_n - \mu|u + n^{\varrho - 1/2}$, for every n. By (7), we have (A.14).

We next show (A.10). We use the Burkholder-Davis-Gundy inequality, which states that for any local martingale M and $p \geq 1$,

$$\mathbb{E}\{(|M|_t^*)^p\} \leq c_p \mathbb{E}\{[M,M]_t^{p/2}\}, \quad t \in [0,\infty),$$

where the constant c_p depends only on p, and $[M,M]$ is the quadratic variation process defined by $[X,X] = X^2 - 2\int X_- dX$ (see [8] p. 58, and p. 175); if X has piecewise smooth sample paths, null at zero, then $[X,X]_t$ is given by $\sum_{s\leq t} \Delta X(s)^2$ (see for example [8], Theorem 22(ii), p. 59). It follows from the martingale assumption that V_n is an $\mathbb{F}_n$-martingale. Since each of its jumps is of size $n^{-1/2}$ and the number of jumps by time t is $D_n(t)$, it follows that

$$\begin{aligned}\mathbb{E}[(|V_n|_u^*)^2] &\leq \frac{c}{n}\mathbb{E}[D_n(u)] = \frac{c}{n}\mathbb{E}[T_n(u)] \\ &= \frac{c}{n}\mathbb{E}\Big[\sum_k \mu_{kn} \int_0^u B_{kn}(s)ds\Big] \\ &\leq cu\,\frac{\mu_n}{n},\end{aligned}$$

and by (7) follows (A.10).

We now prove

$$\text{the random variables} \quad |\widehat{X}_n^+|_{\tau_n}^* \quad \text{are tight.} \tag{A.15}$$

Let $r > 0$ be given. Consider the event $\{|\widehat{X}_n^+|_{\tau_n}^* > 2r\}$. On this event there exist $0 \leq \sigma_n \leq \theta_n \leq \tau_n$ such that $\widehat{X}_n(\sigma_n) \leq r$, $\widehat{X}_n(\theta_n) \geq 2r$, while $\widehat{X}_n(t) > 0$ for $t \in [\sigma_n, \theta_n]$. Note that the last assertion, in view of (11) and (20), implies that $I_n(t) = 0$ for $t \in [\sigma_n, \theta_n]$. Hence by (22), $L_n(\theta_n) = L_n(\sigma_n)$. We can now use (A.6) to bound the probability of the specified event. Indeed, by (16), (17), (21), (A.3) and (A.4), we have that

$$\gamma \int_0^t \widehat{X}_n(s)^+ ds + F_n(t) \equiv n^{-1/2} Z\Big(\gamma_n n^{1/2} \int_0^t \widehat{X}_n(s)^+ ds\Big) = n^{-1/2} Z_n(t) \tag{A.16}$$

is nondecreasing in t. Hence by (A.5) and (A.6), we have, on the event alluded to above, that there exist $0 \leq \sigma_n \leq \theta_n \leq \tau_n$ such that

$$(\widehat{\lambda}_n - \widehat{\mu}_n)(\theta_n - \sigma_n) + W_n(\theta_n) - W_n(\sigma_n) \geq \widehat{X}_n(\theta_n) - \widehat{X}_n(\sigma_n) \geq r.$$

This shows

$$\mathbb{P}(|\widehat{X}_n^+|_{\tau_n}^* > 2r) \leq \mathbb{P}(\text{there exist } 0 \leq \sigma_n \leq \theta_n \leq \tau_n \text{ such that } W_n(\theta_n) - W_n(\sigma_n) \geq r - c_1), \tag{A.17}$$

where

$$c_1 = u \sup_n |\widehat{\lambda}_n - \widehat{\mu}_n|. \tag{A.18}$$

Note that c_1 is finite by (5) and (8). However,

$$|W_n|_u^* \leq |\widehat{A}_n|_u^* + |V_n|_u^*,$$

and since $\widehat{A}_n$ converge and (A.10) holds, it follows that $|W_n|^*_u$ are tight random variables. This shows that the r.h.s. of (A.17) tends to zero as $r \to \infty$, whence follows (A.15).

Now we will use (A.15) to show (A.13). First, (A.15), the convergence of Z^*_n to zero, and (15) imply

$$|F_n|^*_{\tau_n} \to 0, \quad \text{in probability.} \tag{A.19}$$

By (A.6),

$$\mathbb{P}(\tau_n < u) \leq \tag{A.20}$$
$$\mathbb{P}\big(|[\widehat{X}_n(\cdot) - \widehat{X}_n(0)]^+|^*_{\tau_n} \geq n^{\varrho}/4\big) + \mathbb{P}(|\widetilde{W}_n|^*_{\tau_n} \geq n^{\varrho}/4) + \mathbb{P}\Big(\gamma \int_0^{\tau_n} \widehat{X}_n(s)^+ ds \geq n^{\varrho}/4\Big).$$

The tightness of $\widehat{X}_n(0)$ (23) and of $|\widehat{X}^+_n|^*_{\tau_n}$ implies that the first and last terms on the r.h.s. converge to zero. We have already argued that $|W_n|^*_{\tau_n}$ are tight, thus (A.5) and (A.19) imply that the second term converges to zero as well. This proves (A.13).

Finally, in view of (A.13), (A.14) implies (A.9), (A.15) implies (A.11), and (A.19) implies (A.12). □

The first part of the following result is key to the proof of Theorem 2.1. The second part will only be used in the next subsection, in proving Theorem 2.2.

Lemma A.3. *i. The sequence $(\widehat{X}_n(0), \widehat{A}_n, V_n)$ converges in distribution to (ξ_0, w_A, w_S), where w_S is a Brownian motion that starts from zero and has zero drift and diffusion coefficient $\mu^{1/2}$. In addition, the random variable ξ_0, the process w_A and the process w_S and are mutually independent.*

ii. Let $t \in [0, u)$ and let $t^{(n)} \in [t, u]$, $n \in \mathbb{N}$, be a sequence satisfying $t^{(n)} \downarrow t$. For some positive integer r let G_n and G be random variables with values in $(\mathbb{R}^*_+)^r$, where, for each n, G_n is measurable on $\mathcal{F}_n(t^{(n)})$. Assume, in addition, that $(G_n, V_n, \widehat{A}_n)$ converge in distribution to (G, w_S, w_A). Denote

$$\breve{w}_S(s) = w_S(t+s) - w_S(t), \quad \breve{w}_A(s) = w_A(t+s) - w_A(t), \quad s \in [0, u-t].$$

Then G, $\breve{w}_S$, and $\breve{w}_A$ are mutually independent.

