Asymptotics of mean-field closed networks

Alexander Vladimirov †

Institute for Information Transmission Problems
The author gratefully acknowledges the support of grants 16-29-09497, 14-01-00379, 14-01-00319 by Russian Foundation for Sciences.

Abstract

We establish the convergence of equilibria of finite symmetric closed networks with the FIFO service discipline and a general service time with bounded second moment to the unique equilibrium of the non-linear Markov process.

1 Introduction

Let us consider a single-class closed network $\mathcal{S}_{N,M}$ with M particles (customers) that live on a complete graph with N identical nodes (servers). Each node is a server with the FIFO service discipline and with a general distribution F of service time s. The evolution of the network goes on as follows.

It is assumed that each particle j waits in the queue at the server i until all the particles that stay ahead of j in the same queue complete their service. Then, immediately, the server i begins to serve the particle j, and the service time s is a random variable distributed as F and independent of anything else. At the end of the service time the particle j jumps to one of the N servers (may be, to the same one) with equal probability $1/N$ and waits for its turn to be served there. This cycle is repeated infinitely many times for each particle.

Under minimal assumptions each network $\mathcal{S}_{N,M}$ has a unique equilibrium process $\mathcal{E}_{N,M}$ that is a universal attractor. Our goal is to prove the convergence of equilibria $\mathcal{E}_{N,M}$ of the Markov processes on $\mathcal{S}_{N,M}$ to a unique equilibrium $\mathcal{E}$ of a so-called nonlinear Markov process (NLMP) $\mathcal{S}$ as $N \to \infty$ and $M/N \to H$. Note that $\mathcal{E}$ is a single point in the joint state space X of all our processes whence $\mathcal{E}_{N,M}$ is a probability measure on X for each finite N, M.

The NLMP is, in words, the process on the limit network with $N = \infty$ and $M = HN$. Its behavior is in some aspects simpler than that of finite networks. On bounded time intervals, the behavior of $\mathcal{S}_{N,M}$ converges to that of $\mathcal{S}$ as $N \to \infty$ and $M/N \to H$. The convergence of equilibria, nevertheless, is a much harder issue which is the subject of this paper.

We assume $\mathbb{E}s = 1$ and $\mathbb{E}s^2 < \infty$, where the second assumption is necessary for the existence of a nontrivial limit process as $N \to \infty$ and $M/N \to H > 0$. Otherwise it is not hard to see that we get a small number of very long queues for N large and no process at all in the limit.

The main feature of the networks in consideration is symmetry both in N and in M, that is, the system is invariant to all permutations of nodes and of particles. This makes the system a very particular case of a *Jackson-type network*, see [?]. Namely, this is a single-class queueing network with the FIFO service discipline and a general distribution of service time. The specifics of our model, namely, its mean-field nature lies in an especially simple structure of the routing matrix: all the entries of the $N \times N$-matrix P_N are equal to $1/N$.

The stochastic dominance technique had been introduced originally by A. Stolyar [?] for the deterministic service time. In [?] it was extended to a restricted class of general service times. Here we further extended the results to the case of a general service time distribution with the only restriction of finite second moment.

This goal is achieved by means of a new state space that comprises the lengths of queues at

N nodes and the remaining sojourn times of M particles. Similar methods can hopefully be used for the analysis of other mean-field models.

2 Models and parameters

There are different ways to formalize the evolution of $\mathcal{S}_{N,M}$ as a Markov process, that is, to define a state space and a process generator on this space. We begin with a more natural formalization where the current state of the network coincides with the collection of states of its N queues. Then we will give another formalization in terms of N queue heights and M remaining sojourn times. The latter one is more convenient for the proof of our main result though both formalizations describe the same dynamics of $\mathcal{S}_{N,M}$.

Let us begin with a definition of a state space Q of a single FIFO queue. Suppose we know the exogenous inflow of customers to the queue in the future ($t \geq 0$) as a point process of arrival times. Then, in order to know the stochastic queueing process at this server in the future, it suffices to know additionally the current elapsed time of service of the oldest customer in the queue plus the current number of the remaining customers that are currently waiting their turn to be served. Then we may calculate the distribution of future evolution scenarios of the queue.

However, it would be more convenient for several reasons to represent a queue in a more regular way as a finite sequence of *remaining service times* $h_{i,j}$ of all the customers in the queue, in order of their arrivals or in order of their prospective services, which is the same due to the FIFO service discipline. The values of $h_{i,j}$ are not observable, of course, but the dynamics of the process in terms of $h_{i,j}$ has more transparent description than in terms of elapsed service time of the first customer in the queue and the total number of customers in the queue.

Now, the current state of the queue i can be written as

$$q_i(t) = (h_{i,1}, \dots, h_{i,k}),$$

where second indices from 1 up to k mark the order of service of k customers in the queue. Namely, the customer 1 is currently served and the customer k is the last one to be served among the current customers. If new customers arrive, they are, of course, served after the customer k, in the order of their arrival.

