Induced idleness leads to deterministic heavy traffic limits for queue-based random-access algorithms

Eyal Castiel, Sem Borst, Laurent Miclo, Florian Simatos, Phil Whiting

Induced idleness leads to deterministic heavy traffic limits for queue-based random-access algorithms

Eyal Castiel*, Sem Borst, Laurent Miclo, Florian Simatos, and Phil Whiting

ISAE-SUPAERO
10 avenue Edouard Belin
31055 Toulouse
e-mail: eyal.castiel@isae-supaero.fr; florian.simatos@isae-supaero.fr

Eindhoven University of Technology
e-mail: s.c.borst@tue.nl

Toulouse School of Economics
e-mail: laurent.miclo@math.univ-toulouse.fr

Macquarie University
e-mail: philip.whiting@mq.edu.au

Abstract: We examine a queue-based random-access algorithm where activation and deactivation rates are adapted as functions of queue lengths. We establish its heavy traffic behavior on a complete interference graph, which turns out to be nonstandard in two respects: (1) the scaling depends on some parameter of the algorithm and is not the N/N^2 scaling usually found in functional central limit theorems; (2) the heavy traffic limit is deterministic. We discuss how this nonstandard behavior arises from the idleness induced by the distributed nature of the algorithm. In order to prove our main result, we develop a new method for obtaining a fully coupled stochastic averaging principle.

MSC 2010 subject classifications: Primary 60K25; secondary 60K35.
Keywords and phrases: CSMA algorithms, stochastic averaging principle, state space collapse.

Contents

*Authors are listed in alphabetical order except for the first one who is the main contributor.

1. Introduction

In the present paper we investigate the heavy traffic behavior of a queue-based random-access mechanism. Specifically, we analyze the joint queue length process in a critically loaded system where packets arrive at the various nodes as Poisson processes and are transmitted intermittently. When all nodes are inactive, any of them may start a packet transmission at an exponential rate that depends on its local queue length, i.e., the number of packets pending for transmission. Once a node is transmitting, it prevents other nodes from activating and turns inactive at an exponential rate which is also governed by its local queue length.

1.1. Context and motivation

The above model arises in the context of distributed scheduling, i.e., deciding which queues to serve without any central authority having a global knowledge of the network state, in queueing networks with constraints on the set of queues that can be active simultaneously (called constrained queueing networks). This is a fundamental and challenging problem with applications in a wide range of settings. In particular, the random-access mechanism described above captures the dynamics of queue-based versions of the Carrier-Sense Multiple-Access Collision Avoidance (CSMA-CA) protocol as further explained below, which is commonly used in wireless communication networks.

A breakthrough in the area of scheduling in constrained queueing networks was achieved when Tassiulas and Ephremides introduced the Max-Weight algorithm in the early nineties [TE90]. This scheme was the first to provably offer maximum stability guarantees under fairly general conditions, and has been generalized and refined in a huge body of follow-up work. However, the Max-Weight algorithm is inherently centralized in nature, and crucially relies on the solution of a potentially NP-hard global optimization problem at each iteration, namely, finding an independent set of maximum weight which then serves as schedule. This severely limits its implementation in large-scale networks.

Only after nearly twenty further years, Jiang and Walrand [JW08] and Rajagopolan *et al.* [RSS09] proposed the first truly distributed algorithms with the capability to match the throughput optimality of the Max-Weight algorithm. Informally stated, these algorithms aim to mimic the scheduling operations of the Max-Weight algorithm while using only locally available information. Specifically, the individual nodes make fairly autonomous decisions for controlling activity periods (e.g. packet transmissions) and inactivity periods (e.g. back-off intervals), subject to the constraints on simultaneous activity.

For a more detailed description, it is convenient to assume that the latter constraints can be represented in terms of a conflict graph, where the edges indicate which pairs of nodes are prevented from simultaneous activity. Such constraints may for example arise from interference issues preventing simultaneous transmission in wireless networks, in which case the conflict graph is commonly

referred to as interference graph. The operations of distributed scheduling algorithms in these scenarios may be described as follows. Upon completion of a packet transmission, a node either starts a random back-off period or proceeds with the transmission of the next packet, if any, with a probability that depends on the local queue length. When inactive, a node simply runs down its back-off clock, but freezes it whenever any of its neigbhors in the conflict graph are active, ensuring that a back-off period can only end when all its neighbors are inactive. At that point, a node either initiates a packet transmission or proceeds with the next random back-off period with a probability which is also a function of the local queue length.

Now observe that transmission activity in wireless networks can be detected by 'sensing' a shared channel, and that the above back-off mechanism precludes concurrent transmissions of mutually interfering nodes, explaining the term Carrier-Sense Multiple-Access Collision Avoidance. Further note that the idleness and randomized deactivation may seem inefficient from a resource utilization perspective, but play an instrumental role in sharing the medium through 'listening' in the absence of any centralized access control mechanism. The extreme case where a node deactivates only when its queue is empty is referred to as the Random Capture algorithm [FPR10]. Although it may seem to minimize idleness, it was actually shown that it is not always throughput-optimal [GBW14].

