RANK-BASED STOCHASTIC DIFFERENTIAL INCLUSIONS AND DIFFUSION LIMITS FOR A LOAD BALANCING MODEL

RAMI ATAR AND TOMOYUKI ICHIBA

Abstract. In [4], a randomized load balancing model was studied in a heavy traffic asymptotic regime where the load balancing stream is thin compared to the total arrival stream. It was shown that the limit is given by a system of rank-based Brownian particles on the half-line. This paper extends the results of [4] from the case of exponential service time to an invariance principle, where service times have finite second moment. The main tool is a new notion of rank-based stochastic differential inclusion, which may be of interest in its own right.

1. Introduction

This paper studies a randomized load balancing model under an asymptotic regime introduced by Banerjee, Budhiraja and Estevez [4], where N servers, labeled $1, \dots, N$, cater to $N+1$ streams of jobs, labeled $0, \dots, N$. For each $1 \le i \le N$, server i caters to stream i, and stream 0 undergoes randomized load balancing. Specifically, the power-of-choice algorithm is applied to this stream, in which ℓ out of the N queues are chosen at random and the job is routed to the shortest among them. In the regime proposed in [4], the load balancing stream is much thinner than the remaining streams. The motivation to study a thin load balancing stream stems from the fact that the intensity of this stream, along with the parameter ℓ, determines the communication volume between the dispatcher and servers, which one wants to keep low in practical applications. In [4], N is fixed, and, denoting the scaling parameter by n, the server capacity and the arrival rates of streams $1, \dots, N$ scale like n, whereas the arrival rate of stream 0 scales like $n^{1/2}$. Under a critical load condition, it is shown that the N queue lengths, normalized by $n^{1/2}$, converge to a system of Brownian particles on the half line, with rank-dependent drift coefficients. This result thus identifies the minimal order of magnitude of the load balancing stream intensity at which load balancing has a macroscopic effect on the system behavior at the diffusion scale.

The treatment in [4] assumes exponential service times. The goal of this paper is to extend the diffusion limit result to an invariance principle, where service times are only assumed to possess a second moment. This does not amount to a mere technical improvement of the proof ideas of [4]. To briefly explain this point, let $\hat{X}_i^n$, $1 \le i \le N$ denote the normalized queue

Date: September 22, 2024.

2010 *Mathematics Subject Classification.* 60K25, 60J60, 60H10, 34A60.

Key words and phrases. Load balancing, rank-based stochastic differential equations, rank-based stochastic differential inclusions, pathwise uniqueness.

length process of server i. Let the *ranked* normalized queue lengths $\hat{Y}_i^n$, $1 \le i \le N$, be defined as

$$\hat{Y}_i^n(t) = \hat{X}^n_{\pi_t(i)}(t),$$

for a t-dependent permutation π_t of $\{1, \dots, N\}$ that ensures that, for all t,

$$\hat{Y}_1^n(t) \le \hat{Y}_2^n(t) \le \cdots \le \hat{Y}_N^n(t).$$

The limiting dynamics of $\hat{X}_i^n$ is expected to be given in terms of a system of stochastic differential equations (SDEs) on $\mathbb{R}_+^N$ of the form (written here in a slightly simplified manner; see Section 2.3 for a precise definition and the general form of SDEs treated here),

$$X_i(t) = X_i(0) + B_i(t) + \int_0^t b_{\mathcal{R}_i(s)} ds + L_i(t), \tag{1.1}$$

where b_i are constants, $\mathcal{R}_i(t)$ is the rank of $X_i(t)$ (e.g., $\mathcal{R}_i(t) = 1$ if $X_i(t) = \min_j X_j(t)$), B_i are mutually independent Brownian motions, and L_i is a local time term for process X_i at the origin. In particular, $b_{\mathcal{R}_i(t)}$ are rank-dependent drift coefficients. If now one lets Y_i, $1 \le i \le N$ denote the ranked processes corresponding to $\{X_i\}$ (by a transformation like the one above for the normalized queue lengths), one finds that they satisfy a system of SDEs of the form

$$Y_i(t) = Y_i(0) + \tilde{B}_i(t) + b_i t + \frac{1}{2}\tilde{L}_i(t) - 1_{\{i<N\}}\frac{1}{2}\tilde{L}_{i+1}(t) + 1_{\{i=1\}}\frac{1}{2}\tilde{L}_1(t), \tag{1.2}$$

where $\tilde{L}_i$ are local time terms. For background on ranked processes of semimartingales, see [6]. The strategy of the proof in [4] is to work with the approximating processes $\hat{Y}_i^n(t)$ and show convergence to the unique solution of (1.2). In this technique, one needs to show that the intersection time

$$\int_0^\cdot 1_{\{\hat{Y}_i^n(t) = \hat{Y}_{i+1}^n(t)\}} dt$$

converges to zero in probability, as $n \to \infty$ for every $i = 1, \dots, N-1$ (Lemma 4.2 and Corollary 4.3 of [4]), and here the exponential service time assumption is convenient. Estimating it under more general conditions may be quite challenging.

Our proof strategy is to work with the processes $\hat{X}_i^n$ and show that they converge to the unique solution of (1.1). Because (1.1) itself does not involve multiple local-time terms except the one for spending time at zero, there is no need to estimate the intersection times. On the other hand, our approach requires the notion of rank-based stochastic differential inclusions (SDIs), such as

$$X_i(t) = X_i(0) + B_i(t) + \int_0^t \beta_i(s) ds + L_i(t). \tag{1.3}$$

Here, $\beta_i(t) = b_{\mathcal{R}_i(t)}$ at times when all particles are isolated from each other, but a milder condition is required at times when some of the particles collide. The precise condition is formulated by requiring $\{\beta_i(t)\}$ to belong to a set that depends on the current state $\{X_i(t)\}$ (see Section 2.3).

Differential inclusions play an important role in the study of non-smooth dynamical systems in both deterministic [2, 12] and stochastic [1] settings. An intuitive explanation for their effectiveness here is that while a precise formulation of (1.1) must specify the tie-breaking rule for how $\mathcal{R}_i(t)$ are defined when two particles collide, such details should not matter. In fact,

a differential inclusion avoids keeping track of such details. When weak limits are taken for the queueing model, details on tie breaking rules are lost, and one is left with a differential inclusion which, in its simplest form, is given by (1.3).

For a survey on randomized load balancing algorithms and their recent extensive study in asymptotic regimes, see [10]. Rank-based diffusions have also been extensively studied in recent years; see e.g., [3, 5, 9, 11, 14, 18, 19].

1.1. **Notation.** Throughout the paper, we use the following notation. $[N] = \{1, \dots, N\}$, $\mathbb{R}_+ = [0, \infty)$. ι denotes the identity map on $\mathbb{R}_+$. In $\mathbb{R}^N$, the Euclidean norm is denoted by $\|\cdot\|$. For (X, d_X) a Polish space, let $C(\mathbb{R}_+, X)$ and $D(\mathbb{R}_+, X)$ denote the space of continuous paths and, respectively, càdlàg paths, endowed with the topology of uniform convergence on compacts and, respectively, the Skorokhod J_1 topology. Let $C_0^{\uparrow}(\mathbb{R}_+, \mathbb{R}_+)$ denote the subset of $C(\mathbb{R}_+, \mathbb{R}_+)$ of nondecreasing functions that vanish at zero. For $f : \mathbb{R}_+ \to \mathbb{R}^N$, denote

$$\|f\|_t^* := \sup_{s \in [0,t]} \|f(s)\|,$$
$$w_t(f, \delta) := \sup\{\|f(u) - f(s)\| : u, s \in [0, t], |u - s| \le \delta\}.$$

If $v \in \mathbb{R}^N$, then v_i, $i \in [N]$ denote its coordinates, and vice versa: if v_i, $i \in [N]$ are given, then v denotes $(v_1, \dots, v_N)$. The same convention holds for random elements v_i and stochastic processes $v_i(\cdot)$. The symbol $\Rightarrow$ denotes convergence in distribution.

2. Model and main results

2.1. **The load balancing model.** The model we consider is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the corresponding expectation $\mathbb{E}$. It consists of $N + 1$ arrival streams, labeled $0, 1, \dots, N$, as well as N queues and N servers, labeled $1, \dots, N$. For $i \in [N]$, server i serves jobs in queue i in the order of arrival to the queue. The N queues are fed by the $N + 1$ arrival streams, where all jobs from arrival stream $i \in [N]$ are routed to queue i, whereas jobs from stream 0, called the *load balancing stream* (LBS), are routed to the N queues according to a randomized load balancing algorithm described below.