If $k = 0$ then the queue is empty. Otherwise the value of $h_{i,1}$ decreases at rate 1 (the first customer is currently served) and all the other values $h_{i,j}$ do not change for a while.

Two kinds of event at the queue i are possible that make the behavior of the state of queue discontinuous. First (exit), as the value of $h_{i,1}$ hits zero at time t', the first customer is released and the length of the queue drops from k to $k-1$. We write then

$$h_{i,1}(t') = h_{i,2}(t'-1), \dots, h_{i,k-1}(t') = h_{i,k}(t'-1).$$

The released customer arrives immediately to one of the N nodes with equal probability $1/N$ and occupies the last position in the queue at this node. Note that with probability $1/N$ this customer returns to the queue i.

Second (arrival), if a new customer arrives to the queue i at time t'', the length of the queue rises from k to $k+1$ and we write

$$h_{i,k+1}(t'') = s,$$

where s is a random service time distributed as F and independent of anything else. As was mentioned, the events of these two kinds happen simultaneously if the released customer returns to the same queue.

The state space of a single queue is, therefore, the union of a countable number of finite-dimensional orthants:

$$Q = \mathbb{R}_+^* = \emptyset \cup \mathbb{R}_+ \cup \mathbb{R}_+^2 \cup \dots.$$

The state space of the whole network is the product of N copies of Q, that is, the vector space Q^N. However, due to the symmetry of the network, the order of vector components in Q^N is immaterial, that is, the dynamics of the system is invariant to permutations of servers. Therefore, we may consider a current configuration of the network as an atomic measure on the space Q,

where each atom has weight $1/N$ and corresponds to one of N single queues. Clearly, Q is a Polish space in the induced topology.

Denote

$$h_i = \sum_{j=1}^{k} h_{i,k},$$

that is, $h_i(t)$ is the total remaining service time of all the customers in the queue. We say that $h_i(t)$ is the current *queue height*.

The dynamics of a single queue height within the network consists of the deterministic (decreasing) part given by the differential equation with discontinuous right-hand side

$$\dot{h}(t) = \begin{cases} -1 & \text{if} \quad h(t) > 0, \\ 0 & \text{if} \quad h(t) = 0. \end{cases}$$

and the stochastic (increasing) part that consist of instant bursts $h(t) = h(t-0) + s$ with random i.i.d. increments s that happen at random times of arrival of new particles.

3 Alternative description

In what follows we will also use another description of the state space in terms of, again, the queue heights h_i at the N nodes and, additionally, the *remaining sojourn times* g_j of the M particles. The remaining sojourn time of the particle j is defined as the remaining time till the exit of customer j from its current queue i, that is,

$$g_j = \sum_{m \le k} h_{i,m},$$

where k is the current position of particle j in the queue i. Note that, because of the FIFO discipline, the value of g_j decreases at rate 1 until the service of particle j at node i is completed and does not change as new particles arrive to this node.

Thus the current state of the process $\mathcal{S}_{N,M}$ is an $(N+M)$-vector $f = (h, g)$ with non-negative components h_i and g_j. Because of the symmetry, we can reduce the state of the process to a couple of atomic measures μ and ν, both on $\mathbb{R}_+$. Namely, μ has N atoms of equal weight $1/N$ and ν has M atoms of equal weight $1/M$. Note that the value of g_j does not associate uniquely the particle j with some server i apart from special cases where, for instance, $g_j = h_i$ and there is no other $h_{i'} = h_i$ and no other $g_{j'} = g_j$.

Of course, the pair of measures μ and ν cannot be arbitrary pair of atomic measures, that is, μ and ν should be consistent. For instance, the upper customer in each queue has the remaining sojourn time equal to the height of the queue. Note, moreover, that the information contained in measures μ and ν is not sufficient to reconstruct the distribution of remaining service times among the customers nor the distribution of population among the queues. Let us regard the following example.

Let $N = 2$, $M = 4$. Let us arrange the components of measures μ and ν in ascending order. Let $h_1 = h_2 = 3$ and let $g_1 = 1$, $g_2 = 2$, and $g_3 = g_4 = 3$. Then either both queues hold two customers and their remaining service times are 1 and 2 but in different order, or one queue holds a single customer with remaining service time 3 and the other queue holds three customers with remaining service time 1 for each customer.

However, the continuous-time Markov process on measures μ and ν is well defined as we will see immediately. Moreover, if we watch the process for some finite time, we get all the information on the distribution of lengths of queues (number of customers in the queue) and on the distribution of remaining service times.

Let us see how the (continuous-time) Markov process on pairs (μ, ν) evolves. Recall that we consider the system $\mathcal{S}_{N,M}$. All the values $h_i(t)$ and $g_j(t)$ are decreasing at rate 1 as long as they are positive. The service events in the network happen exactly as some $g_j(t)$ vanishes. Suppose this is kth service of particle j. Then we denote this time by t_k^j.