Proof. i. Let us begin by commenting that the convergence of V_n to w_S can be shown in a straightforward manner using the martingale central limit theorem [6]. This, however, does not conveniently lead to an argument that w_A and w_S are mutually independent, for which one has to exploit the independence of the primitives A_n and $\{S_k\}$. We prove the convergence and the independence of w_A and w_S via an argument that a suitable error term (E_n defined below) converges to zero in finite dimensional distributions.

Toward this argument we will use a theorem of Aldous [1], stating that a sequence M_n of martingales converges in distribution to M uniformly over $[0, u]$, provided that M is a continuous martingale, $M_n(t)$ are uniformly integrable in n for each t, and the finite-dimensional distributions of M_n converge to those of M.

Uniform integrability of $V_n(t)$ follows from (A.10). Thus to apply the above result it suffices that we prove convergence in finite-dimensional distributions of V_n to w_S. To this end we write

$$V_n(t) = \widehat{S}_n(t) - E_n(t),$$

where

$$\widehat{S}_n(t) = n^{-1/2} \sum_k (S_k(\mu_{kn}t) - \mu_{kn}t),$$

$$E_n(t) = n^{-1/2} \sum_k [S_k(\mu_{kn}t) - S_k(T_{kn}(t)) - \mu_{kn}t + T_{kn}(t)].$$

Note that, for each n, $\sum_k S_k(\mu_{kn}t)$ is itself a Poisson process of rate μ_n. Also, as follows from the assumption regarding independence of primitive processes and initial conditions, noting that $X_n(0) = Q_n(0) + \sum_k B_{kn}(0)$, the Poisson process alluded to above, the process A_n, and the random variable $X_n(0)$ are mutually independent. Thus $(\widehat{X}_n(0), \widehat{A}_n, \widehat{S}_n)$ converge to (ξ_0, w_A, w_S), and ξ_0, w_A and w_S are mutually independent. It remains to prove that E_n converge to zero in the sense of finite-dimensional distributions. To this end, it suffices to show that for each $v \in [0, u]$, $E_n(v) \to 0$ in probability. Fix v. Recall that, for each n and k, $R_{kn}(t) := S_k(T_{kn}(t)) - T_{kn}(t)$ is an $\mathbb{F}_n$-martingale. Let

$$\tau_{kn} = \inf\{t : T_{kn}(t) = \mu_{kn}v\}.$$

Since $B_{kn} \le 1$, it is easy to verify that $\tau_{kn} \ge v$. Also, it follows from standard ergodic considerations that $\tau_{kn} < \infty$ a.s., for each k, n. To see that, fix n. Let $L > 0$ be such that $\mathbb{P}(A_n(L) > N_n) > 0$. Then with probability 1, there are infinitely many disjoint intervals of the form $[\sigma, \sigma + L + 1] \subset [v, \infty)$ with the properties (a) there are at least $N_n + 1$ arrivals within $[\sigma, \sigma + L]$, (b) there are no abandonments and no service completions within $[\sigma, \sigma + L + 1]$. Clearly, within each of these intervals there are more customers than servers in the system for a unit of time, and since the policy is work conserving, all servers are necessarily occupied for a unit of time. Since there are infinitely many such intervals, this shows $\tau_{kn} < \infty$ for all k, a.s.

Set

$$E'_n(t) = n^{-1/2} \sum_k (R_{kn}(\tau_{kn} \wedge t) - R_{kn}(v)),$$

and note that $E_n(v) = \lim_{t\to\infty} E'_n(t)$. Also, $E'_n(t)$ is an $\mathbb{F}_n$-martingale for $t \in [v, \infty)$, with $E'_n(v) = 0$. To use the Burkholder-Davis-Gundy inequality the way we did before, note that the number of jumps of E'_n over $[v, \infty)$ is given by

$$\sum_k (S_k(\mu_{kn}v) - S_{kn}(T_{kn}(v))) = \sum_k S_k(T_{kn}(\tau_{kn}) - S_k(T_{kn}(v))).$$

Since for each n and k, τ_{kn} is a stopping time on $\mathbb{F}_n$, the expected number of jumps is given by

$$\mathbb{E}[\sum_k (T_{kn}(\tau_{kn}) - T_{kn}(v))] = \mathbb{E}[\sum_k (\mu_{kn}v - T_{kn}(v))].$$

Hence

$$\begin{aligned}
\mathbb{E}[E_n(v)^2] &\le \mathbb{E}[\sup_{t\in[v,\infty)} E'_n(t)^2] \\
&\le \frac{c}{n}\mathbb{E}[\sum_k (\mu_{kn}v - T_{kn}(v))] \\
&= \frac{c}{n}\mathbb{E}[\sum_k \mu_{kn} \int_0^v I_{kn}(t)dt] = \frac{c}{\sqrt{n}}\mathbb{E}[L_n(v)].
\end{aligned}$$

It follows from (22), (A.8) and (A.13) that $n^{-1/2}L_n(v)$ converges in probability to zero. In addition, using (22), it is bounded by the constant $v\sup_n \mu_n/n < \infty$. Hence the last expression in the above display converges to zero. This shows that $E_n(v) \to 0$ in probability. Since $v \in [0,u]$ is arbitrary, the result follows.

ii. Using Assumption 2.1(iv), for each n, $\check{V}_n$, $\check{A}_n$ and G_n are mutually independent, where

$$\check{A}_n(s) = n^{-1/2}[A_n(AT_n(t^{(n)}) + s) - A_n(AT_n(t^{(n)})) - \lambda_n s], \qquad s \geq 0,$$

$$\check{V}_n(s) = V_n(t^{(n)} + s) - V_n(t^{(n)}), \qquad s \geq 0.$$

Note that $\check{A}_n(s) = \widehat{A}_n(AT_n(t) + s) - \widehat{A}_n(AT_n(t))$ for $s \geq 0$. It is easy to see that $AT_n(t^{(n)}) \to t$ in probability. Since the limit in distribution of $(V_n, \widehat{A}_n)$ is a process with continuous sample paths, it follows that $(G_n, \check{V}_n, \check{A}_n)$ converges in distribution, and has the same limit as that of $(G_n, V_n(t+\cdot) - V_n(t), \widehat{A}_n(t+\cdot) - \widehat{A}_n(t))$, namely $(G, \check{w}_S, \check{w}_A)$. It follows that the three components are mutually independent. □