When we now assume the interference graph to be complete and further suppose that the back-off periods and transmission times are all independent and exponentially distributed, the above model reduces to that described in the first paragraph: when all nodes are inactive, any of them may turn active at a queue-dependent exponential rate, and once a node is active, it de-activates at a queue-dependent exponential rate.

The seminal results in [SS12, SST11] showed that the above-described queue-based versions of the CSMA-CA algorithm (henceforth referred to as QB-CSMA) achieve maximum stability, provided that the activation and deactivation probabilities are governed by suitable functions of the local queue lengths. To the best of our knowledge, however, little is known about the queueing dynamics of these algorithms beyond the maximum stability properties.

1.2. A nonstandard heavy traffic behavior

The present paper aims at deepening the understanding of the above-described queue-based random-access algorithms. We prove in particular that, near criticality and for the particular case of a complete interference graph, these scheduling mechanisms exhibit a nonstandard behavior in two ways:

1. the heavy traffic limit is deterministic;
2. the scaling depends on a parameter of the algorithm and is not the usual central limit theorem (CLT) like scaling N/N^2 where the time-scale N^2 is the square of the space scale N.

In particular, the limit is not a reflected Brownian motion and is thus unconventional in the terminology of Harrison [Har95]. The literature on unconventional heavy traffic results is quite scarce, at least compared to the important number of conventional results. Harrison and Williams [HW96] exhibited the first such example in the context of a closed queueing network, and Kruk later provided an example for an open queueing network under the Earliest Deadline First policy [Kru11]. Atar and Cohen [AC19] study a multiclass single-server queue which, subject to the usual CLT scaling, converges to a nonstandard diffusion process (namely, a Walsh Brownian motion). Another example is Puha [Puh15], who studies the Shortest Remaining Processing Time (SRPT) policy: there the scaling is nonstandard but the limiting diffusion is conventional, i.e., the heavy traffic limit is a reflected Brownian motion.

In the model studied in the present paper, the behavior is nonstandard in two ways: (1) the limit is actually deterministic and governed by an ordinary differential equation (ODE), and (2) if N is the space scale, the suitable time scale is N^{1+a} with $a \in (0, 1/2)$ a parameter of the algorithm. In particular, the time scale is in-between the usual fluid and diffusion time scales N and N^2, respectively. This peculiar scaling is due to the idleness which arises as a consequence of the distributed nature of QB-CSMA, see Section 2.4.4 and 8.2 for an illustrative back-of-the-envelope computation.

Despite these two nonstandard features, our model does exhibit a state space collapse property which is commonly associated with the conventional Brownian diffusion limits for a wide range of multi-dimensional queueing processes [Rei84, Rei05]. Indeed, the queue lengths at the various nodes vary according to certain fixed proportions in the heavy traffic limit, meaning that the joint queue length process lives in a one-dimensional space.

1.3. Idleness in random-access settings

As alluded to above, in QB-CSMA, nodes deactivate at a state-dependent rate in order for the system to be able to alternate between different activity states in a distributed way. In particular, nodes may deactivate even when they have work to process. This makes the system non work-conserving and induces additional idleness compared to that owing to queues being empty.

For classical queueing models and stochastic networks, a large body of literature has investigated the impact of idleness on the heavy traffic behavior and performance. For instance, one of the achievements in the study of Jackson networks is the understanding of this impact on the reflection matrix in the limiting multi-dimensional reflected Brownian motion [HR81]. However, in these "classical" settings, idleness occurs when queues are empty or resources get stranded because of concurrency requirements.

In contrast, in random-access settings like ours, idleness occurs even when there are large queues, and is simply part of a distributed mechanism to share resources without explicit information exchange. In this distributed setting, the impact of idleness on heavy traffic behavior is more subtle and model-dependent.

For instance, considering QB-CSMA in a different regime than the one studied here, a lingering effect was highlighted in [SBB14] leading to a heavy traffic scaling $\frac{1}{(1-\rho)^2}$, with ρ the load of the network, compared to the usual $\frac{1}{1-\rho}$ due to idleness. In the present model, the fraction of idleness is inversely proportional to (a power of) the queue lengths, yielding a yet different impact on the heavy traffic behavior. After the model and main results are presented, we will describe this behavior in greater detail in Section 2.4.4, and in particular explain why the heavy traffic behavior is deterministic and the N/N^{1+a} scaling emerges.

It is interesting to compare our results with those on Max-Weight. Indeed, QB-CSMA algorithms were designed with the purpose of mimicking Max-Weight in a decentralized manner, and Shah and Shin [SS12] establish the throughput-optimality of these algorithms by applying the same Lyapunov function as for Max-Weight. Thus, as far as throughput is concerned, QB-CSMA algorithms behave very similarly as Max-Weight. What we show here is that the comparison breaks down at criticality concerning delay. Indeed, Stolyar [Sto04] showed that the critical behavior of Max-Weight is "standard", i.e., consists in the usual CLT scaling and leads to a reflected Brownian motion. Here the behavior is completely different because of the additional idleness induced by the decentralized nature of QB-CSMA.