The sequence of systems is indexed by $n \in \mathbb{N}$. The processes X_i^n, E_i^n, D_i^n and T_i^n, $i \in [N]$, represent the i-th queue length process, arrival process, departure process and cumulative busyness process, respectively. Moreover, A_0^n denotes the LBS arrival process and A_i^n the process counting LBS arrivals routed by the algorithm to server i.

For each $n \in \mathbb{N}$, the $N + 1$ arrival processes E_i^n and A_0^n are mutually independent Poisson processes of intensities λ_i^n and λ_0^n, respectively. Alternatively, they can be viewed as $N + 1$ Poisson thinned streams of a single arrival stream, obtained by random selection. These processes are assumed to have right-continuous sample paths. We have the balance equation

$$X_i^n(t) = X_i^n(0-) + E_i^n(t) + A_i^n(t) - D_i^n(t), \qquad i \in [N],\ t \in \mathbb{R}_+, \tag{2.1}$$

and, assuming work conservation, the cumulative busyness process is given by

$$T_i^n(t) := \int_0^t 1_{\{X_i^n(s) > 0\}} ds, \quad i \in [N],\ t \in \mathbb{R}_+. \tag{2.2}$$

Note that $t - T_i^n(t)\,,\ t \geq 0\,,\ i \in [N]$ are the cumulative idle time processes.

The load balancing algorithm routes LBS jobs to queues according to their relative lengths. The precise construction requires the definition of a rank function. Namely, rank : $[N] \times \mathbb{R}^N \to [N]$ is defined as

$$\text{rank}(i;x) := \#\{j \in [N] : x_j < x_i\} + \#\{j \leq i : x_j = x_i\}, \qquad i \in [N],\ x \in \mathbb{R}^N. \tag{2.3}$$

Thus, $\text{rank}(i;x)$ is the rank of x_i among $\{x_j\}$, where the tie-breaking rule is that the smaller index among the ties is preferred. For example, for $x = (1,1,2,2,3) \in \mathbb{R}^5$, one has $\text{rank}(i;x)_{i=1}^5 = (1,2,3,4,5)$, and for $x = (1,1,3,2,2) \in \mathbb{R}^5$, $\text{rank}(i;x)_{i=1}^5 = (1,2,5,3,4)$.

Next, a probability vector $p = (p_r) \in [0,1]^N$, $p_1 + \cdots + p_N = 1$, is given, assumed throughout to satisfy

$$p_1 \geq p_2 \geq \cdots \geq p_N. \tag{2.4}$$

The load balancing algorithm routes an LBS job to the queue whose length is ranked r with probability p_r (in particular, shorter queues are preferred). For the construction of this part of the queueing model, let θ_k, $k \in \mathbb{N}$ be IID random variables with the common distribution $\mathbb{P}(\theta_1 = r) = p_r$, $r \in [N]$. Then the cumulative number $A_i^n(t)$ of the LBS arrivals routed to server i by time t is given by

$$A_i^n(t) = \int_{[0,t]} 1_{\{\mathcal{R}_i^n(s-) = \theta_{A_0^n(s)}\}} dA_0^n(s), \qquad \mathcal{R}_i^n(t) = \text{rank}(i; X^n(t)), \qquad i \in [N],\ t \in \mathbb{R}_+. \tag{2.5}$$

The most important special cases of this setting are two versions of the well-known *power-of-choice* algorithm. Here, upon an LBS arrival, the queue lengths of ℓ out of the N queues, chosen uniformly at random, are sampled (with or without replacement). The arrival is routed to the queue that is the shortest among the ℓ queues. The volume of communication between the servers and the dispatcher is therefore proportional to ℓ. For this reason, in practice, ℓ is usually chosen much smaller than N.

Under the power-of-choice, the probability p_r that an LBS arrival is routed to the queue whose rank (as defined by (2.3)) is r, is given by

$$p_r = \Big(\frac{N-r+1}{N}\Big)^\ell - \Big(\frac{N-r}{N}\Big)^\ell, \qquad \text{and} \qquad p_r = \frac{\binom{N-r}{\ell-1}}{\binom{N}{\ell}}, \quad \text{respectively} \qquad r \in [N] \tag{2.6}$$

for sampling with and without replacement, respectively. Here, $\binom{k}{j} = 0$ when $j > k$. In both cases, $\sum_r p_r = 1$, and (2.4) hold. Our results are concerned with general p satisfying (2.4), but the main interest is in the cases (2.6).

Initial conditions. The residual times of jobs that have already been processed at time 0 are assumed to satisfy some mild conditions. Denote the (random) set of queues that at time 0 contains no jobs and, respectively, at least one job, by $\mathcal{N}^n = \{i \in [N] : X_i^n(0) = 0\}$ and $\mathcal{P}^n = \{i \in [N] : X_i^n(0) > 0\}$. Note that $\mathcal{N}^n$ and $\mathcal{P}^n$ partition $[N]$ with $\#\mathcal{N}^n + \#\mathcal{P}^n = N$.

For $i \in \mathcal{P}^n$, let $Z_i^n(0)$ denote the initial residual time of the head-of-line job in queue i. For $i \in \mathcal{N}^n$, let fictitious jobs be added, having zero processing time. To this end, rather than specifying $X^n(0)$ as the initial queue length, $X^n(0-)$ is specified (as in (2.1)); and for each

$i \in \mathcal{N}^n$, the queue length is set to $X_i^n(0-) = 1$ and the residual processing time is set to $Z_i^n(0) = 0$. Note that this results in $X_i^n(0) = 0$. Note that by adding the fictitious jobs we attain the following. The first job to enter service after time 0 will have index 1 regardless of the initial status of the queue (empty or non-empty). The initial condition is thus a tuple

$$\mathcal{I}^n = (\{X_i^n(0-), Z_i^n(0), i \in [N]\}, \mathcal{N}^n, \mathcal{P}^n),$$

where $(\mathcal{N}^n, \mathcal{P}^n)$ partitions $[N]$, and

$$X_i^n(0-) = 1,\ Z_i^n(0) = 0,\ i \in \mathcal{N}^n,$$
$$X_i^n(0-) \geq 1,\ Z_i^n(0) > 0,\ i \in \mathcal{P}^n.$$

Service times. Let $\Phi_i^{\rm ser}$, $i \in [N]$, be Borel probability measures on $[0, \infty)$ with mean 1, standard deviation $\sigma_i^{\rm ser} \in (0, \infty)$, and $\Phi_i^{\rm ser}(\{0\}) = 0$. Let $\Phi_i^{{\rm ser},n}$ be defined as scaled versions of these measures, uniquely specified via

$$\Phi_i^{{\rm ser},n}[0, x] := \Phi_i^{\rm ser}[0, \mu_i^n x], \quad x \in \mathbb{R}_+. \tag{2.7}$$

Here, $\mu_i^n > 0$ is the service rate of server i in the n-th system. For $k \geq 1$, let $Z_i^n(k)$ denote the service time of the k-th job to be served by server i after the head-of-line job at time $0-$ (for $i \in \mathcal{N}^n$ this means the k-th job after the fictitious one). It is assumed that, for every i, $\mathcal{Z}_i^n = (Z_i^n(k), k \geq 1)$ is an IID sequence with common distribution $\Phi_i^{{\rm ser},n}$.

The potential service process S_i^n, evaluated at t, represents the number of jobs completed by server i by the time it has worked t units of time. With $\sum_0^{-1} = 0$, it is given by

$$S_i^n(t) = \max\Big\{k \in \mathbb{Z}_+ : \sum_{j=0}^{k-1} Z_i^n(j) \leq t\Big\}, \qquad t \geq 0. \tag{2.8}$$

The departure processes are, therefore, given by $D_i^n(t) = S_i^n(T_i^n(t))$. This is the number of jobs completed by time t by server i.

It is assumed that, for each n, the $2N + 3$ stochastic elements

$$E_i^n, i \in [N], \quad \mathcal{Z}_i^n, i \in [N], \quad \mathcal{I}^n, \quad A_0^n, \quad \{\theta_k\},$$

are mutually independent.