Proof of Theorem 2.1. Recall that $u \in (0,\infty)$ is fixed, but arbitrary. Hence it suffices to prove the convergence in distribution in the uniform topology over $[0,u]$. By Lemma A.3(i), $(\widehat{X}_n(0), \widehat{A}_n, V_n)$ converges in distribution to (ξ_0, w_A, w_S), where the latter three elements are mutually independent. Using equation (A.5), the convergence of the term F_n to zero (A.12) and of the term $\widehat{\lambda}_n - \widehat{\mu}_n$ to $\widehat{\lambda} - \widehat{\mu} = \beta$ (5), (8), it follows that $(\widehat{X}_n(0), \widetilde{W}_n)$ converges in distribution to $(\xi_0, \widetilde{w})$ where $\widetilde{w}$ is a Brownian motion with drift β and diffusion coefficient σ, independent of ξ_0. Using the tightness of the random variables $|\widehat{X}_n^+|_u^*$ in equation (A.6), and the positivity of L_n, shows that $|L_n|_u^*$, and in turn, $|\widehat{X}_n|_u^*$, are tight random variables. We argue that

$$|\widehat{X}_n^-|_u^* \to 0 \quad \text{in probability.} \tag{A.21}$$

Given $\varepsilon > 0$, consider the event

$$\Omega_n^\varepsilon = \{\widehat{X}_n(0) \geq -\varepsilon,\ |\widehat{X}_n^-|_u^* > 3\varepsilon\}.$$

On this event there exist $0 \leq \sigma_n < \theta_n \leq u$ such that $\widehat{X}_n(\sigma_n) \geq -2\varepsilon$, $\widehat{X}_n(\theta_n) \leq -3\varepsilon$, and $\widehat{X}_n(t) \leq -\varepsilon$ for $t \in [\sigma_n, \theta_n]$ (this uses the fact that the jumps of $\widehat{X}_n$ are a.s. bounded by $cn^{-1/2}$). Using (11), we have $I_n(t) \geq \varepsilon n^{1/2}$ for $t \in [\sigma_n, \theta_n]$. Hence by (9) and (22),

$$L_n(\theta_n) - L_n(\sigma_n) \geq \mu_n^{\min}\varepsilon(\theta_n - \sigma_n).$$

Next, let us use equation (A.6). Note that the last term on the r.h.s. of this equation does not vary between the times σ_n and θ_n because $\widehat{X}_n$ is negative. Also, the process L_n is nondecreasing, and so

$$-\varepsilon \geq \widehat{X}_n(\theta_n) - \widehat{X}_n(\sigma_n) \geq \widetilde{W}_n(\theta_n) - \widetilde{W}_n(\sigma_n).$$

Denote

$$\bar{w}_u(x,\delta) = \sup_{s,t\in[0,u];\ |s-t|\leq\delta} |x(t) - x(s)|,$$

for $x : [0,u] \to \mathbb{R}$, $\delta > 0$. We obtain, for each ε and n,

$$\mathbb{P}(\Omega_n^\varepsilon) \leq \mathbb{P}(\text{there exists } \delta > 0 \text{ such that } \bar{w}_u(\widetilde{W}_n, \delta) \geq \varepsilon,\ \varepsilon\delta\mu_n^{\min} \leq |L_n|_u^*).$$

By tightness of $|L_n|_u^*$, there is a function g such that $\lim_{r\to\infty} g(r) = 0$, and, for every $r > 0$,

$$\mathbb{P}(\Omega_n^\varepsilon) \le g(r) + \mathbb{P}\Big(\bar{w}_u\Big(\widetilde{W}_n, \frac{r}{\varepsilon\mu_n^{\min}}\Big) \ge \varepsilon\Big).$$

A sequence of processes defined on $[0, u]$, with sample paths in the Skorohod space, is said to be *C-tight* if it is tight, and every subsequential limit has continuous sample paths with probability one. C-tightness of, say $\{U_n\}$, implies the convergence in probability of $\bar{w}_u(U_n, \delta) \to 0$, for every δ. As processes that converge to a Brownian motion, $\widetilde{W}_n$ are C-tight. Since $\mu_n^{\min} \to \infty$ (9) it therefore follows that the last term on the r.h.s. of the above display converges to zero as $n \to \infty$. Sending $r \to \infty$ shows that $\lim_n \mathbb{P}(\Omega_n^\varepsilon) = 0$. Finally, since the weak limit ξ_0 of $\widehat{X}_n(0)$ is nonnegative (23), it follows that $\lim_n \mathbb{P}(|\widehat{X}_n^-|_u^* > 3\varepsilon) = 0$. This shows (A.21).

For an RCLL path y from $[0, \infty)$ to $\mathbb{R}$, a pair $(\bar{\xi}, \bar{l})$ is regarded a solution to the Skorohod equation

$$\bar{\xi}(t) = y(t) - \gamma \int_0^t \bar{\xi}(s)ds + \bar{l}(t), \tag{A.22}$$

provided that the equation above holds for all $t \ge 0$, that $\bar{\xi}(t) \ge 0$ for all t, $\bar{l}$ is nondecreasing, and $\int_{[0,\infty)} \mathbf{1}_{\{\bar{\xi}(t)>0\}} dl(t) = 0$. Note that the process ξ, that is a solution to the Skorohod equation (24) is, with probability one, a solution to the Skorohod equation (A.22) with data $y(t) = \xi_0 + \beta t + \sigma w(t)$. It is well known that existence and uniqueness hold for the equation (A.22) for y in the space of RCLL paths, and that the solution map is continuous in the topology of uniform convergence on compacts (for example, this statement follows from the results of [3]). Now, by (A.6),

$$\widehat{X}_n(t)^+ = Y_n(y) - \gamma \int_0^t \widehat{X}_n(s)^+ ds + L_n(t), \tag{A.23}$$

where

$$Y_n(t) = \widehat{X}_n(0) + \widetilde{W}_n(t) + \widehat{X}_n(t)^-.$$

It follows from (21) and (22) that $\frac{d}{dt}L_n(t) > 0$ if and only if $\widehat{X}_n(t) < 0$. Hence $(\widehat{X}_n^+, L_n)$ form a solution to the Skorohod equation for Y_n. Recall that $(\widehat{X}_n(0), \widetilde{W}_n) \to (\xi_0, \widetilde{w})$ in distribution. Thus by (A.21), $Y_n \to \xi_0 + \widetilde{w}$ in distribution. As a result of the continuity of the solution map, we obtain convergence in distribution, uniformly on compacts, of $\widehat{X}_n^+$, and in turn, of $\widehat{X}_n$, to the solution to (A.22) with $y = \xi_0 + \widetilde{w}$. If w is a standard Brownian motion independent of ξ_0, then $(\xi_0, \widetilde{w})$ is equal in law to $(\xi_0, \{\beta t + \sigma w(t), t \ge 0\})$. Thus, with (ξ, l) denoting the unique solution to (24) with data (ξ_0, w), we have shown that $\widehat{X}_n \Rightarrow \xi$. Recalling that $\hat{Q}_n$ and $\hat{I}_n$ are the positive and, resp., negative parts of $\widehat{X}_n$ (21), and noting that L_n is, by (A.23), given explicitly in terms of $\widehat{X}_n$, shows $(\widehat{X}_n, L_n, \widehat{Q}_n, \widehat{I}_n) \Rightarrow (\xi, l, \xi, 0)$. □