1.4. Link with polling systems

When run on a complete interference graph, only one server can be active at a time and so QB-CSMA can be viewed as a particular polling system with state-dependent non-zero switchover times and switching decisions. This equivalence has in fact been exploited to use results for polling systems with a so-called 1-limited service discipline and a probabilistic routing policy in analyzing CSMA algorithms where nodes deactivate at a fixed (non-queue-based) rate, see for instance [CBvW16, DBBV15].

There is a significant body of heavy traffic results for polling systems by now, starting with the seminal papers [CPR95, CPR98]. However, the model in [CPR95] did not include any switchover times, so that the total amount of work behaves as in a work-conserving single-server queue and in particular exhibits the standard heavy traffic scaling behavior. The model in [CPR98] did incorporate non-zero switchover times, but involved an exhaustive service discipline, which implies that the fraction of idleness is basically reciprocal to the queue length, rather than the queue length raised to a power $a \in (0, 1/2)$. While the total amount of work is substantially larger than in a work-conserving single-server queue due to the non-zero switchover times, it exhibits a similar $\frac{1}{1-\rho}$ scaling behavior because of the rapid decay of the idleness as function of the queue length. Moreover, the exhaustive service discipline causes the work to rapidly shift among the various queues, causing fundamentally different dynamics than the state space collapse that we observe in our model.

Heavy traffic results for a broader class of polling systems with so-called Bernoulli-exhaustive and Bernoulli-gated service disciplines are established in [vdM07].

However, these concern stationary distributions rather than process-level limits, and again pertain to disciplines where the idleness scales inversely proportional to the queue length, yielding qualitatively similar scaling behavior as in [CPR98].

Finally, heavy traffic results for polling systems with k-limited service disciplines are presented in [BW14]. In these systems the idleness essentially approaches a constant, positive fraction as the queue lengths grow, again causing fundamentally different scaling behavior from what we encounter in our model.

1.5. Methodological contribution

The seemingly simple case of a complete interference graph actually turns out to be challenging to analyze. Technically, the main difficulty lies in controlling the so-called stochastic averaging principle, or homogenization. This principle asserts that when two processes interact but evolve on different time scales, then the 'slow' process only interacts with the 'fast' process through the instantaneous equilibrium distribution of the fast process. The most difficult case is the so-called fully coupled stochastic averaging principle which arises when this instantaneous distribution depends on the state of the slow process, which is the case here.

Controlling such an approximation is in general a difficult problem, and numerous methods have been developed for that, see for instance the classical monograph of Freidlin and Wentzell [FW84]. However, in our case we were not able to apply any standard method, in particular the ones developed by Kurtz [Kur92] and Luczak and Norris [LN13]. This led us to develop a new method. It is close in spirit to that of Luczak and Norris but is more tailored to Markov processes. The stochastic averaging principle is controlled by martingale arguments and leverages properties of solutions to the Poisson equations associated with the fast generators, see Section 2.4.3 for more details. We believe that this new approach has the potential of being applied to a wide class of problems and its more general applicability will be studied elsewhere.

To give a more precise idea of our technique to control the homogenization, imagine the Markov process under study is (Q^N, σ^N) with Q^N the 'slow' process and σ^N the 'fast' one. Controlling homogenization amounts to controlling the approximation

$$\int_0^t F\left(Q^N(s), \sigma^N(s)\right) \mathrm{d}s \approx \int_0^t \pi^{Q^N(s)}\left[F\left(Q^N(s), \cdot\right)\right] \mathrm{d}s$$

with π^q the stationary distribution of the fast process when the slow process is in state q and $\nu[f]$ the integral of a measurable function f with respect to the measure ν. To do so, we first rewrite the difference

$$\int_0^t \left(F\left(Q^N(s), \sigma^N(s)\right) - \pi^{Q^N(s)}\left[F\left(Q^N(s), \cdot\right)\right]\right) \mathrm{d}s$$

in the form

$$V(Q^N(t),\sigma^N(t)) - V(Q^N(0),\sigma^N(0)) \\ - \int_0^t L_{\mathrm{s}}^{N,\sigma^N(s)}(V(\,\cdot\,,\sigma^N(s)))(Q^N(s))\mathrm{d}s + (\text{martingale term})$$

with $L_{\mathrm{s}}^{N,\sigma}$ the generator of the slow process when the slow process is in state σ. Here the function V that appears is linked to solutions to the Poisson equation $L_{\mathrm{f}}^{N,q}\phi = g - \pi^q[g]$ with unknown ϕ, and so the above expression indeed makes it possible to cast the problem of homogenization in terms of control of solutions to Poisson equations. Moreover, this control is achieved by expressing solutions ϕ in the form

$$\phi(\sigma) = \int_0^\infty [\mathbb{E}_\sigma(g(X(t))) - \mathbb{E}_\sigma(g(X(\infty)))]\,\mathrm{d}t$$

with $(X(t))$ the fast Markov process started at σ under $\mathbb{P}_\sigma$, and $X(\infty)$ is stationary distribution. To the best of our knowledge, this approach for controlling homogenization, and in particular the bounds on the solutions to the Poisson equation that we establish are new.