2.2. **The scaling and critical load condition.** The arrival and service rates are assumed to satisfy the following. There are constants $\lambda_i > 0$ and $\hat{\lambda}_i \in \mathbb{R}$ such that

$$\hat{\lambda}_i^n := n^{-1/2}(\lambda_i^n - n\lambda_i) \to \hat{\lambda}_i, \quad \text{as } n \to \infty, \qquad i \in [N], \tag{2.9}$$

a constant $\lambda_0 > 0$ such that

$$\hat{\lambda}_0^n := n^{-1/2}\lambda_0^n \to \lambda_0, \quad \text{as } n \to \infty, \tag{2.10}$$

and constants $\mu_i > 0$ and $\hat{\mu}_i \in \mathbb{R}$ such that

$$\hat{\mu}_i^n := n^{-1/2}(\mu_i^n - n\mu_i) \to \hat{\mu}_i, \quad \text{as } n \to \infty, \qquad i \in [N]. \tag{2.11}$$

Each of the queues is critically loaded, namely

$$\lambda_i \equiv \mu_i, \qquad i \in [N]. \tag{2.12}$$

The scaled initial residual time $n^{1/2}Z_i^n(0)$ are assumed to satisfy

$$n^{1/2}Z_i^n(0) \to 0 \quad \text{in probability}, \qquad i \in [N], \tag{2.13}$$

as $n \to \infty$.

Finally, we specify conditions regarding the initial queue lengths $X_i^n(0-)$, representing two different scenarios. One with initial queue lengths of the order $n^{1/2}$, and another with queue lengths of a larger order of magnitude. To state these conditions, let us define

$$\hat{X}^n(0-) = (\hat{X}_1^n(0-), \dots, \hat{X}_N^n(0-)), \quad \hat{X}_i^n(0-) := n^{-1/2}X_i^n(0-),\ i \in [N],$$

and, for a fixed sequence α_n satisfying $n^{-1/2}\alpha_n \to \infty$,

$$\check{X}^n(0-) = (\check{X}_1^n(0-), \dots, \check{X}_N^n(0-)), \quad \check{X}_i^n(0-) := n^{-1/2}(X_i^n(0-) - \alpha_n),\ i \in [N].$$

It will be assumed that one of the following holds: Either

$$\hat{X}^n(0-) \Rightarrow X_0 := (X_{0,1}, \dots, X_{0,N}),\ \text{an } \mathbb{R}_+^{[N]}\text{-valued random vector}, \tag{IC$_0$}$$

or

$$\check{X}^n(0-) \Rightarrow X_0 := (X_{0,1}, \dots, X_{0,N}),\ \text{an } \mathbb{R}^{[N]}\text{-valued random vector}. \tag{IC$_\alpha$}$$

The main results of this model are concerned with diffusion-scale versions of the queue length processes $X_i^n(t)$, $t \geq 0$ and cumulative idle time processes $t - T_i^n(t)$, $t \geq 0$. In particular, under (IC$_0$), we will be interested in

$$\hat{X}_i^n(t) := n^{-1/2}X_i^n(t), \qquad \hat{L}_i^n(t) := n^{-1/2}\mu_i^n(t - T_i^n(t)), \quad t \geq 0,\ i \in [N], \tag{2.14}$$

whereas under (IC$_\alpha$), we will study the behavior of

$$\check{X}_i^n(t) := n^{-1/2}(X_i^n(t) - \alpha_n), \quad t \geq 0,\ i \in [N], \tag{2.15}$$

where the queue lengths are centered around the same constants α_n as the initial conditions.

2.3. **Rank-based SDE and SDI.** We are concerned with a rank-based diffusion defined via a system of SDEs or SDIs, with and without reflection.

Let constants $b := (b_1, \dots, b_N) \in \mathbb{R}^N$, $m := (m_1, \dots, m_N) \in \mathbb{R}^N$, $\sigma := (\sigma_1, \dots, \sigma_N) \in (0, \infty)^N$ be given, and consider rank-based SDE without, and, respectively, with reflection,

$$\begin{aligned} &X_i(t) = X_{0,i} + \sigma_i B_i(t) + m_i t + \int_0^t b_{\mathcal{R}_i(s)} ds, \\ &\mathcal{R}_i(t) = \text{rank}(i; X(t)), \end{aligned} \qquad t \geq 0,\ i \in [N]. \tag{SDE}$$

$$\begin{aligned} &X_i(t) = X_{0,i} + \sigma_i B_i(t) + m_i t + \int_0^t b_{\mathcal{R}_i(s)} ds + L_i(t) \geq 0, \\ &\mathcal{R}_i(t) = \text{rank}(i; X(t)), \\ &\int_0^\infty X_i(t) dL_i(t) = 0, \end{aligned} \qquad t \geq 0,\ i \in [N]. \tag{SDER}$$

Here, rank is the function defined in (2.3) and L_i are the continuous, nondecreasing, adapted processes, starting from $L_i(0) = 0$, that make the N-dimensional process $X(\cdot)$ stay in the non-negative orthant $\mathbb{R}_+^N$.

The precise notions of a solution, in strong and weak form, are standard, but we give them for completeness. On a given stochastic basis $(\Omega, \mathcal{F}, \mathbb{F} := \{\mathcal{F}_t, t \geq 0\}, \mathbb{P})$ satisfying the usual conditions, with an $\mathbb{F}$-Brownian motion B in dimension N, and an initial condition $X_0 \in \mathcal{F}_0$, a strong solution of (SDER) is an $\mathbb{F}$-adapted process (X, L) with sample paths in $C(\mathbb{R}_+, \mathbb{R}_+^N) \times C_0^{\uparrow}(\mathbb{R}_+, \mathbb{R}_+)^N$ that satisfies (SDER). A weak solution to (SDER) is a stochastic basis $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ that satisfies the usual conditions, along with processes (X, L, B) defined on it, satisfying (SDER) a.s., where B is an $\mathbb{F}$-Brownian motion in dimension N, $X_0 \in F_0$, and (X, L) is $\mathbb{F}$-adapted, with sample paths in $C(\mathbb{R}_+, \mathbb{R}_+^N) \times C_0^{\uparrow}(\mathbb{R}_+, \mathbb{R}_+)^N$.

With a slight abuse of terminology, we will sometimes say that a tuple (X, L, B), or even (X, L) is a weak solution, without specifying the stochastic basis (or the Brownian motion).

We say that uniqueness in distribution holds for (SDER) if for any two weak solutions (X, L, B), $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ and $(\widetilde{X}, \widetilde{L}, \widetilde{B})$, $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{F}}, \widetilde{\mathbb{P}})$, with the same initial distribution, the two processes (X, L) and $(\widetilde{X}, \widetilde{L})$ have the same distribution. We say that pathwise uniqueness holds for (SDER), if for any two weak solutions (X, L, B), $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ and $(\widetilde{X}, \widetilde{L}, B)$, $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, with common initial value, the two processes X and $\widetilde{X}$ are indistinguishable, that is, $\mathbb{P}(X_t = \widetilde{X}_t, t \geq 0) = 1$, and so are L and $\widetilde{L}$.

Analogous notions are defined for (SDE) in a similar manner, with X having the sample path in $C(\mathbb{R}_+, \mathbb{R}^N)$.

Remark 2.1. *It is shown by Yamada and Watanabe in [20] that the pathwise uniqueness and the weak existence of SDE imply the strong existence of the solution and the uniqueness in law. Following the weak and strong solutions of stochastic system developed by Kurtz in [16], we claim that the pathwise uniqueness and the weak existence of SDI imply the strong existence of the solution and the uniqueness in law. Particularly, the pathwise uniqueness and weak existence of* (SDIR) (*(SDI), respectively*) *imply the strong existence of the solution of* (SDIR) (*(SDI), respectively*) *and uniqueness in law.*

To introduce our notion of a rank-based SDI, let Π denote the set of all permutations of $[N]$. For each $\pi \in \Pi$, write b_π for the vector $(b_{\pi(1)}, \ldots, b_{\pi(N)})$. Let us consider a set-valued map $\mathfrak{P} : \mathbb{R}^N \to 2^\Pi$ defined by

$$\mathfrak{P}(x) = \{\pi \in \Pi : x_i < x_j \text{ implies } \pi(i) < \pi(j) \text{ for every } i, j \in [N]\}, \tag{2.16}$$

for $x \in \mathbb{R}^N$. Denote by $\mathrm{conv}(A)$ the convex hull of set $A \subset \mathbb{R}^N$. The SDI without and, respectively, with reflection, that will be of interest here, are

$$\text{(SDI)} \qquad \begin{aligned} X_i(t) &= X_{0,i} + \sigma_i B_i(t) + m_i t + \int_0^t \beta_i(s) ds, && t \geq 0,\ i \in [N], \\ \beta(t) &= (\beta_1(t), \ldots, \beta_N(t)) \in \mathrm{conv}\{b_\pi : \pi \in \mathfrak{P}(X(t))\} && \text{a.e. } t \in \mathbb{R}_+, \end{aligned}$$

and a similar one on the state space $\mathbb{R}^N_+$,

$$\text{(SDIR)}\qquad \begin{aligned} &X_i(t) = X_{0,i} + \sigma_i B_i(t) + m_i t + \int_0^t \beta_i(s)ds + L_i(t), && t \geq 0,\ i \in [N],\\ &\beta(t) \in \mathrm{conv}\{b_\pi : \pi \in \mathfrak{P}(X(t))\} && \text{a.e. } t \in \mathbb{R}_+,\\ &\int_0^\infty X_i(t)dL_i(t) = 0, && i \in [N]. \end{aligned}$$

We call β the *rank-dependent drift process*. We say that X (respectively, (X, L)) is a solution to (SDI) (respectively, (SDIR)) if there exist an $\mathbb{F}$-progressively measurable process β so that (SDI) (respectively, (SDIR)) holds. We extend the notions of weak/strong solution, weak uniqueness and pathwise uniqueness, given above, to (SDI) and (SDIR).