A.2 Convergence of delay and service time

This subsection contains the proof of Theorem 2.2. Lemma A.4(iv) below shows that $\mu\widehat{\Delta}_n$ is close to $\widehat{Q}_n$, that is the positive part of $\widehat{X}_n$. This translates the problem into that of analyzing the joint limit law of $\widehat{X}_n$ and $\widehat{\Sigma}_n$, where $\widehat{X}_n \Rightarrow \xi$ is already established by Theorem 2.1. The main ingredient of the proof is an argument that the η_i's, which represent the limit law of the scaled service times, are independent across i and independent of ξ. Intuitively, given some $t_1 < t_2$, the value that the delay experienced by the customer $C_n(t_1)$ takes should have nearly no effect on the delay and service time experienced by $C_n(t_2)$ because the system contains many customers. However this intuition does not appear to lead to an efficient way of proving independence. We approach this problem via an argument by induction on the number of customers. Among other considerations, the argument crucially uses the independence result of Lemma A.3(ii) and the uniqueness of solutions to the Skorohod equation.

Throughout, let $u \in (0, \infty)$, $j \in \mathbb{N}$ and $0 < t_1 < \cdots < t_j < u$ be fixed. To simplify the notation we write $\widehat{X}_n^i$ for $\widehat{X}_n(t_i)$, $1 \le i \le j$, and use similar convention for $\widehat{Q}_n^i$, $\widehat{\Delta}_n^i$ and $\widehat{\Sigma}_n^i$, $1 \le i \le j$. $\widehat{X}_n(0)$ will be denoted by $\widehat{X}_n^0$.

Recall that $AT_n(t)$ and $RT_n(t)$ denote the arrival and, resp., routing time of the customer $C_n(t)$. Given t, let $\Omega_n(t)$ denote the event that the customer $C_n(t)$ does not abandon, namely $\Omega_n(t) = \{AB_n(t) = \infty\} = \{\Delta_n(t) < \infty\}$. Since the process Q_n is right-continuous, the quantity $Q_n(AT_n(t))$ represents the number of customers in the queue at the time of arrival of the customer $C_n(t)$, including this customer. The first item of the lemma below shows that abandonment of an individual customer occurs according to an exponential clock. The combination of items (ii)–(iv) shows that abandonment occurs with small probability, and that $\mu\widehat{\Delta}_n$ is close to $\widehat{Q}_n$. In what follows we use the notation $M[s,t] := M(t) - M(s)$ for any process M.

Lemma A.4. *Let $t \ge 0$ be fixed. Then one has the following.*
i. Write $\overline{AB}_n = AB_n(t) - AT_n(t)$ and $\Delta_n = \Delta_n(t) \equiv RT_n(t) - AT_n(t)$. Then

$$\mathbb{P}(\overline{AB}_n > s) \equiv \mathbb{P}(\overline{AB}_n > s \wedge \Delta_n) = \mathbb{E}[e^{-\gamma_n(s\wedge\Delta_n)}], \qquad s \ge 0.$$

ii. On $\Omega_n(t)$,

$$|\widehat{Q}_n(AT_n(t)) - \bar{\mu}_n\widehat{\Delta}_n(t)| \le J_n(t) := |V_n[AT_n, RT_n]| + L_n[AT_n, RT_n] + n^{-1/2}Z_n[AT_n, RT_n].$$

iii. $\mathbb{P}(\Omega_n(t)) \to 1$.
iv. $|\widehat{Q}_n(AT_n(t)) - \bar{\mu}_n\widehat{\Delta}_n(t)| \to 0$ in probability.

Proof. i. Using the representation

$$\mathbb{P}(\overline{AB}_n > s) = \mathbb{P}(\bar{N}_s = 0),$$

where

$$\bar{N}_s = \int_{[t,t+s]} 1_{\{U_{Z_n(v)}Q_n(v-)\in[POS_n(t,v-),POS_n(t,v-)+1)\}}dZ_n(v),$$

the claim follows from a standard result on thinning of a Poisson point process (cf. [9, Proposition 1.13]).

ii. On the event $\Omega_n(t)$, $RT_n(t)$ represents the time when $C_n(t)$ starts being served. Equivalently, this is the time when all the $Q_n(AT_n(t))$ customers that are in the queue at time $AT_n(t)$, have left the queue, either by starting service or by abandoning. As a result, we have on $\Omega_n(t)$,

$$D_n[AT_n, RT_n] \le Q_n(AT_n) \le D_n[AT_n, RT_n] + Z_n[AT_n, RT_n], \qquad \text{(A.24)}$$

where we omit the dependence of AT_n and RT_n on t from the notation. Using (A.2) and then (A.7),

$$
\begin{aligned}
n^{-1/2}D_n[AT_n, RT_n] &= V_n[AT_n, RT_n] + n^{-1/2}T_n[AT_n, RT_n] \\
&= V_n[AT_n, RT_n] - L_n[AT_n, RT_n] + \frac{\mu_n}{n}\widehat{\Delta}_n(t),
\end{aligned}
$$

holds on the event $\Omega_n(t)$. This shows (ii).

iii. Let

$$
\theta_n = \theta_n(t) = \inf\{s \geq AT_n : D_n[AT_n, s] \geq Q_n(AT_n)\}.
$$

Let us first argue that $\theta_n - AT_n \to 0$ in probability. Roughly speaking, this should hold since the queue-length is of order $n^{1/2}$ while the rate at which the departure process increases is of order n. Toward a rigorous argument, note that $Q(AT_n) = D[AT_n, \theta_n]$. Thus given $\varepsilon > 0$,

$$
\begin{aligned}
\mathbb{P}(\theta_n - AT_n > \varepsilon) &\leq \mathbb{P}(\theta_n > t + 2\varepsilon,\ AT_n < t + \varepsilon) + \mathbb{P}(AT_n \geq t + \varepsilon) \\
&\leq \mathbb{P}\Big(\frac{D_n[t+\varepsilon, t+2\varepsilon]}{n} \leq \frac{|Q_n|^*_u}{n}\Big) + \mathbb{P}(AT_n \geq t + \varepsilon).
\end{aligned}
$$

Recall from Lemma A.2 the notation $\bar{T}(s) = \mu s$, $s \geq 0$. By (A.2), (A.9), (A.10), we have that $n^{-1}D_n \to \bar{T}$ in probability, uniformly in $[0, u]$. Hence $D_n[t+\varepsilon, t+2\varepsilon]/n \to \varepsilon\mu$ in probability. By Theorem 2.1, $|\widehat{Q}_n|^*_u$ are tight r.v.s, hence $|Q_n|^*_u/n \to 0$ in probability, and the first term on the r.h.s. of the above display converges to zero. So does the second term, by the fact that $AT_n \to t$ (as follows from the speeding up of the inter-arrival times). Since ε is arbitrary, this shows that $\theta_n - AT_n \to 0$ in probability, as claimed.