Instantly, the particle j is routed to a random queue. Denote the height of its new queue by $h \ge 0$. The particle is allotted an F-distributed service time s there. The values of h and s are

random and independent. The distribution of h is given by the current value of N-atomic measure μ. The values of $h_i = h$ and of g_j both are updated to $h + s$

Then the next jump time for particle j is

$$t_{k+1}^j = t_k^j + h + s$$

and the evolution of $g_j(t)$ and of $h_i(t)$ goes on deterministically on the time interval $(t_k^j, t_{k+1}^j]$ as usual:

$$\dot{g}_j(t) = -1, \quad \dot{h}_i(t) = -1$$

The linear decrease of $g_j(t)$ is completely deterministic while the evolution of $h_i(t)$ may have positive bursts if new particles arrive.

We denote the space of pairs of atomic measures $\mu = \mu_N$ and $\nu = \nu_M$ by $X_{N,M}$. We embed all $X_{N,M}$ into the space X of all pairs of probability measures on $\mathbb{R}_+$. As we will see soon, X is the configuration space of the NLMP. For the NLMP (that is, for $N = \infty$), we have a limit dynamics of general probability measures $\mu(t)$ and $\nu(t)$ on $\mathbb{R}_+$. Again, these measures should be consistent, see below.

Formally, a series of continuous-time Markov processes $\mathcal{S}_{N,M}$ is defined on X and it can be proved that their generators converge to that of the NLMP $\mathcal{S}$ ensuring the convergence of processes on bounded time intervals. We will, however, use other tools for justification of this fact, that is, stochastic dominance methods. We will also see that the NLMP conserves single-point measures (it is a deterministic dynamical system on X) while finite processes $\mathcal{S}_{N,M}$, obviously, do not.

4 The NLMP

As we have mentioned above, the NLMP $\mathcal{S}$ is the limit process for $\mathcal{S}_{N,M}$ as $N \to \infty$ and $M/N \to H$. Formally we will define two limit dynamical systems for each form of Markov process (with two different state spaces) that were used for the description of the evolution of finite systems $\mathcal{S}_{N,M}$ and demonstrate that they are equivalent, that is, they describe the evolution of the same limit process $\mathcal{S}$ (the NLMP).

To begin with, let us use the more intuitive state space based on Q for the primary description of the NLMP. The current state of the process is now a probability measure η on Q. Clearly, the weak limit points of atomic measures on Q cover the space $\mathcal{M}(Q)$ of all probability measures on Q, hence, η is an arbitrary point of $\mathcal{M}(Q)$.

In order to get an intuitive notion of the NLMP, one may imagine the situation where there are infinitely many queues in the system, that is, $N = \infty$ and the distribution of states of these queues is η. Note, however, that we cannot formally define a mean-field routing process on a countable number of servers since there is no uniformly distributed probability measure on such a set. One may think of the queues of the NLMP as of elements of a continuous measurable space, say, of the interval [0,1] with Borel measure but, in fact, we do not need such a specialization.

The non-linear Markov process goes on as follows. Let $\eta(0)$ be given. Let us distinguish a single queue $q = q_\omega$ which is in the state $q(0)$ at $t = 0$. Its evolution within the NLMP is a stochastic process which can be described completely if we know the evolution of the measure $\eta(t)$ for $t \geq 0$. The evolution of different queues are independent.

It would be easier to begin with a definition of the NLMP for a given time-dependent Poisson inflow $\lambda(t)$, that is, without feedback. Let us assume that the inflows to all the nodes are independent Poisson flows of rate $\lambda(t)$ (all the inflows in the closed NLMP are Poisson ones because they have infinitely many additive sources). Then we will have a deterministic dynamics of $\eta(t)$ since the evolution of each particular queue is a Markov process and these processes are independent for different queues.

In turn, if we know $\eta(t)$ for $t \geq 0$, we can find the resulting mean outflow rate

$$b(t) = \lim_{\Delta \to 0} \frac{P(\Delta)}{\Delta},$$

where $P(\Delta)$ is the probability of service event at a random queue during the time interval $[0,\Delta)$. The value of $b(t)$ must be equal to $\lambda(t)$ for all t since the network is closed.

Definition 4.1 *The pair $(\lambda(t),\eta(t))$ is a solution of the NLMP if $\lambda(t)$ is the outflow generated by $\eta(t)$ and $\eta(t)$ is the evolution of the measure on queue states generated by the inflow of rate $\lambda(t)$.*

Theorem 4.2 *For each η_0 there exists a unique solution $(\lambda(t),\eta(t))$ of the NLMP such that $\eta(0)=\eta_0$.*

Proof. We are going to construct a series of processes $\mathcal{P}_n$, $n=0,1,...$, whose solutions converge to the required solution of $\mathcal{S}$. In the process $\mathcal{P}_0$, there is no feedback, that is, the customers that are served do not return to the system. The evolution of the corresponding measure $\eta_0(t)$ is simple: each queue drains out and it is empty forever since time $t=h(0)$.