2.4. **Results.** A vector $b \in \mathbb{R}^N$ is said to be nonincreasing if the sequence $b_1, \ldots, b_N$ is.

Theorem 2.2. *Let data $b, m \in \mathbb{R}^N$, $\sigma \in (0,\infty)^N$ be given and assume that b is nonincreasing. Then pathwise uniqueness and strong existence hold for* (SDE), (SDER), (SDI) *and* (SDIR).

Theorem 2.3. *Consider data defined in terms of the load balancing model, as follows*

$$b_r = \lambda_0 p_r,\ r \in [N], \qquad m_i = \hat{\lambda}_i - \hat{\mu}_i, \qquad \sigma_i = (\lambda_i + \mu_i(\sigma_i^{\mathrm{ser}})^2)^{1/2}, \qquad i \in [N]. \tag{2.17}$$

i. Assume (IC_0). *Then $(\hat{X}^n, \hat{L}^n) \Rightarrow (X, L)$ in $D(\mathbb{R}_+, \mathbb{R}^N) \times C(\mathbb{R}_+, \mathbb{R}^N)$ as $n \to \infty$, where (X, L) is the solution to* (SDIR), *equivalently,* (SDER), *with data* (2.17).
ii. Fix a sequence α_n with $n^{-1/2}\alpha_n \to \infty$ and assume (IC_α). *Then $(\check{X}^n, \hat{L}^n) \Rightarrow (X, 0)$ in $D(\mathbb{R}_+, \mathbb{R}^N) \times C(\mathbb{R}_+, \mathbb{R}^N)$ as $n \to \infty$, where X is the solution to* (SDI), *equivalently,* (SDE), *with data* (2.17).

2.5. **Proof outline.** The proofs of Theorems 2.2 and 2.3 are intertwined: The existence of a limit for the queueing model is based on the uniqueness provided by Theorem 2.2, whereas existence of solutions to both differential inclusions is a consequence of the convergence proved in Theorem 2.3. The steps are as follows.

1. Pathwise uniqueness of (SDI) and (SDIR). This is shown in Section 3, specifically, in Proposition 3.1.

2. This automatically gives the pathwise uniqueness of (SDE) and (SDER). See Remark 3.2.

3. Weak convergence of the rescaled queueing model to a solution of (SDIR) and (SDI) (under (IC_0) and (IC_α), respectively). This is argued by showing that tightness holds and that subsequential weak limits satisfy the differential inclusions, which, in view of step 1, imply the existence of a limit of the entire sequence. This is carried out in Section 4.

4. Step 3 immediately gives weak existence of solutions to both differential inclusions.

5. For (SDI) and (SDIR), the set of times when two or more components X_i meet is shown to have Lesbegue measure zero. This gives weak existence of solutions to (SDE) and (SDER). This is proved in Section 5.

6. The Yamada-Watanabe Theorem now gives strong existence for the four equations (SDI), (SDIR), (SDE) and (SDER) (cf. Remark 2.1). This is also proved in Section 5.

3. SDI and SDE uniqueness

Proposition 3.1. *Pathwise uniqueness holds for* (SDI) *and* (SDIR).

Remark 3.2. *Note that every solution to* (SDE) *is a solution to* (SDI). *The same holds for* (SDER) *and* (SDIR). *Hence, the above immediately gives the pathwise uniqueness to* (SDE) *and* (SDER).

Proof. The starting point for this proof is the *rearrangement inequality* [13], which states that if u_i, $i \in [N]$ is nonincreasing and v_i, $i \in [N]$ is nondecreasing, then

$$\sum_{i=1}^{N} u_i v_i \leq \sum_{i=1}^{N} u_{\pi(i)} v_i \tag{3.1}$$

for any permutation $\pi \in \Pi$.

Let two weak solutions (X, L, B) and (Y, M, B), defined on the same stochastic basis, be given, and let β and γ be the corresponding rank-dependent drift processes. The difference process $V = (V_1, \dots, V_N)$ with $V_i(\cdot) := X_i(\cdot) - Y_i(\cdot)$, $i \in [N]$ satisfies

$$V_i(\cdot) := X_i(\cdot) - Y_i(\cdot) = \int_0^{\cdot} (\beta_i - \gamma_i) ds + L_i(\cdot) - M_i(\cdot),$$

and hence, the squared norm satisfies

$$\begin{aligned}\frac{1}{2}\|V(t)\|^2 &= \int_0^t V(s) \cdot [(\beta(s) - \gamma(s))ds + dL(s) - dM(s)] \\ &= \sum_{i=1}^{N} \int_0^t (X_i(s) - Y_i(s))(\beta_i(s) - \gamma_i(s))ds + \sum_{i=1}^{N} \int_0^t (X_i(s) - Y_i(s))(dL_i(s) - dM_i(s)).\end{aligned}$$

The proof of pathwise uniqueness for (SDIR) will be complete if one shows that the right-hand above is non-positive for all t. The second sum is clearly nonpositive because $\int_0^{\cdot} X_i(t) dL_i(t) = 0$ while $\int_0^{\cdot} X_i(t) dM_i(t) \geq 0$, and similarly for the Y_i parts in the second sum. Thus, it suffices to show that

$$\sum_{i=1}^{N} (X_i(t) - Y_i(t))(\beta_i(t) - \gamma_i(t)) \leq 0, \quad \text{a.e. } t \in \mathbb{R}_+. \tag{3.2}$$

A similar argument for (SDI), only slightly simpler as it does not involve the boundary terms L and M, also leads to the conclusion that proving (3.2) suffices. We therefore proceed to show (3.2) for both (SDI) and (SDIR).

Following the definition of the SDI, let us write

$$\beta(t) = \sum_{\pi \in \Pi} g_\pi(t) b_\pi, \quad \gamma(t) = \sum_{\pi \in \Pi} h_\pi(t) b_\pi, \quad t \geq 0,$$

where $g_\pi(t) \geq 0$, $\pi \in \Pi$, $\sum_{\pi \in \Pi} g_\pi(t) = 1$, and $g_\pi(t) = 0$ for $\pi \notin \mathfrak{P}(X(t))$ and similarly, $h_\pi(t) \geq 0$, $\pi \in \Pi$, $\sum_{\pi \in \Pi} h_\pi(t) = 1$, and $h_\pi(t) = 0$ for $\pi \notin \mathfrak{P}(Y(t))$ for $t \geq 0$. By the definition of $\mathfrak{P}(\cdot)$ in (2.16), we have therefore that for a.e. t, the conjunction of conditions $g_\pi(t) > 0$

and $X_i(t) < X_j(t)$ implies $\pi(i) < \pi(j)$. A similar statement holds for $h_\pi(t)$ and $Y(t)$. Then, omitting t in the formulas below, we now have

$$\begin{aligned}\sum_{i=1}^N (X_i - Y_i)(\beta_i - \gamma_i) &= \sum_{i=1}^N (X_i - Y_i) \sum_{\pi \in \Pi} [g_\pi b_{\pi(i)} - h_\pi b_{\pi(i)}] \\ &= \sum_{i=1}^N (X_i - Y_i) \sum_{\pi,\sigma \in \Pi} g_\pi h_\sigma (b_{\pi(i)} - b_{\sigma(i)}) \\ &= \sum_{\pi,\sigma \in \Pi} \sum_{i=1}^N (X_i - Y_i) g_\pi h_\sigma (b_{\pi(i)} - b_{\sigma(i)}).\end{aligned}$$

Consider the terms involving only the X_i's in the above sum, namely

$$\sum_{\pi,\sigma} \sum_{i=1}^N X_i g_\pi h_\sigma (b_{\pi(i)} - b_{\sigma(i)}).$$

To prove nonpositivity of this sum, it suffices to prove that, for any π for which $g_\pi > 0$ (that is, $\pi \in \mathfrak{P}(X)$), and for an arbitrary $\sigma \in \Pi$,

$$\sum_{i=1}^N X_i (b_{\pi(i)} - b_{\sigma(i)}) \le 0.$$

Because of the monotonicity: if $\pi(i) > \pi(j)$ then $X_i \ge X_j$, reordering i's so that $\pi(i)$ is increasing will give that X_i are nondecreasing. It will also give (by the assumption on b) that $b_{\pi(i)}$ is nonincreasing. Hence, the last display follows from the rearrangement inequality (3.1). Interchanging the roles of X, π, g_π and those of Y, σ, h_σ, we also obtain the inequality for Y. Therefore, we conclude that (3.2) is true. This shows that $\|V(\cdot)\| \equiv 0$ for both (SDI) and (SDIR). □

4. Diffusion limits

This section provides the main step toward proving Theorem 2.3. In Subsection 4.1, under condition (IC_0), it is shown that $(\hat{X}^n, \hat{L}^n)$ converge to the unique solution of (SDIR). This proves Theorem 2.3(i) except for the statement regarding (SDER), which is treated later in Section 5. Subsection 4.2 assumes (IC_α), and, similarly, proves Theorem 2.3(ii) except the statement on (SDE), whose proof is deferred to Section 5. Most of the work is done in Subsection 4.1.