Note that on the event $\Omega_n(t)^c = \{AB_n(t) < \infty\}$, one necessarily has $\overline{AB}_n \leq \theta_n - AT_n$. Hence, given $\varepsilon > 0$, there exist $\alpha_n \downarrow 0$, such that

$$
\begin{aligned}
\mathbb{P}(\Omega_n(t)^c) &\leq \mathbb{P}(\overline{AB}_n \leq \theta_n - AT_n) \leq \mathbb{P}(\overline{AB}_n \leq \varepsilon) + \alpha_n \\
&= 1 - \mathbb{P}(\overline{AB}_n = \infty) - \mathbb{P}(\overline{AB}_n \in (\varepsilon, \infty)) + \alpha_n \\
&= 1 - \mathbb{P}(\overline{AB}_n = \infty) - \mathbb{P}(\overline{AB}_n > \varepsilon \wedge \Delta_n, \Delta_n = \infty) + \alpha_n,
\end{aligned}
$$

where we used the fact that $\Delta_n = \infty$ if and only if $\overline{AB}_n < \infty$. Thus, using item (i),

$$
\begin{aligned}
\mathbb{P}(\Omega_n(t)^c) &\leq 1 - \mathbb{P}(\overline{AB}_n > \varepsilon \wedge \Delta_n) + \alpha_n \\
&= 1 - \mathbb{E}[e^{-\gamma_n(\varepsilon \wedge \Delta_n)}] + \alpha_n \\
&\leq 1 - e^{-\gamma_n \varepsilon} + \alpha_n \leq 1 - e^{\bar{\gamma}\varepsilon} + \alpha_n,
\end{aligned}
$$

where $\bar{\gamma} = \sup_n \gamma_n < \infty$. Taking $n \to \infty$ and then $\varepsilon \to 0$ gives zero on the r.h.s. This proves item (iii).

iv. In view of items (ii) and (iii), it suffices to prove that $J_n(t) \to 0$ in probability. An argument as in item (iii) above, using the fact $D_n[AT_n, RT_n] \leq Q_n(AT_n)$ on $\Omega_n(t)$ (A.24), shows that $RT_n - AT_n \to 0$ in probability. Hence by item (ii) of the lemma, the convergence of $J_n(t)$ to zero will follow provided C-tightness of the processes V_n, L_n and $n^{-1/2}Z_n$. And indeed, V_n are C-tight as processes that converge to a Brownian motion (Lemma A.3(i)), and so are L_n as processes that converge to the boundary term l of equation (24) (Theorem 2.1). Finally, by (A.11), (A.12) and (A.16), the processes $n^{-1/2}Z_n$ are C-tight. This shows $J_n(t) \to 0$ in probability and establishes item (iv). □

The theorem will be proved by first showing

$$(\widehat{X}_n^1, \widehat{\Sigma}_n^1, \widehat{X}_n^2, \widehat{\Sigma}_n^2 \ldots, \widehat{X}_n^j, \widehat{\Sigma}_n^j) \Rightarrow (\xi(t_1), \eta_1, \xi(t_2), \eta_2, \ldots, \xi(t_j), \eta_j). \tag{A.25}$$

Once the above is established, the result will follow by an easy implementation of Lemma A.4.

Toward proving (A.25), consider the statements

$$\widehat{X}_n^1 \Rightarrow \xi(t_1), \tag{A.26}$$

$$(\widehat{X}_n^1, \widehat{\Sigma}_n^1, \widehat{X}_n^2, \widehat{\Sigma}_n^2, \ldots, \widehat{X}_n^i) \Rightarrow (\xi(t_1), \eta_1, \xi(t_2), \eta_2, \ldots, \xi(t_i)), \tag{A.27}$$

$$(\widehat{X}_n^1, \widehat{\Sigma}_n^1, \widehat{X}_n^2, \widehat{\Sigma}_n^2, \ldots, \widehat{X}_n^i, \widehat{\Sigma}_n^i) \Rightarrow (\xi(t_1), \eta_1, \xi(t_2), \eta_2, \ldots, \xi(t_i), \eta_i), \tag{A.28}$$

where ξ and η are as in the statement of the theorem.

Claim A. (A.26) *holds.*

Claim B. *Given* $i \in \{1, 2, \ldots, j\}$*, if* (A.27) *holds then* (A.28) *holds.*

Claim C. *Given* $i \in \{1, 2, \ldots, j-1\}$*, if* (A.28) *holds then* (A.27) *holds for* $i+1$.

Clearly, by induction on i, Claims A–C imply (A.25). In what follows we shall prove these claims.

First, note that Claim A follows directly from Theorem 2.1.

To prove Claim C we will need the following.

Lemma A.5. *Let ξ be a solution to the Skorohod equation (24) with data (ξ_0, w). Given $0 \le t < u$ there exists a measurable map $H : \mathbb{R} \times C([t,u] : \mathbb{R}) \to C([t,u] : \mathbb{R})$ such that, with probability one, $\xi|_{[t,u]} = H(\xi(t), (w|_{[t,u]} - w(t)))$.*

Proof. It is well-known that the result follows from pathwise uniqueness and weak existence of solutions (cf. [7, Cor. 5.3.23]). By [2], pathwise uniqueness and, in fact, strong existence do hold for equation (24). □

Let us now prove Claim C. Fix $1 \le i \le j-1$. We are given that (A.28) holds and are required to prove that

$$(\widehat{X}_n^1, \widehat{\Sigma}_n^1, \widehat{X}_n^2, \widehat{\Sigma}_n^2, \ldots, \widehat{X}_n^{i+1}) \Rightarrow (\xi(t_1), \eta_1, \xi(t_2), \eta_2, \ldots, \xi(t_{i+1})). \tag{A.29}$$

Moreover, by Lemma A.3(i) and Theorem 2.1, we have $(X_n, V_n, \widehat{A}_n) \Rightarrow (\xi, w_S, w_A)$. We write w for the standard Brownian motion related to w_S and w_A by $w_S - w_A = \sigma w$ (note that $\beta s + \sigma w(s)$ is what we denoted by $\widetilde{w}(s)$ in the previous section). As a result, every subsequence of $(\widehat{X}_n, V_n, \widehat{A}_n, (\widehat{\Sigma}_n^l)_{l=1}^i)$ has a further subsequence that is convergent, and if $(\xi^*, w_S^*, w_A^*, (\eta_l^*)_{l=1}^i)$ is a subsequential limit then the following relations must hold:

- (ξ^*, w_S^*, w_A^*) has the same law as (ξ, w_S, w_A); in particular, ξ^* is a solution to the Skorohod equation with data $(\xi^*(0), w^*)$, $(\xi^*(0), w^*)$ is equal in law to (ξ_0, w), and $w_A^* - w_S^* = \sigma w^*$;
- $((\xi^*(t_l))_{l=1}^i, (\eta_l^*)_{l=1}^i)$ has the same law as $((\xi(t_l))_{l=1}^i, (\eta_l)_{l=1}^i)$.