The resulting outflow normalized by the number of queues has rate $\lambda_0(t)$, $t\geq 0$. Let us now construct the next process $\mathcal{P}_1$ as follows: we only allow particles to return to the system once. Formally, we consider a random queue that is distributed as $\eta(0)$ initially and then receives the inflow of rate $\lambda_0(t)$ (these inflows to different queues are mutually independent). Clearly, it becomes empty eventually with probability 1. We denote the corresponding outflow rate by $\lambda_1(t)$.

Now, we make an important remark: under an appropriate coupling, all the exit events in the process $\mathcal{P}_0$ happen at the same queues at the same time in the process $\mathcal{P}_1$ (because of the FIFO service discipline). Additionally, there are secondary exit events as the secondary particles that have returned to the system after the first service are served the second time in their lives and leave the system forever. Hence, $\lambda_1(\cdot)\geq\lambda_0(\cdot)$ in the following sense:

$$\int_0^t \lambda_1(s)ds \geq \int_0^t \lambda_0(s)ds \quad \text{forall} \quad t\geq 0.$$

Next we recall a simple monotonicity property of a FIFO server.

Lemma 4.3 *Suppose we have two identical FIFO servers 1 and 2 with the same initial states at $t=0$. Let u_1 and u_2 be point processes on $\mathbb{R}_+$ and let u_1 dominates u_2 stochastically (denoted $u_1 \pm u_2$). Denote by w_1 and w_2 the departure point processes of servers 1 and 2, respectively, where the server i receives the inflow u_i, $i=1,2$. Then $w_1 \pm w_2$.*

Recall that u_1 dominates u_2 stochastically if there is a coupling between the two processes such that $t_1^k \geq t_2^k$ for all coupled pairs of configurations $(t_i^1, t_i^2, \dots)$ and all $k=1,2,....$

Then we iterate the construction of λ_n for $n=2,3,....$ From Lemma 4.3, we get inequalities

$$\lambda_n(\cdot)\geq\lambda_{n-1}(\cdot) \quad \text{forall} \quad n=1,2,....$$

On the other hand, there is a finite upper bound $\bar{\lambda}(t)$ for all $\lambda_n(t)$ in the integral sense. Indeed, the maximum of service rate at each queue is 1. Hence there is a convergence and the limit is a solution $(\lambda(t),\eta(t))$ of $\mathcal{S}$.

Suppose there exists another solution $(\lambda'(t),\eta'(t))$. Then, by construction, $\lambda'(t)\geq\lambda(t)$ for all t and the inequality is strict for some $t<\infty$. We come to a contradiction easily since the mean mass of queues at this time t must be different for the inflows $\lambda(\cdot)$ and $\lambda'(\cdot)$.

Moreover, we can construct Markov processes on finite networks $\mathcal{S}_{N,M}$ by the same monotone iteration procedure as in $\mathcal{S}$. As a result, we also conclude that a finite-time convergence of $\mathcal{S}_{N,M}$ to $\mathcal{S}$ takes place.

Namely, it is not hard to prove that the dynamics of $\mu_N(t)$ and $\nu_M(t)$ in the process $\mathcal{S}_{N,M}$ is close to that of $\mu(t)$ and $\nu(t)$ in the NLMP on finite time intervals if N and M are large and if respective initial values of measures are close to each other. Analogous results can be found, for instance, in [?] under stronger assumptions on the service time distribution F.

Theorem 4.4 *Suppose that the sequence of probability measures $\varphi_{N,M}(0)$ on $X_{N,M}\subseteq X$ converges weakly to a probability measure $\varphi(0)$ on X as $N\to\infty$ and $M/N\to H$. Then the solutions of $\mathcal{S}_{N,M}$ from the initial states $\varphi_{N,M}(0)$ converge to the solution of the NLMP $\mathcal{S}$ from the initial state $\varphi(0)$ on any bounded time interval $[0,T]$.*

Here we do not prove the theorem and do not specify the notion of convergence of processes since these are rather technical issues.

In the same framework as for the finite systems, we may give a description of the NLMP as evolution of two probability measures on $\mathbb{R}_+$. They are the measure $\mu(t)$ on queue heights and the measure $\nu(t)$ on remaining sojourn times. Clearly, the resulting flow rate $\lambda(t)$ in the (h,g)-representation is the same as in (λ,η)-representation.

Again, there exist some constraints on possible pairs (μ,ν) caused by the fact that the remaining sojourn time of a particle at the top of the queue coincides with the height of this queue. Let us write down this constraint explicitly.