4.1. The limit under condition (IC_0).

In this subsection we prove Proposition 4.1 below. The proof relies on the uniqueness of (SDIR) proved in the previous section. Let

$$\hat{E}_i^n(t) = n^{-1/2}(E_i^n(t) - \lambda_i^n t), \qquad \hat{S}_i^n(t) = n^{-1/2}(S_i^n(t) - \mu_i^n t), \tag{4.1}$$

$$\hat{A}_0^n(t) = n^{-1/2} A_0^n(t), \qquad \hat{A}_i^n(t) = n^{-1/2} A_i^n(t), \tag{4.2}$$

$$\hat{P}_i^{\#,n}(t) = \lambda_0 \int_0^t p_{\mathcal{R}_i^n(s)} ds, \tag{4.3}$$

$$\hat{m}_i^n = \hat{\lambda}_i^n - \hat{\mu}_i^n. \tag{4.4}$$

Throughout this subsection, let condition (IC_0) hold.

Proposition 4.1. *i. The sequence $(\hat{X}^n, \hat{L}^n, \hat{E}^n, \hat{S}^n, \hat{P}^{\#,n})$ is C-tight.*
ii. If (X, L, E, S, P) is a subsequential weak limit, then (X, L, B) forms a weak solution to (SDIR) *with the data indicated in Theorem 2.3, and where $B_i = \sigma_i^{-1}(E_i - S_i)$ (with σ_i as in* (2.17)*) and the progressively measurable rank-dependent drift β is the a.e. derivative of P.*
iii. Consequently, denoting $\hat{B}_i^n = \sigma_i^{-1}(\hat{E}_i^n - \hat{S}_i^n)$, one has $(\hat{X}^n, \hat{L}^n, \hat{B}^n) \Rightarrow (X, L, B)$, where the latter is a (weak) solution of (SDIR).

Let the Skorokhod map on the half-line $\Gamma : D(\mathbb{R}_+, \mathbb{R}) \to D(\mathbb{R}_+, \mathbb{R}_+)^2$ be defined by

$$\Gamma(y) = (x, z) \qquad \text{where} \qquad x(t) = y(t) + z(t), \qquad z(t) = \sup_{s \in [0,t]} y^-(s), \qquad t \geq 0.$$

Note that if $(x, z) = \Gamma(y)$ then

$$z(t) \leq \|y\|_t^*, \qquad w_t(z, \delta) \leq w_t(y, \delta), \qquad t > 0,\ \delta > 0. \tag{4.5}$$

Lemma 4.2. *Let $(y, x, z) \in D(\mathbb{R}_+, \mathbb{R}) \times D(\mathbb{R}_+, \mathbb{R}_+)^2$ satisfy $x = y + z$. Then the condition z is nondecreasing and $\int_{[0,\infty)} x(t) dz(t) = 0$ (with the convention $z(0-) = 0$) holds if and only if $(x, z) = \Gamma(y)$.*

Proof. This is known as Skorokhod's lemma [8, §8]. □

The tuple

$$\mathcal{S}^n(t) = (E_i^n(t), A_i^n(t), D_i^n(t), X_i^n(t), T_i^n(t),\ i \in [n],\ A_0^n(t), \theta_{A_0^n(t)})$$

is used to define the 'history' of the system, namely the filtration

$$\mathcal{F}_t^n = \sigma\{\mathcal{I}^n, \mathcal{S}^n(s), s \in [0, t]\}, \quad t \geq 0.$$

Let

$$\hat{P}_i^n(t) = \hat{\lambda}_0^n \int_0^t p_{\mathcal{R}_i^n(s)} ds \qquad \text{and} \qquad \hat{M}_i^n(t) = \hat{A}_i^n(t) - \hat{P}_i^n(t). \tag{4.6}$$

Lemma 4.3. *The process $\hat{M}_i^n$ is an $\{\mathcal{F}_t^n\}$-martingale, with optional quadratic variation $[\hat{M}_i^n](t) = n^{-1} A_i^n(t)$.*

Proof. Clearly A_i^n and $\mathcal{R}_i^n$, defined in (2.5), are $\mathcal{F}_t^n$-adapted, and $A_i^n(t)$ is integrable for all t. Hence, the same is true for $\hat{M}_i^n$. Next, let

$$t^n(k) = \inf\{t \geq 0 : A_0^n(t) \geq k\}, \qquad k = 1, 2, \ldots.$$

These are the stopping times on $\{\mathcal{F}_t^n\}$. Hence

$$t^n(k) \in \mathcal{F}_{t^n(k)-}^n, \qquad k \geq 1, \tag{4.7}$$

where we recall that for a stopping time τ,

$$\mathcal{F}^n_{\tau-} = \mathcal{F}^n_0 \vee \sigma\{A : A \cap \{\tau < t\} \in \mathcal{F}^n_t, t \geq 0\}$$

(see [15, I.1.11 and I.1.14]). To show the martingale property, we can write, using $\int_0^t p_{\mathcal{R}^n_i(s-)} ds = \int_0^t p_{\mathcal{R}^n_i(s)} ds$,

$$\begin{aligned} n^{1/2}\hat{M}^n_i(t) &= \int_{[0,t]} 1_{\{\mathcal{R}^n_i(s-)=\theta^n_{A^n_0(s)}\}} dA^n_0(s) - \lambda^n_0 \int_0^t p_{\mathcal{R}^n_i(s)} ds \\ &= \int_{[0,t]} (1_{\{\mathcal{R}^n_i(s-)=\theta^n_{A^n_0(s)}\}} - p_{\mathcal{R}^n_i(s-)}) dA^n_0(s) + \int_0^t p_{\mathcal{R}^n_i(s-)} (dA^n_0(s) - \lambda^n_0 ds) \\ &=: M^n_{i,1}(t) + M^n_{i,2}(t). \end{aligned}$$

For $M^n_{i,1}$, write

$$A^n_i(t) = \sum_{k=1}^{A^n_0(t)} 1_{\{\mathcal{R}^n_i(t^n(k)-)=\theta^n_k\}}.$$

The history of the system up to $t^n(k)-$, namely $\{\mathcal{S}^n(t), t < t^n(k)\}$, can be recovered from the tuple $\mathcal{I}^n$, $(E^n_i(t), t \in \mathbb{R}_+, i \in [N])$, $(A^n_0(t), t \in \mathbb{R}_+)$, $(Z^n_i(j), j \in \mathbb{N}, i \in [N])$ and finally, $(\theta^n_j, j \leq k-1)$. By our assumptions, θ^n_k is independent of this tuple. As a result, it is independent of $\mathcal{F}^n_{t^n_i(k)-}$. Therefore, for $0 \leq s < t$, we have

$$\begin{aligned} \mathbb{E}[A^n_i(t)|\mathcal{F}^n_s] - A^n_i(s) &= \sum_{k=1}^{\infty} \mathbb{E}[1_{\{\mathcal{R}^n_i(t^n(k)-)=\theta^n_k\}} 1_{\{s<t^n(k)\leq t\}} |\mathcal{F}^n_s] \\ &= \sum_{k=1}^{\infty} \mathbb{E}[\mathbb{E}[1_{\{\mathcal{R}^n_i(t^n(k)-)=\theta^n_k\}} 1_{\{s<t^n(k)\leq t\}} |\mathcal{F}^n_{t^n(k)-}]|\mathcal{F}^n_s] \\ &= \sum_{k=1}^{\infty} \mathbb{E}[p_{\mathcal{R}^n_i(t^n(k)-)} 1_{\{s<t^n(k)\leq t\}} |\mathcal{F}^n_s] \\ &= \mathbb{E}[C^n_i(t)|\mathcal{F}^n_s] - C^n_i(s), \end{aligned}$$

where

$$C^n_i(t) = \sum_{k=1}^{A^n_0(t)} p_{\mathcal{R}^n_i(t^n(k)-)} = \int_{[0,t]} p_{\mathcal{R}^n_i(s-)} dA^n_0(s),$$

showing that $A^n_i - C^n_i = M^n_{i,1}$ is a martingale.