Clearly, Claim C will follow once we show that $((\xi^*(t_l))_{l=1}^{i+1}, (\eta_l^*)_{l=1}^i)$ and $((\xi(t_l))_{l=1}^{i+1}, (\eta_l)_{l=1}^i)$ are equal in law, which, in view of the second bullet, is equivalent to the statement

$$(\xi^*(t_l))_{l=1}^{i+1} \text{ and } (\eta_l^*)_{l=1}^i \text{ are mutually independent.} \tag{A.30}$$

And indeed by Lemma A.5 and the first bullet, there exists a map H' such that, with probability one, $\xi^*(t_{i+1}) = H'(\xi^*(t_i), \check{w}^*)$ and $\xi(t_{i+1}) = H'(\xi(t_i), \check{w})$, where $\check{w}^* = w^*|_{[t_i,u]} - w^*_{t_i}$ and a similar definition is used for $\check{w}$. Thus by the second bullet, (A.30) will follow provided we show that $\check{w}^*$ and $((\xi^*(t_l))_{l=1}^i, (\eta_l^*)_{l=1}^i)$ are mutually independent. This is established in the following lemma, which completes the argument.

Lemma A.6. *Let (G, w_S, w_A), where $G = (\xi(t_l), \eta_l)_{l=1}^i$, be a limit in distribution of a convergent subsequence of*

$$((\widehat{X}_n^l, \widehat{\Sigma}_n^l)_{l=1}^i, V_n, \widehat{A}_n).$$

Then G, $\check{w}_S := w_S|_{[t_i,u]} - w_S(t_i)$ and $\check{w}_A := w_A|_{[t_i,u]} - w_A(t_i)$ are mutually independent.

Proof. The result will follow from Lemma A.3(ii) once its hypotheses are verified. To this end, for each n, let $t^{(n)} = t_i + n^\gamma$, where $\gamma \in (-1/2, 0)$ is a constant. All statements regarding convergence will be understood as convergence along the given subsequence.

Recall that $EX_n(t)$ denotes the time when customer $C_n(t)$ leaves the system (either by completing service or by abandoning), namely

$$EX_n(t) = DEP_n(t) \wedge AB_n(t). \tag{A.31}$$

Define, for $1 \le l \le i$,

$$E_n^l = \{EX_n(t_l) < t^{(n)}\}, \qquad \widetilde{\Sigma}_n^l = \widehat{\Sigma}_n^l \mathbf{1}_{E_n^l},$$

where we interpret infinity times zero as zero. It will be convenient to work with the random variables $\widetilde{\Sigma}_n^l$ because they are (as we will show) measurable on $\mathcal{F}_n(t^{(n)})$, and at the same time close to the original quantities $\widehat{\Sigma}_n^l$.

The assumptions of the lemma imply

$$G_n := (\widehat{X}_n^1, \widetilde{\Sigma}_n^1, \widehat{X}_n^2, \widetilde{\Sigma}_n^2 \dots, \widehat{X}_n^i, \widetilde{\Sigma}_n^i) \Rightarrow G = (\xi(t_1), \eta_1, \xi(t_2), \eta_2, \dots, \xi(t_i), \eta_i), \tag{A.32}$$

provided we show, for $1 \le l \le i$,

$$\mathbb{P}(E_n^l) \to 1. \tag{A.33}$$

Let us then argue that (A.33) holds. Fix l. By (A.31), it suffices to show that (i) $\mathbb{P}(AT_n(t_l) - t_l < n^\gamma/3) \to 1$, (ii) $\mathbb{P}(\Delta_n(t_l) < n^\gamma/3) \to 1$ and (iii) $\mathbb{P}(\Sigma_n(t_l) < n^\gamma/3) \to 1$. Claim (i) is an easy consequence of the fact that the acceleration parameter λ_n of the arrival process A_n is of the order of n (5). Claim (iii) holds due to the validity of (A.28), by which $\mathbb{P}(n^{1/2}\Sigma_n(t_l) > a_n) \to 0$, for any sequence $a_n \uparrow \infty$. To see that claim (ii) holds, note that $\widehat{Q}_n \le |\widehat{X}_n|$ (by (11)), hence using the fact $AT_n(t_l) \to t_l$ and Theorem 2.1, we have that $\widehat{Q}_n(AT_n)$ form a tight sequence of random variables. By Lemma A.4 and the fact $\bar{\mu}_n \to \mu$ (7), it follows that $\widehat{\Delta}^l_n$ are tight. As a result, $n^{-\gamma}\Delta_n(t_l) \to 0$ in probability, which gives (ii). We have thus shown that items (i)–(iii) hold. This implies (A.33) and, in turn, (A.32).

Next we show that all components of G_n are measurable on $\mathcal{F}_n(t^{(n)})$, which, for this discussion we abbreviate by $\mathcal{F}$. We also fix l and drop the l and t_l from the notation of $\widehat{X}^l_n$, $\widehat{\Sigma}^l_n$, E^l_n, $\Delta_n(t_l)$, etc. By Assumption 2.1(i), $\widehat{X}_n$ is $\mathcal{F}$-measurable. To show that so is $\widetilde{\Sigma}_n$, note that for arbitrary $a \in [0, \infty)$,

$$\{\widetilde{\Sigma}_n < a\} = (E_n \cap \{\widehat{\Sigma}_n < a\}) \cup E^c_n,$$

where one uses the fact that, on the specified event, $EX_n = DEP_n$ (because $\Delta_n = \infty$ in case the minimizer on the r.h.s. of (A.31) is the second term). First, by Assumption 2.2, $\{DEP_n \le \delta\} \cup \{AB_n \le \delta\}$ are both in $\mathcal{F}$ provided $\delta < t^{(n)}$, hence so is the union of these events over $\delta = \delta_k \uparrow t^{(n)}$. This shows $E_n \in \mathcal{F}$. Next, denoting $a_n = an^{-1/2}$, note that

$$\begin{aligned} E_n \cap \{\widehat{\Sigma}_n < a\} &= \{RT_n + \Sigma_n < t^{(n)}, \, \Sigma_n < a_n\} \\ &= \bigcup_{b \in \mathbb{Q}:\, b < t^{(n)} - a_n} \{RT_n > b\} \cap \{RT_n + \Sigma_n < b + a_n\}. \end{aligned}$$

By Assumption 2.2, each of the sets over which the union is taken, is $\mathcal{F}$-measurable. Consequently, so is the event on the l.h.s. of the above display. This shows $\widetilde{\Sigma}_n$ is $\mathcal{F}$-measurable.