1.6. Organization of the paper

We introduce our model and state our main result in Section 2. This section also presents a discussion of the result, in particular why we consider polynomial activation functions, a back-of-the-envelope computation to provide an intuition for the result, and also a more detailed discussion of the stochastic averaging principle. Section 3 gathers the notation used throughout the paper, and in particular the generators and their associated Poisson equations as well as important stopping times used in localization arguments. The three main steps of the proof are then presented. Sections 4 to 7 contain the technical arguments: Sections 4 and 5 describe the arguments controlling the stochastic averaging principle, Section 6 the arguments controlling the state space collapse, and Section 7 gathers the arguments to provide the full proof. The paper is concluded with different extensions and directions for future research in Section 8.

2. Model description and main result

2.1. Model description with fixed arrival rates

We have a set of n nodes labeled by $V = \{1,\dots,n\}$. Each node $v \in V$ represents an $M/M/1$ queue with the FIFO service discipline and vacations, its arrival rate is denoted by $\lambda_v > 0$. We denote by $Q_v(t) \in \mathbb{N} \coloneqq \{0,1,\dots,\}$ the length of v's backlog at time t and by $\sigma_v(t) \in \{0,1\}$ the activity process: the server at v is active and processing pending requests at unit rate if $\sigma_v(t) = 1$, and

$\sigma_v(t) = 0$ otherwise. Put differently, $\sigma_v(t)$ is the instantaneous service rate of node v at time t. We define $\lambda := (\lambda_v, v \in V)$, $Q(t) := (Q_v(t), v \in V)$ and $\sigma(t) := (\sigma_v(t), v \in V)$.

We impose that only one node can be active at a time, and so whenever convenient we will identify σ with the active node, or put $\sigma = 0$ if no node is active (empty schedule). We will thus either consider $\sigma \in \{0,1\}^V$ when seeing σ as the vector of instantaneous service rates, or $\sigma \in V_0$ with $V_0 = V \cup \{0\}$ when seeing σ as the current schedule. Because a schedule is associated with a node, we will sometimes use the notation q_σ to denote the vth coordinate of the vector $q \in \mathbb{R}_+^V$, with v the only non-zero coordinate of σ, and in this case we will adopt the convention $q_0 = 0$. Note that with this convention, we have $\sigma_0 = 1 - \sum_{v \in V} \sigma_v$.

Given the current schedule σ, the queue-length process Q evolves as n independent $M/M/1$ queues with service rates σ and input rates λ. On the other hand, given the current value Q of the queue-length process, σ evolves according to the following dynamic, which is a particular case of the Glauber dynamics for the hard-core model [Dob68, vdBS94]: an active node v with $\sigma_v = 1$ deactivates at rate $\Psi_-(Q_v)$ for some deactivation function Ψ_-, and an inactive node v with $\sigma_v = 0$ activates at rate $\Psi_+(Q_v)$ for some activation function Ψ_+, provided no other node is active.

To be more formal, (Q, σ) is a Markov process on $\mathbb{N}^V \times \{0,1\}^V$ with infinitesimal generator L that can be decomposed as the sum of two generators:

- the generator L_{s}^σ of the *slow* queue-length process Q whose dynamic depends on σ;
- and the generator L_{f}^q of the *fast* activity process σ whose dynamic depends on q.

The terminology *slow* and *fast* will be justified in Section 2.4.3 when discussing the stochastic averaging principle. Thus, L acts on functions $f : \mathbb{N}^V \times \{0,1\}^V \to \mathbb{R}$ as

$$Lf(\sigma, q) = L_{\mathrm{s}}^\sigma(f(\sigma, \cdot))(q) + L_{\mathrm{f}}^q(f(\cdot, q))(\sigma)$$

with

$$L_{\mathrm{s}}^\sigma(g)(q) = \sum_{v \in V} \lambda_v \left(g(q + e^v) - g(q)\right) + \sum_{v \in V} \sigma_v \mathbb{1}_{q_v > 0} \left(g(q - e^v) - g(q)\right) \quad (2.1)$$

and

$$L_{\mathrm{f}}^q(h)(\sigma) = \sum_{v \in V} \sigma_v \Psi_-(q_v) \left(h(\sigma - e^v) - h(\sigma)\right) + \prod_{w \in V} (1 - \sigma_w) \sum_{v \in V} \Psi_+(q_v) \left(h(\sigma + e^v) - h(\sigma)\right) \quad (2.2)$$

with $g : \mathbb{N}^V \to \mathbb{R}$ and $h : \{0,1\}^V \to \mathbb{R}$ arbitrary functions and $e^v \in \{0,1\}^V$ with 0's everywhere except at the vth coordinate equal to 1. Note that a server