In the expression for $M^n_{i,2}$, the integrand is $\{\mathcal{F}^n_t\}$-adapted and has LCRL sample paths, while the integrator is a martingale on this filtration. As a result, $M^n_{i,2}$ is a local martingale [17, Theorem II.20]; Using the estimate $\|M^n_{i,2}\|^*_t \leq A^n_0(t) + c$ shows that it is, in fact, a martingale. As a result, so is $\hat{M}^n_i$. Finally, the expression for the quadratic variation is straightforward. □

Lemma 4.4. *i. One has*

$$\hat{X}^n_i = \hat{U}^n_i + \hat{L}^n_i \qquad \textit{where} \qquad \hat{U}^n_i(t) = \hat{X}^n_i(0-) + \hat{E}^n_i(t) + \hat{A}^n_i(t) - \hat{S}^n_i(T^n_i(t)) + \hat{m}^n_i t. \tag{4.8}$$

Moreover, the sample paths of $\hat{L}_i^n$ *are in* $C_0^\uparrow$ *and*

$$\int_0^\infty \hat{X}_i^n(t)d\hat{L}_i^n(t) = 0. \tag{4.9}$$

In particular,

$$(\hat{X}_i^n, \hat{L}_i^n) = \Gamma(\hat{U}_i^n), \qquad i \in [N]. \tag{4.10}$$

ii. One has $(\hat{E}^n, \hat{S}^n) \Rightarrow (E, S)$*, where the latter is a pair of mutually independent* N*-dimensional Brownian motions starting at zero, with zero drift and diffusion coefficients* $\mathrm{diag}(\lambda_i^{1/2})$ *and* $\mathrm{diag}(\mu_i^{1/2}\sigma_i^{\mathrm{ser}})$*, respectively (where we recall* $\lambda_i = \mu_i$*).*

iii. $\hat{P}_i^n$*,* $\hat{L}_i^n$ *and* $\hat{X}_i^n$ *are* C*-tight, and* $\hat{M}_i^n \to 0$ *in probability.*

Proof. i. By (2.1) and (2.14),

$$\begin{aligned} n^{-1/2}X_i^n(t) = n^{-1/2}X_i^n(0-) + n^{-1/2}(E_i^n(t) - \lambda_i^n t) + n^{-1/2}(\lambda_i^n - n\lambda_i)t + n^{1/2}\lambda_i t + n^{-1/2}A_i^n(t) \\ - n^{-1/2}(S_i^n(T_i^n(t)) - \mu_i^n T_i^n(t)) - n^{-1/2}\mu_i^n T_i^n(t), \end{aligned}$$

and

$$-n^{-1/2}\mu_i^n T_i^n(t) = -n^{-1/2}(\mu_i^n - n\mu_i)t - n^{1/2}\mu_i t + \hat{L}_i^n(t).$$

Using (2.12), (4.1), (4.2) and (4.4) gives (4.8). The properties of $\hat{L}_i^n$ and (4.9) follow from (2.2). The identity (4.10) follows from Lemma 4.2.

ii. The fact that $\hat{E}^n \Rightarrow E$ follows from the central limit theorem for renewal processes [7, §17], and the fact that, by (2.9) $n^{-1}\lambda_i^n \to \lambda_i$, for each i. For $\hat{S}_i^n$, one has to be careful about the fact that the assumptions about $Z_i^n(0)$ differ from those about $Z_i^n(k)$, $k \geq 1$. By (2.8), S_i^n is the inverse of

$$Z_i^n(0) + \sum_{j=1}^{k-1} Z_i^n(j).$$

In the expression

$$n^{1/2}(Z_i^n(0) - (\mu_i^n)^{-1}) + n^{1/2}\sum_{j=1}^{[(n-1)t]}(Z_i^n(j) - (\mu_i^n)^{-1}),$$

the first term converges to 0 in probability by (2.13). In the second term, the summands are IID, and thus its limit in law is a zero drift Brownian motion with diffusion σ_i^{ser}. Hence again by [7, §17], (2.11) and the independence of $\{Z_i^n(j)\}$ across i, we have $\hat{S}^n \Rightarrow S$. The mutual independence of E and S follows from that of $\hat{E}^n$ and $\hat{S}^n$.

iii. Because $\hat{\lambda}_0^n \to \lambda_0$, the processes $\hat{P}_i^n$ are all (λ_0+1)-Lipschitz, null at zero, for n sufficiently large. Hence, they are C-tight. Moreover, by Lemma 4.3 and the calculation

$$\mathbb{E}[n^{-1}A_i^n(t)] = n^{-1}\lambda_0^n \int_0^t \mathbb{E}[p_{\mathcal{R}_i^n(s)}]ds \leq n^{-1}(\lambda_0+1)n^{1/2}t,$$

one has $\hat{M}_i^n \to 0$ in probability. By (4.6), this shows that $\hat{A}_i^n$ are also C-tight.

In view of the identity $(\hat{X}_i^n, \hat{L}_i^n) = \Gamma(\hat{U}_i^n)$ and (4.5),

$$\hat{L}_i^n(t) \le \|\hat{U}_i^n\|_t^*, \qquad w_t(\hat{L}_i^n, \delta) \le w_t(\hat{U}_i^n, \delta). \tag{4.11}$$

Because T_i^n are 1-Lipschitz and $\hat{S}_i^n$ are C-tight, so are $\hat{S}_i^n(T_i^n(\cdot))$. We have already shown that $\hat{E}_i^n$ and $\hat{A}_i^n$ are C-tight. Hence, by the convergence in law of $\hat{X}_i^n(0-)$ assumed in (IC_0) and the convergence $\hat{m}_i^n \to m_i$, which follows from (2.9), (2.11) and (4.4), $\hat{U}_i^n$ are C-tight. In view of (4.11) and the fact $\hat{X}_i^n = \hat{U}_i^n + \hat{L}_i^n$, it follows that $\hat{L}_i^n$ and $\hat{X}_i^n$ are also C-tight. □

Proof of Proposition 4.1.

i. The C-tightness of the sequence follows from Lemma 4.4 parts (ii) and (iii).

ii. Consider a convergent subsequence of $(\hat{X}^n, \hat{L}^n, \hat{E}^n, \hat{S}^n, \hat{P}^n, \hat{U}^n)$, with limit (X, L, E, S, P, U), where E and S are as before. Recall (4.3) and note that one has $\hat{P}_i^n - \hat{P}_i^{\#,n} \to 0$ in probability. Also note by the second part of (2.14) and the tightness of $\hat{L}_i^n$, that $T_i^n \to \iota$ in probability. This and the C-tightness of $\hat{S}_i^n$ implies that $\hat{S}_i^n(T_i^n) - \hat{S}_i^n \to 0$ in probability. Hence, letting $B_i = \sigma_i^{-1}(E_i - S_i)$, one has

$$U_i(t) = X_{0,i} + \sigma_i B_i(t) + P_i(t) + m_i t.$$

The map Γ is continuous in the topology of uniform convergence on compact. Hence, by (4.10), $(X_i, L_i) = \Gamma(U_i)$ for $i \in [N]$, and then by Lemma 4.2,

$$X_i(t) = X_{0,i} + \sigma_i B_i(t) + m_i t + P_i(t) + L_i(t), \qquad \int_0^\infty X_i(t) dL_i(t) = 0.$$

Now, P_i is a.s. Lipschitz and therefore a.s. a.e. differentiable. The existence of a progressively measurable process that is a.e. the derivative of P_i follows by an argument that was given in the proof of [1, Theorem 3.4]. We denote this derivative by β_i.

The goal now is to prove that (SDIR) is satisfied. By invoking Skorokhod's representation theorem, we may assume without loss of generality that convergence holds a.s. In what follows, fix ω in the full $\mathbb{P}$-measure set where convergence holds.