We can now apply Lemma A.3(ii). We have $(G_n, V_n, \widehat{A}_n) \Rightarrow (G, w_S, w_A)$, and, for each n, G_n is $\mathcal{F}_n(t^{(n)})$-measurable. Consequently, G, $\check{w}_S$ and $\check{w}_A$ are mutually independent. □

As argued above, this completes the proof of Claim C.

Finally we turn to the proof of Claim B. Recall that t_i is fixed. The main idea of the proof is that if $C_n(t_i)$ ever gets to be the first in line, since service time and abandonment are exponential, the probability that it is routed to server k is proportional to μ_k. Another important part of the proof is to show that, with probability tending to one, this customer indeed reaches the head of the line (recall that in Subsection 2.2 we have imposed the first-come-first-served discipline).

Recall that $POS_n(s,t)$ represents the position $\in \{1, 2, 3, \ldots\}$ at time t of $C_n(s)$ if this customer is in the queue at time t, and it takes the value ∞ otherwise. Define

$$\kappa_n = \inf\{t : POS_n(t_i, t) = 1\},$$

$$\bar{\kappa}_n = \kappa_n \wedge EX_n(t_i).$$

Lemma A.7. *i. $\bar{\kappa}_n$ is a finite $\mathbb{F}_n$-stopping time.*
ii. One has $\{\kappa_n < \infty\} \in \mathcal{F}_n(\bar{\kappa}_n)$.
iii. $\mathbb{P}(\kappa_n < \infty) \to 1$.

Proof. i. Since $EX_n(t_i) < \infty$ a.s., $\bar{\kappa}_n$ is finite a.s. By assumption, $POS_n(t_i, t) \in \mathcal{F}_n(t)$, and therefore $\{\kappa_n < t\} \in \mathcal{F}_n(t)$ for every t. Hence κ_n is a stopping time, and since we have already shown that EX_n is a stopping time, so is $\bar{\kappa}_n$.

ii. To prove the statement we need to show that, for any u, $E_u = \{\kappa_n < \infty\} \cap \{\bar\kappa_n < u\} \in \mathcal{F}_n(u)$. Now,

$$E_u = \{\kappa_n < u\} \cup \{\kappa_n < \infty, EX_n < u\} = \{\kappa_n < u\},$$

since clearly $\kappa_n < \infty$ implies $\kappa_n \le EX_n$. Since we have already shown that κ_n is a stopping time, it follows that $E_u \in \mathcal{F}_n(u)$.

iii. Given $t > 0$, denote by $\Omega'_n(t)$ the event $\{Q_n(AT_n(t)) > 0\}$, namely that, at the time when the customer $C_n(t)$ arrives, the queue is non-empty. Let us show that the probability of this event, for fixed t, tends to one. Indeed,

$$\mathbb{P}(\Omega'_n(t)^c) = \mathbb{P}(Q_n(AT_n) = 0) = \mathbb{P}(\widehat{X}_n(AT_n) \le 0),$$

and, since $\widehat{X}_n(AT_n) \Rightarrow \xi(t)$, $\limsup_n \mathbb{P}(\Omega'_n(t)^c) \le \mathbb{P}(\xi(t) \le 0) = \mathbb{P}(\xi(t) = 0)$. It is well-known for a non-degenerate reflected diffusion such as (24) that $\mathbb{P}(\xi(t) = 0) = 0$, for every $t > 0$. As a result,

$$\mathbb{P}(\Omega'_n(t)) \to 1, \qquad t > 0. \tag{A.34}$$

Now, $\mathbb{P}(\kappa_n < \infty) \ge \mathbb{P}(\Omega'_n(t_i) \cap \{AB_n(t_i) = \infty\})$, because any customer that arrives to a non-empty queue and never abandons, will, with probability one, eventually get to position 1. Hence by (A.34) and Lemma A.4, $\mathbb{P}(\kappa_n < \infty) \to 1$. □

As mentioned earlier, the calculation of the chances that customer $C_n(t_i)$ is routed to each of the servers is based on the fact that, starting at κ_n, a competition between exponentials takes place. Note however that it is the conditional distribution of the service time given G_n that we are required to compute, and there is no guarantee that the service times that are elements of G_n correspond to services completed earlier than κ_n. We need a truncated version of service times.

Write $\bar E^l_n = \{EX_n(t_l) < t_i\}$, for $1 \le i \le i-1$. An argument similar to the one

that proves (A.33) gives

$$\mathbb{P}(\bar E^l_n) \to 1, \qquad 1 \le l \le i-1. \tag{A.35}$$

Let

$$\bar G_n = (\widehat{X}^1_n, \widehat{\Sigma}^1_n \mathbf{1}_{\bar E^1_n}, \widehat{X}^2_n, \widehat{\Sigma}^2_n \mathbf{1}_{\bar E^2_n}, \dots, \widehat{X}^i_n), \qquad \bar G = (\xi(t_1), \eta_1, \xi(t_2), \eta_2, \dots, \xi(t_i)).$$

Recall that we are given that (A.27) holds. Because the measure induced by $\bar G$ on $\mathbb{R}^{2i-1}$ does not charge the boundary of any set of the form $M = (-\infty, a_1] \times (-\infty, a_2] \times \cdots \times (-\infty, a_{2i-1}]$, and in view of (A.35), we have, for every such set M, with the notation $M_n = \{\bar G_n \in M\}$,

$$\mathbb{P}(M_n) \to \mathbb{P}(\bar G \in M) =: p(M). \tag{A.36}$$

We are required to prove (A.28). For a similar reason, (A.28) is equivalent to the statement that, for every M as above, and $a \in [0, \infty)$,

$$\mathbb{P}(M_n \cap \{\widehat{\Sigma}^i_n > a\}) \to p(M)\mathbb{P}(\eta_i > a). \tag{A.37}$$

We will prove (A.37) in what follows.