does not deactivate immediately when its queue gets empty, which makes the indicator term $\mathbb{1}_{q_v>0}$ necessary in (2.1). Since the graph associated with L_{f}^q is a star centered at 0, this generator admits a reversible distribution denoted π^q. For reasons explained in Section 2.4, we consider polynomial activation and deactivation functions of the form

$$\Psi_+(x) = \frac{(x+1)^a}{1+(x+1)^a} \in [0,1] \quad \text{and} \quad \Psi_-(x) = 1 - \Psi_+(x), \ x \in \mathbb{N},$$

with $a > 0$ a parameter of the algorithm. In this case, π^q is given by

$$\pi^q(\sigma) = \frac{(1+q_\sigma)^a}{\sum_{\eta \in V_0}(1+q_\eta)^a}, \ \sigma \in V_0.$$

Let $\rho = \sum_{v \in V} \lambda_v$. Under the above assumptions, it is not hard to establish that (Q, σ) is positive recurrent if $\rho < 1$ and transient if $\rho > 1$. Transience for $\rho > 1$ can be proved by lower bounding $X := \sum_v Q_v$ by an $M/M/1$ queue with arrival rate ρ and service rate 1. Positive recurrence for $\rho < 1$ can be proved using the Foster–Lyapunov criterion and showing that X is a Lyapunov function. Indeed, as soon as one queue is active, arrivals make X increase in the mean by ρ while the queue in service makes it decrease by 1, so that overall X decreases at rate $\rho - 1 < 0$.

Thus, the regime where $\sum_v \lambda_v = 1$ will be referred to as the critical case and the rest of the paper will be devoted to the study of the near-critical case where $\sum_v \lambda_v \approx 1$, which we introduce now.

2.2. Near-critical regime and heavy traffic scaling

Throughout the paper, we fix $V = \{1, \dots, n\}$, $a > 0$, $\lambda^\infty \in \mathbb{R}_+^V$ with $\sum_v \lambda_v^\infty = 1$ and $\gamma \in \mathbb{R}^V$. For each $\varepsilon > 0$, we define $N = \varepsilon^{-1/a}$ and consider

$$\lambda^N = \lambda^\infty - \varepsilon\gamma = \lambda^\infty - N^{-a}\gamma. \tag{2.3}$$

The parameter $\varepsilon > 0$ represents the 'distance' between λ^N and the boundary of the stability region. We introduce $N = \varepsilon^{-1/a}$ because it will be simpler to index the processes by N rather than by ε, as is for instance reflected by the notation λ^N instead of λ^ε. As will be seen shortly, $N = \varepsilon^{-1/a}$ is the 'right' order of magnitude of the queue length process (see the discussion in Section 8.2 for more details).

Our main object of interest is the Markov process with infinitesimal generator given by (2.1) and (2.2) but with λ^N instead of λ. Thus, the generator L_{s}^σ that we will consider actually depends on N and is given by

$$L_{\mathrm{s}}^\sigma(g)(q) = \sum_{v \in V} \lambda_v^N \left(g(q + e^v) - g(q)\right) + \sum_{v \in V} \sigma_v \mathbb{1}_{q_v > 0} \left(g(q - e^v) - g(q)\right),$$

but in order to avoid cumbersome notation we will omit this dependency in N. Likewise, we will denote by L the generator $L = L_{\mathrm{f}}^q + L_{\mathrm{s}}^\sigma$ introduced above but

with λ^N instead of λ, and in the sequel we will denote by (Q, σ) the Markov process with this generator (again, omitting the dependency in N for ease of notation).

In contrast, we will keep the dependency in N for the scaled processes. More precisely, we consider (Q^N, σ^N) the Markov process obtained from (Q, σ) by speeding up time by a factor N^{1+a} and scaling the Q-components by N in space:

$$Q^N(t) = \frac{1}{N} Q\left(N^{a+1} t\right) \quad \text{and} \quad \sigma^N(t) = \sigma\left(N^{a+1} t\right), \ t \geq 0.$$

The infinitesimal generator of (Q^N, σ^N) will be denoted by L^N, see Section 3.1.2 for an explicit formula.

Remark 2.1. *Other scalings are possible: actually, when the arrival rates are still given by (2.3) and ε is the distance to the boundary of the stability region, we investigate in Section 8.2 what happens on the space scale $\varepsilon^{-1/a'}$ with $a' > 0$ not necessarily equal to a.*

2.3. Main result

For $x \in \mathbb{R}^V$ and $b > 0$, let in the sequel

$$\|x\|_b = \left(\sum_{v \in V} |x_v|^b\right)^{1/b} \quad \text{and} \quad s(x) = \sum_{v \in V} x_v.$$

As will be seen shortly, the limiting process lives in the one-dimensional vector space

$$\begin{aligned} I &= \left\{x \in \mathbb{R}_+^V : \lambda_w^\infty x_v^a = \lambda_v^\infty x_w^a, \ v, w \in V\right\} \\ &= \left\{x \in \mathbb{R}_+^V : x_v = \left(\frac{\lambda_v^\infty}{\mu}\right)^{1/a} s(x), \ v \in V\right\}, \end{aligned} \tag{2.4}$$

where here and in the sequel, $\mu = \|\lambda^\infty\|_{1/a}$. Intuitively, I is the space where the mean service rate at each node matches the corresponding arrival rate. In the sequel we use $\Rightarrow$ to denote weak convergence as $N \to \infty$. The following result is the main result of the paper, which describes the behavior of the queue-length process in the near-critical case.