Fix a time interval $[0, t]$. The proof will be complete once it is shown that

$$\mathrm{Leb}(G) = t \qquad \text{where} \qquad G = \{s \in [0,t] : \beta(s) \in \mathrm{conv}\{b_\pi : \pi \in \Pi(X(s))\}\}.$$

For $\varepsilon > 0$ let

$$\Pi^\varepsilon(x) = \{\pi \in \Pi : x_i < x_j - 4\varepsilon \text{ implies } \pi(i) < \pi(j)\}, \qquad x \in \mathbb{R}^N,$$

$$G^\varepsilon = \{s \in [0,t] : \beta(s) \in \mathrm{conv}\{b_\pi : \pi \in \Pi^\varepsilon(X(s))\}\}.$$

Then $\varepsilon \mapsto \Pi^\varepsilon(x)$ is setwise increasing, and

$$\bigcap_{\varepsilon > 0} \Pi^\varepsilon(x) = \Pi(x).$$

As a consequence, $\varepsilon \mapsto G^\varepsilon$ is setwise increasing and

$$\bigcap_{\varepsilon > 0} G^\varepsilon = G.$$

By continuity of measure, it suffices to show that for every $\varepsilon > 0$, $\mathrm{Leb}(G^\varepsilon) = t$. To this end, we first show the following.

$$\text{For every } \varepsilon > 0 \text{ there are } \delta_0 \text{ and } n_0 \text{ such that for } s \in (0,t],\ n > n_0 \text{ and } \delta < \delta_0, \quad \mathcal{R}^n(\theta) \in \Pi^\varepsilon(X(s)), \qquad \theta \in [s, s+\delta]. \tag{4.12}$$

To show (4.12), let n_0 be so large that for all $n > n_0$,

$$\max_i \|\hat{X}_i^n - X_i\|_t^* < \varepsilon.$$

Let $\delta_0 > 0$ be so small that

$$\max_i w_t(X_i, \delta_0) < \varepsilon.$$

To show that $\mathcal{R}^n(\theta)$ is in $\Pi^\varepsilon(X(s))$ is to show that whenever $X_i(s) < X_j(s) - 4\varepsilon$, one has $\mathcal{R}_i^n(\theta) < \mathcal{R}_j^n(\theta)$, that is, $\mathrm{rank}(i; \hat{X}^n(\theta)) < \mathrm{rank}(j; \hat{X}^n(\theta))$. The latter will be guaranteed if $\hat{X}_i^n(\theta) < \hat{X}_j^n(\theta)$ for all $\theta \in [s, s+\delta_0]$. But

$$\hat{X}_j^n(\theta) - \hat{X}_i^n(\theta) > X_j(\theta) - X_i(\theta) - 2\varepsilon > X_j(s) - X_i(s) - 4\varepsilon > 4\varepsilon - 4\varepsilon = 0.$$

This proves (4.12).

In view of (4.12), and recalling $b = \lambda_0 p$, we have for $s \in (0,t]$ and δ and n as above,

$$\delta^{-1}(\hat{P}^{\#,n}(s+\delta) - \hat{P}^{\#,n}(s)) = \delta^{-1}\lambda_0 \int_s^{s+\delta} p_{\mathcal{R}^n(\theta)} d\theta \in \mathrm{conv}\{b_\pi : \pi \in \Pi^\varepsilon(X(s))\}.$$

Since for every $x \in \mathbb{R}^N$, $\mathrm{conv}\{b_\pi : \pi \in \Pi^\varepsilon(x)\}$ is a closed subset of $\mathbb{R}^N$, we also have

$$\delta^{-1}(P(s+\delta) - P(s)) \in \mathrm{conv}\{b_\pi : \pi \in \Pi^\varepsilon(X(s))\}.$$

For a.e. s, the limit of the lefthand is $\beta(s)$. This shows $\mathrm{Leb}(G^\varepsilon) = t$ and completes the proof of part (ii).

iii. The tightness shown in part (i), the fact that limits are supported on solutions to (SDIR), as shown in part (ii), and the uniqueness stated in Proposition 3.1, imply that the entire sequence converges in distribution to the unique weak solution of (SDIR). □

4.2. **The limit under condition** (IC_α)**.** The goal here is to prove Proposition 4.5, which is the analogue of Proposition 4.1. In this subsection, (IC_α) is assumed throughout.

Proposition 4.5. *i. The sequence* $(\check{X}^n, \hat{E}^n, \hat{S}^n, \hat{P}^{\#,n})$ *is C-tight, and* $\hat{L}^n \to 0$ *in probability.*
ii. If (X, E, S, P) *is a subsequential weak limit, then* (X, B) *forms a weak solution to* (SDI) *with the data indicated in Theorem 2.3, and where* $B_i = \sigma_i^{-1}(E_i - S_i)$ *(with* σ_i *as in* (2.17)*) and the progressively measurable rank-dependent drift* β *is the a.e. derivative of* P*.*
iii. Consequently, denoting $\hat{B}_i^n = \sigma_i^{-1}(\hat{E}_i^n - \hat{S}_i^n)$*, one has* $(\hat{X}^n, \hat{B}^n) \Rightarrow (X, B)$*, where the latter is a (weak) solution of* (SDI)*.*

Proof. The arguments are very similar to, only somewhat simpler than those given in Subsection 4.1. Hence, we only indicate the differences.

Lemma 4.3 holds, and its proof is valid as is, as it has nothing to do with conditions (IC_0) or (IC_α). The same is true with respect to Lemma 4.4 parts i and ii.

Also, in part iii of Lemma 4.4, the statements regarding $\hat{P}_i^n$ and $\hat{M}_i^n$ and their proof are valid without any change. As for the remaining content of Lemma 4.4 part iii, we prove instead

$$\hat{L}_i^n \to 0 \text{ in probability, and } \check{X}_i^n \text{ are } C\text{-tight.} \tag{4.13}$$

To this end, subtract $n^{-1/2}\alpha_n$ from both sides of (4.8) to obtain

$$\check{X}_i^n = \check{U}_i^n + \hat{L}_i^n \qquad \text{where} \qquad \check{U}_i^n(t) = \check{X}_i^n(0-) + \hat{E}_i^n(t) + \hat{A}_i^n(t) - \hat{S}_i^n(T_i^n(t)) + \hat{m}_i^n t. \tag{4.14}$$

Now, $\check{U}_i^n$ are C-tight because $\check{X}^n(0-)$ converge under condition (IC_α), and the remaining terms in the definition of $\check{U}_i^n$ are C-tight as already shown.

To prove the claim regarding $\hat{L}_i^n$, note that if $\hat{L}_i^n(T) > 0$ then there exists a time $t \in [0,T]$ such that $X_i^n(t) = 0$ and $L_i^n(t) = 0$. Therefore, $0 = \hat{X}_i^n(t) = \check{X}_i^n(t) + n^{-1/2}\alpha_n$, hence by (4.14), $\check{U}_i^n(t) = -n^{-1/2}\alpha_n$. Thus

$$\mathbb{P}(\hat{L}_i^n(T) > 0) \le \mathbb{P}(\|\check{U}_i^n\|_T^* \ge n^{-1/2}\alpha_n) \to 0,$$

where the last statement follows from the tightness of $\|\check{U}_i^n\|_T^*$ and the fact that $n^{-1/2}\alpha_n \to \infty$. This proves that $\hat{L}_i^n \to 0$ in probability. In view of (4.14), this also shows that $\check{X}^n$ are C-tight, and (4.13) is proved.

Next, if (X,U) is a limit point of $(\check{X}^n, \check{U}^n)$, then we have shown that $X = U$ (compare with $(X_i, L_i) = \Gamma(U_i)$ in the case of Subsection 4.1).

Based on these statements, the completion of the proof of Proposition 4.5 follows closely that of Proposition 4.1, where, in particular, the satisfiability of (SDI) is completely analogous to that of (SDIR). □

5. Proof of main results

Going back to the steps listed in subsection 2.5, note that Proposition 3.1 and Remark 3.2 establish steps 1 and 2, and Propositions 4.1 and 4.5 give steps 3 and 4. Steps 5 and 6 are carried out below, which completes the proof of the main results.

Proof of Theorem 2.3. The convergence to the solution of (SDIR) and (SDI) has already been shown in Proposition 4.1(iii) and Proposition 4.5(iii), respectively. To prove the theorem, it remains to show that if (X, L, B) is a weak solution of (SDIR) then it is also a weak solution of (SDER), and similarly, if (X, B) is a weak solution of (SDI) then it is also a weak solution of (SDE). To this end, it suffices to show that, a.s., for a.e. t, for every $i \neq j$, $X_i(t) \neq X_j(t)$.