Denote by $\alpha=\alpha(\mu)$ the fraction of empty queues, that is, $\alpha=\mu(\{0\})$. Denote by $\grave{u}=\grave{u}(\mu)$ the measure $(1-\alpha)\frac{N}{M}\mu$ which is not a probability measure, of course. The constraint on the measure ν can be then written as

$$\nu\geq \grave{u},\quad \text{thatis},\quad \nu([a,b])\geq \grave{u}([a,b])\quad \text{whenever}\quad 0\leq a\leq b<\infty.$$

The dynamics of the pair $(\mu(t),\nu(t))$ for $t\geq 0$ is completely defined by the value $(\mu(0),\nu(0))$ as follows. Recall that the mean number of particles per node in the NLMP is exactly H. By the current value of $\nu(t)$ we can find the current service rate $\gamma(t)$ per particle:

$$\gamma(t)=\lim_{Delta\to+0}\frac{\nu[0,\Delta)}{\Delta}.$$

The current service rate per node is then equal to $\lambda(t)=\gamma(t)/H$. Note that the parameter H is already included in the constitutive relations for the process. As soon as we know $\lambda(t)$ and the current distribution of queue heights, we can write evolution equations for $\mu(t)$ and $\nu(t)$.

A rigorous way to prove the existence and uniqueness of a solution from any pair $(\mu(0),\nu(0))$ is, again, to realize a recursive construction where we allow 1, 2,, k lumps to each particle. The same method demonstrates the equivalence of two forms of the NLMP.

Note that an essential difference exists with the dynamics of a pair (μ_N,ν_M) in the finite network $\mathcal{S}_{N,M}$. Namely, the evolution of (μ_N,ν_M) is stochastic for each pair N,M and the evolution of $(\mu(t),\nu(t))$ is deterministic.

Now let us lift the consistency restrictions for the pair (μ,ν). We will define a solution of the NLMP that starts from an arbitrary pair $\mu,\nu\in\mathbb{R}_+$ in the same manner as before. Actually, such a solution has the following physical sense.

We do not suppose any longer that the system contains H particles per node straightaway. Instead we assume that these particles are elsewhere at $t=0$ and that, initially, each particle has its delay g distributed as ν and each queue has its height h distributed as μ. Then the process is started. The queue heights decrease at rate 1 while positive and the particles enter the system at times g.

Then the process goes on exactly by the rules of closed system, that is, the particles, as they jump, are given the new value of g equal to $h+s$ and the value of h is replaced by $h+s$ as well. Clear enough, the dynamics of such a system approaches that of the closed system as the number of particles per node in the system approaches H and the remaining initial height of queues vanishes.

5 Stationary solutions and ergodicity

Note that continuous-time Markov processes $\mathcal{S}_{N,M}$ are ergodic (if F is a non-lattice distribution) since there is a renewal event where all the particles get in the same queue and the oldest one begins its service, see [?, ?]. This event obviously happens with a positive frequency.

In the case of a lattice distribution F, we may consider a discrete-time process and repeat the argument. The discrete-time version of the process is, again, ergodic. For definiteness, in what follows we assume that F is a non-lattice distribution.

Our goal is to study the asymptotic behavior of unique equilibrium solutions $\mathcal{E}_{N,M}$ as the size N tends to infinity and as $M/N\to H$ for some $H>0$ (mean population of a node or, in other

words, mean length of a queue). Namely, we ask if the sequence $\mathcal{E}_{N,M}$ converges to the unique equilibrium $\mathcal{E}$ of the NLMP with the same parameter H. The two parameters of this problem is the service time distribution F and the mean length H of a queue which is time-independent since the networks are closed.

The process $\mathcal{E}_{N,M}$ can be written in (h,g)-notation. Then the unique equilibrium distribution $\varphi_{N,M}$ is defined on the configuration space X_N. Since all X_N are embedded into

$$X = \mathcal{M}(\mathbb{R}_+) \times \mathcal{M}(\mathbb{R}_+),$$

we write $\varphi_{N,M} \in \mathcal{M}(X)$.

The convergence of equilibria $\mathcal{E}_{N,M}$ to $\mathcal{E}$ can be understood in different senses. In particular, one may expect that the stationary measures $\varphi_{N,M}$ converge weakly to the δ-measure on a unique stationary point (μ^*, ν^*) of $\mathcal{S}$ as $N \to \infty$ (in (h,g)-notation). The rest of the paper is concerned with the proof of this fact.

Note that, in general, there is no explicit information on the invariant value of H in a solution $(\mu(\cdot), \nu(\cdot))$ of the NLMP. In equilibrium, however, it is not hard to calculate H from the stationary measure (μ^*, ν^*) as follows. Denote the jump rate per node by λ. Then the jump rate per customer is λ/H. Next, we compare the rates of continuous decrease of h and g. The first one is equal to $1-\alpha$ and the second one is 1.