Theorem 2.2. *Assume that the three following assumptions hold:*

- *$a < 1/2$;*
- *condition (2.3) holds, i.e., $\lambda^N = \lambda^\infty - N^{-a}\gamma$ with λ^∞ and γ introduced above;*
- *$Q^N(0) \Rightarrow q^0$ for some $q^0 \in I \setminus \{0\}$.*

Then $Q^N \Rightarrow q$ uniformly on compact time-sets, where q is uniquely characterized as follows: $q(t) \in I$ for every $t \geq 0$ and $s \circ q$ is the unique solution to the ODE

$$\dot{x} = \mu x^{-a} - s(\gamma)$$

with initial condition $x(0) = s(q^0)$ *and where* $\mu = \|\lambda^\infty\|_{1/a}$.

Except when $s(\gamma) = 0$, there does not seem to be an explicit formula for the solution of the previous ODE. For $s(\gamma) = 0$, the solution to the ODE $\dot{x} = \mu x^{-a}$ is

$$x(t) = \left(x(0)^{a+1} + (a+1)\mu t\right)^{1/(a+1)}, \ t \geq 0,$$

and so using the fact that $q(t) \in I$, the limit q in the previous statement is given in this case by

$$q_v(t) = \left(\frac{\lambda_v^\infty}{\mu}\right)^{1/a} \left(s(q^0)^{a+1} + (a+1)\mu t\right)^{1/(a+1)}, \ v \in V, t \geq 0.$$

From now on, we assume that the conditions of this theorem are enforced, i.e., we assume throughout that $a < 1/2$, that (2.3) holds and that $Q^N(0) \Rightarrow q^0 \in I \setminus \{0\}$.

With some extra work, but without giving much more insight on the system's behavior, the previous result could be generalized to an arbitrary initial condition $q^0 \in \mathbb{R}_+^V$. If $q^0 = 0$ nothing changes in the statement of the above result, while if $q^0 \notin I$ then the convergence holds uniformly on compact time-sets from $(0, +\infty)$ because the limiting process immediately jumps at time $0+$ to the invariant manifold I even if it does not start there. The rest of this introduction is devoted to discussing this result in more details.

2.4. Intuition and discussion

We discuss here in more details the context and implications of our result. We begin by justifying our interest in polynomial activation functions, then give an intuition behind the state space collapse result based on the stochastic averaging principle, and we finally discuss the nonstandard scaling that emerges from it.

2.4.1. Polynomial activation functions

The literature on optimal CSMA algorithms is very rich and the interested reader is for instance referred to the thorough survey by Yun et al. [YYSE12] for more details. In this paper we are interested in the class of QB-CSMA algorithms initially proposed by Rajagopalan, Shah and Shin [RSS09]. The main idea of these algorithms is to have activation and deactivation rates Ψ_+ and Ψ_- being adapted as a function of queue lengths. Rajagopalan, Shah and Shin study in particular the case where $\Psi_+ + \Psi_- = 1$ with

$$\Psi_+(q) = \frac{f(q_v)}{1 + f(q_v)}$$

for some function f. The main result of [GS10, RSS09, SS12] is that this algorithm is throughput-optimal for any interference graph provided f increases

slowly enough, namely sub-polynomially[1]. However, results of [GBW14] suggest that if f grows polynomially, then it is only throughput-optimal for some interference graphs, depending on the relation between the graph topology and the exponent of the polynomial growth of f.

The rationale for seeking fast-increasing functions f is that a folklore result has it that delay is improved with faster increasing functions f, an intuition which is backed up by results in [BBvL11]. Polynomial activation and deactivation functions should therefore achieve the optimal trade-off between throughput and delay for this class of algorithms, which is the reason why we focus on this case here. Note that in the case of a complete interference graph as considered here, the algorithm is throughput-optimal for any functions Ψ_+ and Ψ_- satisfying $\Psi_+(q) \to 1$ and $\Psi_-(q) \to 0$ as $q \to \infty$, so that we need not worry about stability issues for such polynomial activation and deactivation functions, as may be the case in a more general setting.

2.4.2. State space collapse from the stochastic averaging principle

The reason behind the state space collapse property is simple to understand based on the *stochastic averaging principle*. Put simply, when queue lengths are large, say of the order of N, then the typical time scale of σ is much faster than the one of Q which makes Q interact with σ only through the stationary distribution π^q of its corresponding instantaneous Glauber dynamics. The latter depends on Q, which gives rise to a so-called fully coupled stochastic averaging principle, which essentially amounts to the approximation

$$\int_0^t F\left(Q^N(s), \sigma^N(s)\right) \mathrm{d}s \approx \int_0^t \pi^{NQ^N(s)}\left[F\left(Q^N(s), \cdot\right)\right] \mathrm{d}s \tag{2.5}$$

with $\nu[f] = \int f \mathrm{d}\nu$ for any positive measure ν and integrable function f.