Fix $i \neq j$. For the solution X to (SDIR), the difference $X_i(t) - X_j(t)$ is given by

$$X_i(t) - X_j(t) = X_{0,i} - X_{0,j} + \sigma_i B_i(t) - \sigma_j B_j(t) + (m_i - m_j)t + \int_0^t (\beta_i(s) - \beta_j(s))ds + L_i(t) - L_j(t)$$

for $t \ge 0$. By Tanaka's formula, we obtain

$$|X_i(t) - X_j(t)| = |X_{0,i} - X_{0,j}| + \int_0^t \mathrm{sgn}(X_i(s) - X_j(s))d(X_i(s) - X_j(s)) + L^{i,j}(t)$$

for $t \ge 0$, where $\mathrm{sgn}(x) := 1_{\{x>0\}} - 1_{\{x\le 0\}}$, $x \in \mathbb{R}$ and $L^{i,j}(\cdot)$ is the local time accumulated at the origin for the semimartingale $X_i(\cdot) - X_j(\cdot)$. Take a nonnegative function $\phi \in C_b^2(\mathbb{R}_+)$ with

the nonincreasing, nonnegative second derivative ϕ'' that satisfies $\phi''(u) = 1$ for $u \in [0, 1/2]$, $\phi''(u) = 0$ for $u \geq 1$ and $\phi(u) = \phi'(u) = 0$ for $u \geq 1$. Then applying Itô's formula to $\phi(\varepsilon^{-1}|X_i(t) - X_j(t)|)$, we obtain

$$\phi\Big(\frac{1}{\varepsilon}|X_i(t) - X_j(t)|\Big) - \phi\Big(\frac{1}{\varepsilon}|X_{0,i} - X_{0,j}|\Big)$$

$$= \frac{1}{\varepsilon}\int_0^t \phi'\Big(\frac{1}{\varepsilon}|X_i(u) - X_j(u)|\Big)d|X_i(u) - X_j(u)| + \frac{1}{2\varepsilon^2}\int_0^t \phi''\Big(\frac{1}{\varepsilon}|X_i(u) - X_j(u)|\Big)(\sigma_i^2 + \sigma_j^2)du$$

for $\varepsilon > 0$ and $t \geq 0$. This implies that

$$\int_0^t \phi''\Big(\frac{1}{\varepsilon}|X_i(u) - X_j(u)|\Big)du = \frac{2\varepsilon^2}{(\sigma_i^2 + \sigma_j^2)}\Big(\phi\Big(\frac{1}{\varepsilon}|X_i(t) - X_j(t)|\Big) - \phi\Big(\frac{1}{\varepsilon}|X_{0,i} - X_{0,j}|\Big)\Big)$$

$$-\frac{2\varepsilon}{\sigma_i^2 + \sigma_j^2}\int_0^t \phi'\Big(\frac{1}{\varepsilon}|X_i(u) - X_j(u)|\Big)d|X_i(u) - X_j(u)|$$

for $\varepsilon > 0$ and $t \geq 0$. Taking the limits as $\varepsilon \downarrow 0$, we obtain

$$\liminf_{\varepsilon\downarrow 0}\int_0^t \phi''\Big(\frac{1}{\varepsilon}|X_i(u) - X_j(u)|\Big)du = 0$$

almost surely. Since $\varphi''(0) = 1$, this implies, by Fatou's lemma, that

$$\begin{aligned}\int_0^t 1_{\{X_i(u)=X_j(u)\}}du &= \int_0^t \liminf_{\varepsilon\downarrow 0}\phi''\Big(\frac{1}{\varepsilon}|X_i(u) - X_j(u)|\Big)du \\ &\leq \liminf_{\varepsilon\downarrow 0}\int_0^t \phi''\Big(\frac{1}{\varepsilon}|X_i(u) - X_j(u)|\Big)du = 0\end{aligned} \tag{5.15}$$

Therefore, the set of times when two or more components X_i meet is shown to have Lesbegue measure zero. If (X, L, B) is a weak solution of (SDIR), then it is also a weak solution of (SDER).

The solution of (SDI) can be handled in a similar manner. If (X, B) is a weak solution of (SDI), then it is also a weak solution of (SDE). □

Proof of Theorem 2.2. Pathwise uniqueness for (SDI), (SDIR), (SDE) and (SDER) has been shown in Proposition 3.1 and Remark 3.2. Weak existence for the four equations follows from Theorem 2.3. It remains to prove strong existence. To this end we employ the Yamada-Watanabe Theorem (see Remark 2.1). □

Acknowledgement. The first author is partially supported by ISF grant 1035/20. The second author is partially supported by NSF grant DMS-2008427. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Stochastic systems for anomalous diffusions (SSD), where a part of the work on this paper was undertaken.

References

[1] R. Atar, A. Budhiraja, and K. Ramanan. Deterministic and stochastic differential inclusions with multiple surfaces of discontinuity. *Probab. Th. Rel. Fields*, 142:249–283, 2008.

[2] J.-P. Aubin and A. Cellina. *Differential inclusions : set-valued maps and viability theory*. Springer-Verlag, Berlin, Heidelberg, 1984.

[3] S. Banerjee and A. Budhiraja. Domains of attraction of invariant distributions of the infinite Atlas model. *Ann. Probab.*, 50(4):1610–1646, 2022. ISSN 0091-1798. URL https://doi.org/10.1214/22-aop1570.

[4] S. Banerjee, A. Budhiraja, and B. Estevez. Load balancing in parallel queues and rank-based diffusions. *Math. Oper. Res., to appear. arXiv:2302.10317*, 2023.

[5] A. D. Banner, R. Fernholz, and I. Karatzas. Atlas models of equity markets. *Ann. Appl. Probab.*, 15(4): 2296–2330, 2005. ISSN 1050-5164. URL https://doi.org/10.1214/105051605000000449.

[6] A. D. Banner and R. Ghomrasni. Local times of ranked continuous semimartingales. *Stochastic Process. Appl.*, 118(7):1244–1253, 2008. ISSN 0304-4149. URL https://doi.org/10.1016/j.spa.2007.08.001.

[7] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013.

[8] K. L. Chung and R. J. Williams. *Introduction to Stochastic Integration*, volume 2. Springer, 1990.

[9] C. Cuchiero, L. Di Persio, F. Guida, and S. Svaluto-Ferro. Measure-valued affine and polynomial diffusions. *Stochastic Process. Appl.*, 175:Paper No. 104392, 29, 2024. ISSN 0304-4149. URL https://doi.org/10.1016/j.spa.2024.104392.

[10] M. V. der Boor, S. C. Borst, J. S. Van Leeuwaarden, and D. Mukherjee. Scalable load balancing in networked systems: A survey of recent advances. *SIAM Review*, 64(3):554–622, 2022.

[11] E. R. Fernholz. *Stochastic portfolio theory*, volume 48 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 2002. ISBN 0-387-95405-8. xiv+177 pp. URL https://doi.org/10.1007/978-1-4757-3699-1. Stochastic Modelling and Applied Probability.

[12] A. F. Filippov. *Differential equations with discontinuous righthand sides: control systems*, volume 18. Springer Science & Business Media, 2013.

[13] G. Hardy, J. Littlewood, and G. Polya. *Inequalities*. Cambridge University Press, 1952.

[14] T. Ichiba, I. Karatzas, and M. Shkolnikov. Strong solutions of stochastic equations with rank-based coefficients. *Probab. Theory Related Fields*, 156(1-2):229–248, 2013. ISSN 0178-8051. URL https://doi.org/10.1007/s00440-012-0426-3.

[15] J. Jacod and A. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer-Verlag, Berlin, 1987.

[16] T. G. Kurtz. Weak and strong solutions of general stochastic models. *Electron. Commun. Probab.*, 19:no. 58, 16, 2014. URL https://doi.org/10.1214/ECP.v19-2833.

[17] P. Protter. *Stochastic Integration and Differential Equations. A New Approach*. Springer, Berlin, 1990.

[18] A. Sarantsev and L.-C. Tsai. Stationary gap distributions for infinite systems of competing Brownian particles. *Electron. J. Probab.*, 22:Paper No. 56, 20, 2017. URL https://doi.org/10.1214/17-EJP78.

[19] L.-C. Tsai. Stationary distributions of the Atlas model. *Electron. Commun. Probab.*, 23:Paper No. 10, 10, 2018. URL https://doi.org/10.1214/18-ECP112.

[20] T. Yamada and S. Watanabe. On the uniqueness of solutions of stochastic differential equations. *J. Math. Kyoto Univ.*, 11:155–167, 1971. ISSN 0023-608X. URL https://doi.org/10.1215/kjm/1250523691.

Viterbi Faculty of Electrical and Computer Engineering, Technion – Israel Institute of Technology

Department of Statistics and Applied Probability, University of California Santa Barbara