Since we consider an equilibrium, these rates must be equal to the the rates of increase of h and g at jump events. They are, respectively, λ (since the mean service time is 1) and $\lambda(\mathbb{E}h+1)/H$ (since the mean value of new g after the jump equals the mean height of a queue plus the mean service time). Therefore,

$$H = (1-\alpha)(\mathbb{E}h + 1).$$

Lemma 5.1 *The NLMP has a unique equilibrium and all other solutions approach this equilibrium.*

Proof. For the proof we will use a fundamental result on "smoothing effect" of the FIFO server. It was proved under some additional restrictions on F in [?] and will be proved in the general case in a subsequent paper by the author.

The main idea is the monotonicity argument. Suppose there is no convergence and come to a contradiction. Indeed, in this case the mean population of a queue cannot be constant.

In more details, suppose there is a non-converging solution of the NLMP. Then $\lambda(t)$ does not converge as $t \to \infty$. By limit transition we can construct a non-constant solution $\lambda'(t)$ on the whole time axis $\mathbb{R}$ such that

$$\int_0^1 \lambda'(s)ds \geq \int_t^{t+1} \lambda(s)ds, \quad t \in \mathbb{R}.$$

By monotonicity, there is a coupling between the inflows $\lambda'(t)$ and $\lambda''(t) = \lambda'(t-1)$ on $(-\infty, 0]$ such that all the particles arrive not earlier in the first case and some of them arrive strictly later. Then the mean mass of a queue at $t=0$ is strictly larger than that at $t=-1$, which is a contradiction.

Now, since there is a unique equilibrium measure on solutions of the NLMP (actually, this is a δ-measure on the unique equilibrium solution), it suffices to prove tightness of the family (μ_N^*, ν_M^*), $N = 1,2,\ldots$ in order to derive convergence.

There exists a well-known criterion of tightness for random measures on $Y = \mathcal{M}(X)$, that is, on the space of probability measures on a Polish space X, where the topology of weak convergence of measures generates the measurable structure on Y. Recall that X itself is the space of pairs of probability measures on $\mathbb{R}_+$.

Proposition 5.2 *The family (μ_N^*, ν_M^*) is tight if and only if two following conditions hold. The probability of $h_i^N > K$ tends to zero as $K \to \infty$ uniformly on N. The probability of $g_j^M > K$ tends to zero as $K \to \infty$ uniformly on M.*

We will see that it suffices to get a "stable" upper bound on the inflow that does not

depend on the parameters in order to prove the required tightness.

6 Dominance

We use Baccelli–Foss theorems [?] for Jackson-type networks with implications for closed networks. The following theorem is an extension of Lemma 4.3 from a single FIFO node to a finite open Jackson-type network.

Theorem 6.1 *In a finite open FIFO network, if all the arrivals happen earlier and all the services happen faster, then all the events happen earlier.*

Here the notions "earlier" and "faster" should be understood in the sense of stochastic dominance. For closed networks, we may assume that the particles arrive once from the outside and then circulate within the network.

Theorem 6.2 *Suppose there is an infinite program at each node of the network G, that is, an infinite sequence of pairs (x_k, i_k), where x_k is the service time of kth particle and i_k is its address after the service. Suppose there is another instance G' of the same network with a program (x_k', i_k') that dominates the first one in the following sense: $x_k' \geq x_k$ and $i_k' = i_k$ for each k and each node of the network. Suppose the initial positions of particles coincide in both cases. Then all the events in G' happen not earlier than their counterparts in G.*

In particular, if the programs are identical but the exogenous particles arrive later to the same nodes in the second case, then, again, all the events in the primed case happen later. This assertion can be reduced to Theorem 6.2 by introduction of additional virtual nodes where the exogenous particles reside initially.

Theorem 6.2 can be proved by the following argument. Suppose the contrary. We make a coupling that preserves the order of initial events. Then there exists the first pair of events with the reverse order. And this, clearly, cannot happen.

Our next goal is to find a universal stochastic upper bound on the number of arrivals to a single node of a finite network $\mathcal{S}_{N,M}$ during the time interval $[0,T]$, that is, a bound that does not depend on N (if N is lrge enough) and on the initial state of the network. Suppose this upper bound B satisfies the *stability condition* $\mathbb{E}B < T$. Then the required tightness would follow.

Let us study processes $\mathcal{S}_{N,M}$ and $\mathcal{S}$ on a bounded time interval $[0,T]$ and count the number of services per node that happen within this interval. By coupling and monotonicity, there is the "worst" initial configuration that entails more services than any other initial configuration (either in the sense of stochastic dominance or for any fixed sequence of service times at each node). This is the zero configuration where $h_i = 0$ and $g_j = 0$ for all i and j. In words, all the M particles are jumping immediately (at $t = 0$). Then the process goes on by standard rules.