Recall that in our case, the stationary probability $\pi^q(v)$ of node $v \in V$ being active is given by

$$\pi^q(v) = \frac{(1+q_v)^a}{1+\sum_{w\in V}(1+q_w)^a}.$$

According to the stochastic averaging principle, this should represent the instantaneous service rate of node v which should thus behave as a subcritical $M/M/1$ queue when $\pi^q(v) < \lambda_v^\infty$ and as a supercritical $M/M/1$ queue when $\pi^q(v) > \lambda_v^\infty$. As $\sum_v \lambda_v^\infty = 1$ and $\sum_{v\in V} \pi^q(v) = 1 - \pi^q(0) \approx 1$ for large q, we see that the only way for the network to behave smoothly is that each average service rate $\pi^q(v)$ matches its incoming service rate, i.e., $\pi^q(v) \approx \lambda_v^\infty$. When q is large, this forces q to live in the invariant manifold I because $\pi^q(v) \approx q_v^a$ up to a multiplicative constant. Thus, the state space collapse phenomenon can be directly understood as a consequence of the stochastic averaging principle together with the criticality assumption.

[1] Actually, these algorithms also use some information on the current maximum queue length, whether the exact maximum or an estimation thereof.

2.4.3. The stochastic averaging principle

In the context of stochastic networks, the *stochastic averaging principle* was put forth for loss networks in the famous work by Hunt and Kurtz [HK94] but, as mentioned in Feuillet and Robert [FR14], "outside this class of networks, there are, up to now, few examples of stochastic networks for which a fully coupled stochastic averaging principle occurs". Establishing a fully coupled stochastic averaging principle is in general a challenging task and, in the queueing literature, many works actually restrict their study to the so-called homogenized process, assuming that timescale separation indeed occurs.

Rigorous proofs of stochastic averaging principles were established for polling systems times [CPR95, CPR98, Jen10], for models of distributed hash tables [FR14] and for the X model [PW13]. Luczak and Norris [LN13] also developed a new method which they applied to a variant of the supermarket model.

Most of these works, in particular [FR14, HK94, PW13], rely on the machinery developed by Kurtz [Kur92]. It relies on martingale arguments and identifies the asymptotic occupation measure of the fast process as the invariant measure of a limiting averaged generator. In our case this identification step is not clear because some rates go to 0 in the limit. In particular, the limiting scheduling process is degenerate: it starts at 0 and then jumps to one of the possible states $v \in V$ where it is absorbed. In the absence of uniqueness, it is known that any accumulation point must be a linear combination of the different stationary measures but no general method seem to exist to characterize this combination.

The method of Luczak and Norris [LN13] does not yield this problem. However, we have not been able to apply their results to our case. It seems plausible to modify their arguments in order to obtain Theorem 2.2 but only for $a < \frac{1}{3}$. The method that we develop here is close in spirit to theirs but is more tailored to Markov processes. The approximation (2.5) is controlled by martingale arguments and leverages properties of solutions to the Poisson equations (in ϕ) $L^q_{\mathrm{f}}\phi = g - \pi^q[g]$ associated with the fast generators L^q_{f} and to functions $g : V_0 \to \mathbb{R}$.

2.4.4. Nonstandard behavior

Taking the state space collapse and the stochastic averaging principle for granted, back-of-the-envelope computation can give insight into the nonstandard critical behavior observed for our system. As mentioned above, a consequence of the stochastic averaging and the criticality assumption is that $\pi^q(v) \approx \lambda^\infty_v$. However, taking into account the idle time induced by the necessary scheduling of the empty state which, when queue lengths are of the order of N, is of the order $\pi^q(0) \approx N^{-a}$, gives rise to the second-order approximation where $\lambda^\infty_v - \pi^q(v)$ is of the order of N^{-a} (see Section 8.2 for a more detailed heuristic). This suggests that node $v \in V$ behaves as a near-critical $M/M/1$ queue with arrival rate $\lambda^N_v = \lambda^\infty_v - N^{-a}\gamma_v$ and service rate $\lambda^\infty_v - N^{-a}$.

What is the right time scale for such a queue? A first-order asymptotic expansion of its generator can give a clue, namely, if time is sped up by N^b then

the action on its generator on a function f is given by

$$N^b\left(\lambda_v^\infty - N^{-a}\gamma_v\right)\left(f\left(q+\frac{1}{N}\right)-f(q)\right) + N^b\left(\lambda_v^\infty - N^{-a}\right)\left(f\left(q-\frac{1}{N}\right)-f(q)\right).$$

The leading term is $N^{b-a-1}f'(q)$ which suggests to take $b=a+1$, as turns out to be indeed the case. Moreover, we see that only first-order terms are dominant, which explains why the limiting process is deterministic and no diffusion term arises. This discussion also clearly highlights the key impact of idleness on the system performance at criticality, as without idleness, i.e., if we had $\lambda^\infty - \pi^q$ of the order of $1/N$, then we would see the usual N/N^2 scaling and a diffusion process in the limit.