To this end, write $M'_n = M_n \cap \{\kappa_n < \infty\}$, and note that

$$\mathbb{P}(\widehat{\Sigma}^i_n > a | M'_n) = \sum_{k \in K_n} \mathbb{P}(\widehat{\Sigma}^i_n > a | M'_n, RD_n = k)\mathbb{P}(RD_n = k | M'_n).$$

Toward computing $\mathbb{P}(RD_n = k|M'_n)$ for some fixed $k \in K_n$, denote, for $r \in K_n$, $v_{rn} = T_{rn}(\bar{\kappa}_n)$, and set

$$\tau_{rn} = \inf\{s > v_{rn} : \Delta S_r(s) > 0\}, \quad \bar{\tau}_{rn} = \mu_{rn}^{-1}(\tau_{rn} - v_{rn}).$$

On the event $\{\kappa_n < \infty\}$, the routing decision RD_n takes the value k if and only if

$$\bar{\tau}_{kn} < \bar{\tau}_{rn} \text{ for all } r \neq k, \text{ and } AB_n < \infty.$$

Now, since $\bar{\kappa}_n$ is a finite stopping time (Lemma A.7(i)), it follows from Assumption 2.1(iii) that, conditioned on $\mathcal{F}_n(\bar{\kappa}_n)$, the random variables $(\tau_r - v_r)$ are i.i.d. standard exponentials. Using this along with Lemma A.4(i), we have a.s. on $\{\kappa_n < \infty\}$,

$$\mathbb{P}(RD_n = k|\mathcal{F}_n(\bar{\kappa}_n)) = \frac{\mu_{kn}}{\gamma_n + \sum_{r \in K_n} \mu_{rn}} = \frac{\mu_{kn}}{\gamma_n + \mu_n}.$$

Let us argue that $M'_n \in \mathcal{F}_n(\bar{\kappa}_n)$. First, the argument provided in the proof of Lemma A.6 which shows that G_n are measurable on $\mathcal{F}_n(t^{(n)})$ can be used, with small modifications, to show that $\bar{G}_n$ are measurable on $\mathcal{F}_n(t_i)$. Since $\bar{\kappa}_n \geq t_i$ a.s., it follows by $\mathbb{P}$-completeness of the filtration that $\bar{G}_n \in \mathcal{F}_n(t_i) \subset \mathcal{F}_n(\bar{\kappa}_n)$. Also, $\{\kappa_n < \infty\} \in \mathcal{F}_n(\bar{\kappa}_n)$ by Lemma A.7. Thus $M'_n \in \mathcal{F}_n(\bar{\kappa}_n)$. It follows that

$$\mathbb{P}(RD_n = k|M'_n) = \frac{\mu_{kn}}{\gamma_n + \mu_n}.$$

Next, Denote $\overline{RT}_n = RT_n \wedge EX_n$. Since $\kappa_n \leq RT_n$, clearly $\bar{\kappa}_n \leq \overline{RT}_n$. Thus $M'_n \in \mathcal{F}_n(\overline{RT}_n)$. Now, $\overline{RT}_n$ is a finite stopping time, and so $S_k(T_k(\overline{RT}_n) + \cdot) - S_k(T_k(\overline{RT}_n))$, under the conditioning on $\mathcal{F}_n(\overline{RT}_n)$, is a Poisson process with rate μ_k. In addition, on $\{RD_n = k\}$, $\overline{RT}_n = RT_n$. Hence a.s. on $\{RD_n = k\}$,

$$\mathbb{P}(\Sigma_n^i > a|\mathcal{F}_n(\overline{RT}_n)) = e^{-a\mu_{kn}}.$$

Therefore

$$\mathbb{P}(\Sigma_n^i > a|M'_n, RD_n = k) = e^{-a\mu_{kn}}.$$

As a result,

$$\mathbb{P}(\widehat{\Sigma}_n^i > a|M'_n) = \sum_k e^{-\frac{a\mu_{kn}}{\sqrt{n}}} \frac{\mu_{kn}}{\gamma_n + \mu_n} = \frac{1}{\bar{\mu}_n + n^{-1}\gamma_n} \frac{N_n}{\sqrt{n}} \frac{1}{N_n} \sum_k e^{-a\widehat{\mu}_{kn}} \widehat{\mu}_{kn}.$$

Using (7), (15), and (25), the limit of the above expression is

$$\frac{\nu}{\mu} \int e^{-ay} y m(dy).$$

We have thus shown convergence in distribution to a nonnegative r.v., say η, for which $\mathbb{P}(\eta > a) = \int_{(a,\infty)} f(x)dx$, $a > 0$, where

$$f(a) = \frac{\nu}{\mu} \int e^{-ay} y^2 m(dy).$$

In other words, comparing with (30)–(31), we have shown $\mathbb{P}(\widehat{\Sigma}_n > a|M'_n) \to \mathbb{P}(\eta_i > a)$. In view of (A.36) and $\mathbb{P}(\kappa_n < \infty) \to 1$ (Lemma A.7(ii)), this implies (A.37). This completes the proof of Claim B.

Proof of Theorem 2.2. Claims A–C establish (A.25). Since $\widehat{Q}_n$ and $\widehat{X}_n$ both converge in distribution to the same limit, and since $AT_n(t) \to t$ in probability for each t, the theorem follows by Lemma A.4(iv) and the fact $\bar{\mu}_n \to \mu$. □

References

[1] D. Aldous. Stopping times and tightness, II. *Ann. Probab.*, 17, No. 2, 586–595 (1989)

[2] R. F. Anderson and S. Orey. Small random perturbation of dynamical systems with reflecting boundary. *Nagoya Math. J.* Vol. 60 (1976), 189–216

[3] R. Atar, A. Budhiraja and K. Ramanan. Deterministic and stochastic differential inclusions with multiple surfaces of discontinuity. *Probab. Theory Related Fields* 142 (2008), no. 1-2, 249–283

[4] R. Atar and M. I. Reiman. Asymptotically optimal dynamic pricing for network revenue management. Work in progress.

[5] P. Billingsley. *Convergence of probability measures. Second edition.* Wiley, New York, 1999.

[6] S. N. Ethier and T. G. Kurtz. *Markov processes. Characterization and convergence.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986

[7] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus.* Second edition. Springer-Verlag, New York, 1991.

[8] P. E. Protter. *Stochastic integration and differential equations. Second edition.* Springer-Verlag, Berlin, 2004

[9] P. Robert. *Stochastic networks and queues.* Applications of Mathematics (New York), 52. Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2003.

Department of Electrical Engineering
Technion–Israel Institute of Technology
Haifa 32000, Israel