For given N and M and for the "worst" initial distribution, there is a distribution $B_{N,M,T}$ of the number of arrivals to a given node i during $[0,T]$. This distribution does not depend on i but if we assume the sequence of service times at node i to be known, then the number of arrivals to i has a different distribution, that is, there is a correlation between the service times at node i and the number of arrivals to i.

In order to cope with this inconvenience we consider the "fastest" program at a given node i, that is we assume that all the particles at this node are served immediately. Then the new distribution $B'_{N,M,T}$ of the number of arrivals certainly dominates the distribution $B_{N,M,T}$ since it dominates the distribution $B''_{N,M,T}$ for any other service time sequence at the node i.

Now, in order to find a uniform "stable" upper bound on the inflow to a given node, we need to find such T that distributions $B'_{N,M,T}$ are dominated by some B_T for all relevant pairs N, M (if N is large enough) and that B_T is "T-stable" (its mean total service time is strictly less than T).

Note that the queueing process with a special node i whose service time is always zero coincides in certain sense with the process $\mathcal{S}_{N-1,M}$. Namely, on the nodes different from i, the

process does not differ from $\mathcal{S}_{N-1,M}$. As $N \to \infty$ and $M/N \to H$, the difference between $\mathcal{S}_{N-1,M}$ and $\mathcal{S}_{N,M}$ vanishes. Hence, we will use upper bounds on $B_{N,M,T}$ instead of $B'_{N,M,T}$ and prove that they are all uniformly T-stable for some $T > 0$ and all N large enough.

Then we break the time half-axis $\mathbb{R}_+$ in intervals of length T and study the discrete-time process that dominates all equilibria $\mathcal{E}_{N,M}$ in certain sense. First of all, we look at the NLMP and the corresponding arrival process.

Theorem 6.3 *For the NLMP $\mathcal{S}$, there is a $T > 0$ and $\varepsilon > 0$ such that the mean number of arrivals to a node from any initial state is less than $T - \varepsilon$.*

Proof. We omit detailed proof and give just a bare idea. The mean number of arrivals to a node on the time interval $[0, T]$ is

$$\Lambda(t) = \int_0^T \lambda(t)dt.$$

Let us fix a $\lambda^* < 1$ such that the stationary Poisson arrival of rate λ^* leads to the stationary mean length $H' > H$ of a single queue with service time distribution F. Then, eventually, we have

$$\Lambda(t) < t\lambda^*.$$

Otherwise, by monotonicity argument, we can find $t > 0$ such that the mean length of the queue under the Poisson inflow with rate $\lambda(t)$ exceeds H. This proves the lemma.

Denote by B_T^* the corresponding distribution of arrival events at a given node. By approximation argument, any finite part of this distribution is close to the corresponding part of $B_{N,M,T}$. Then we handle the remaining part by dominance argument again and use some combinatorics to handle the tails.

7 Uniform upper bound for the inflow

The uniform dominance for all $N \geq \overline{N}$ follows from the finite-time convergence Theorem 4.4 and the NLMP stability Theorem 6.3. For the proof, let us prove the dominance separately for medium flows and for large flows.

If the size of flows (number of particles in the flow) in consideration is bounded from above by the same constant D, then we deal with a compact part of the state space. The probability of a single server to receive an inflow of k particles within the time interval $[0, T]$ satisfies the relation

$$\lim_{N\to\infty} P_k^N = P_k,$$

which implies immediately the required dominance.

Next we assume that all the initial queues are infinite and prove the second-moment bounds on the inflows in $\mathcal{S}_{N,M}$, $N \geq N_0$. These bounds dominate the inflows from any initial distribution of M particles among N nodes.

We have N independent identically distributed integer-valued variables m_i^N (numbers of services at individual queues) whose mean values are $T + 1$ (to be sure) and variances are bounded. Their sum

$$M^N = \sum_{i=1}^N m_i^N$$

is then distributed uniformly among N queues, producing the number of arrival n_i^N.

The mean value of n_i^N for any i is equal to $T + 1$, and we are going to find upper bounds for the probabilities of large values of n_i^N. To this end let us first choose a special service time distribution F' that is stochastically dominated by F and then find upper bounds on the probabilities of k arrivals to a single node during the time interval $[0, T]$.

The distribution F' has two atoms, at 0 and at 1. The probability p of 1 is strictly positive. Now note that the probability of k arrivals from N infinite queues with service time distribution F' equals the probability of drawing k ones in a series of $[T + 1]N$ independent draws, where the chance to draw 1 equals p/N.

Next, we find the asymptotic bounds on the probability of massive arrival. We study a discrete-time random walk on the positive orthant $\mathbb{N}_+^2$. Namely, we start at the origin and make one of the three steps: (1,0) with probability u, (0,1) with probability v, and (1,1) with probability $w = 1 - u - v$.

The values of u, v, w depend on the parameter $N = 1,2, \ldots$. We also have a an integer constant $T > 0$ (the length of interval) and a real constant $\alpha > 0$ (probability of service time 1). The process is defined as follows.

Let us make a vertical step with probability α, otherwise we make a horizontal step. Each step is then marked with probability $1/N$, independently. We count the number of marked steps (axis x) and the number of vertical steps (axis y). As a result we draw a path on the lattice $\mathbb{Z}^2$ from (0,0) to infinity.

We are interested in the measure on paths that is induced by the uvw-rules of random walk. Our goal is to assess the slope of the path and the probability that this slope is below certain value as the path reaches some vertical bound. This is a model of the inflow at a given node of the mean field closed FIFO network.

The next step is to draw required upper bounds. Since we need an upper bound, we can use another scheme of path construction: let us replace marked vertical edges by marked horizontal ones. This will be performed for $N > 1$. Then we have the following probability of a horizontal step:

$$\alpha'_N = \frac{(N-1)\alpha+1}{N}, \quad 1 - \alpha'_N = \frac{(N-1)(1-\alpha)}{N}. \tag{7.1}$$

We look for an upper bound on $p(k,n)$, that is, on the probability that the path from the origin passes through the point (k,n).

The event of passing through (k,n) is equivalent to the following one. The first $k+n$ steps contain exactly n vertical and k horizontal steps. The number of paths from (0,0) to (k,n), hence, equals to

$$S(k,n) = \frac{(k+n)!}{k!n!}.$$

All these paths have the same probability $(\alpha_N')^n(1-\alpha_N')^k$, therefore

$$p(k,n) = \frac{(k+n)!}{k!n!}(\alpha_N')^n(1-\alpha_N')^k.$$

Here

$$\alpha_N' = \frac{\alpha(N-1)}{\alpha(N-1)+1}, \quad 1 - \alpha_N' = \frac{1}{\alpha(N-1)+1}$$

For simplicity we may assume $\alpha_N' = (N-1)/N$ and $1 - \alpha_N' = 1/N$ (this will not change the asymptotics). Next, we substitute $n = TN$ and get

$$p(k,TN) = \frac{(k+TN)!}{k!(TN)!}\left(\frac{N-1}{N}\right)^{TN}\left(\frac{1}{N}\right)^k.$$

Actually, we are interested in an upper bound on the sum

$$P(k,n) = \sum_{m=k}^{\infty} p(m,n)$$

or another sum

$$Q(k,n) = \sum_{m=0}^{k} p(m, k+n-m).$$

The second sum is more efficient since no two points on the diagonal

$$D_{k+n} = \{(a,b) \in \mathbb{Z}_+^2 : a + b = k + n\}$$

can lie on the same path from the origin, that is, each path hits D_{k+n} exactly once.

Our goal is to find an appropriate upper bound that is uniform for all $N \geq N_0$ for some finite N_0. Again, for simplicity, let $T = 1$ (for a while). Then we have

$$p(k,N) = \frac{(k+N)!}{k!N!}\left(\frac{N-1}{N}\right)^{N}\left(\frac{1}{N}\right)^k.$$

By means of the Stirling formula, we reduce the bound to

$$p(k,N) \simeq \frac{\sqrt{k+N}(k+N)^{k+N}}{\sqrt{kN}k^kN^{k+N}} \tag{7.2}$$

(up to a multiplicative constant). We need the maximum of (7.2) over $N \geq N_0$. For the main part of (7.2), let us write

$$\frac{(k+N)^{k+N}}{k^kN^{k+N}} = \frac{\left(1+\frac{k}{N}\right)^{k+N}}{k^k} =$$

$$= \left(1 + \frac{k}{N}\right)^N \left(\frac{1}{k} + \frac{1}{N}\right)^k \leq e^k \left(\frac{1}{k} + \frac{1}{N_0}\right)^k \leq \frac{e^k e^{N_0}}{N_0^k}.$$

For $N_0 > e$ this bound vanishes exponentially as $k \to \infty$. The required dominance follows.

References

1. F. Baccelli and S. Foss. Ergodicity of Jackson-type queueing networks. *Queueing Systems*, 17(1-2):5{72, 1994.
2. R. Durrett. *Probability: Theory and Examples*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press, 2010.
3. Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. With a prologue by Peter W. Glynn.
4. S. Pirogov, A. Rybko, S. Shlosman, and A. Vladimirov. Propagation of Chaos and Poisson Hypothesis. *ArXiv e-prints*, October 2016.
5. A.N. Rybko and S.B. Shlosman. Poisson hypothesis for information networks. *Moscow Math. J.*, 5:679{704, 2005.
6. A. L. Stolyar. The asymptotics of stationary distribution for a closed queueing system. (Russian). Problems Inform. Transmission, 25(4):321{333, 1990.
7. A. A. Vladimirov, A. N. Rybko, and S. B. Shlosman. The self-averaging propert