3. Notation and main steps of the proof

We introduce in this section further notation, and then explain the main steps of the proof of Theorem 2.2.

3.1. Notation

We first gather notation used throughout the paper.

3.1.1. General notation

For $b>0$ and $x\in\mathbb{R}^V$ recall the notation $\|x\|_b = (|x_1|^b+\cdots+|x_n|^b)^{1/b}$ and $s(x)=x_1+\cdots+x_n$. We write $\|\cdot\|_\infty$ for the supremum norm, thus $\|f\|_\infty = \sup|f|$ for $f:\mathbb{R}^V\to\mathbb{R}$ and $\|q\|_\infty = \max_v|q_v|$ for $q\in\mathbb{R}^V$. If $U\subset\mathbb{R}^V$ and $f:\mathbb{R}^V\to\mathbb{R}$ we also define $\|f\|_{U,\infty}=\sup_{x\in U}|f(x)|$.

Whenever f is smooth enough, we denote by ∂_v its partial derivative along q_v and $\partial^2_{v,w}$ its second-order derivative along q_v and q_w, i.e.,

$$\partial_v f = \frac{\partial f}{\partial q_v} \quad\text{and}\quad \partial^2_{v,w} f = \frac{\partial^2 f}{\partial q_v \partial q_w}.$$

We will also consider the discrete differences $\Delta^N_{\pm,v}f$ for a function $f:\mathbb{R}^V\times V_0\to\mathbb{R}$, given by

$$\Delta^N_{\pm,v}f(q,\sigma) = f\left(q\pm\frac{e^v}{N},\sigma\right)-f(q,\sigma).$$

Thus, $N\Delta^N_{\pm,v}f\to\pm\partial_v f$ as $N\to\infty$ for f differentiable.

3.1.2. Generators

Let $E^N = \frac{1}{N}\mathbb{N}^V$ be the state space of the scaled process Q^N. We define L^N, $L_{\mathrm{f}}^{N,q}$ and $L_{\mathrm{s}}^{N,\sigma}$ for $q \in E^N$ and $\sigma \in V_0$ the scaled generators with arrival rates λ^N: for $f : E^N \times V_0 \to \mathbb{R}$, $g : E^N \to \mathbb{R}$, $h : V_0 \to \mathbb{R}$, $q \in E^N$ and $\sigma \in V_0$,

$$L^N f(q,\sigma) = N^{a+1} L f^{(N)}(Nq,\sigma),\ L_{\mathrm{s}}^{N,\sigma} g(q) = N^{a+1} L_{\mathrm{s}}^{\sigma} g^{(N)}(Nq)$$

and

$$L_{\mathrm{f}}^{N,q} h(\sigma) = N^{a+1} L_{\mathrm{f}}^{Nq} h(\sigma)$$

with $f^{(N)}(q,\sigma) = f(q/N,\sigma)$ and $g^{(N)}(q) = g(q/N)$ for $q \in \mathbb{N}^V$. Note that the stationary distribution of $L_{\mathrm{f}}^{N,q}$ is π^{Nq}. Let Γ^N be the *carré du champ* operator associated with L^N: for any $f : E^N \times V_0 \to \mathbb{R}$, . We have

$$\Gamma^N(f) = L^N(f^2) - 2fL^N(f)$$

and elementary computation shows that for $(q,\sigma) \in E^N \times V_0$ we have

$$\begin{aligned}\Gamma^N f(q,\sigma) &= N^{a+1} \sum_{v\in V} \lambda_v^N \left(f\left(q + \frac{e^v}{N}, \sigma\right) - f(q,\sigma)\right)^2 \\ &+ N^{a+1} \sum_{v\in V} \sigma_v \mathbb{1}_{q_v>0} \left(f\left(q - \frac{e^v}{N}, \sigma\right) - f(q,\sigma)\right)^2 \\ &+ N^{a+1} \sum_{v\in V} \left(f(q,0) - f(q,e^v)\right)^2 \left(\frac{\sigma_v}{1+(Nq_v+1)^a} + \frac{\sigma_0}{1+(Nq_v+1)^{-a}}\right).\end{aligned} \tag{3.1}$$

From standard Markov process theory, for any function f the process

$$M_f^N(t) = f(Q^N(t),\sigma^N(t)) - f(Q^N(0),\sigma^N(0)) - \int_0^t L^N f(Q^N(s),\sigma^N(s))\mathrm{d}s$$

is a local martingale with increasing process

$$\left\langle M_f^N \right\rangle (t) = \int_0^t \Gamma^N f(Q^N(s),\sigma^N(s))\mathrm{d}s.$$

For $N \geq 1$ we consider the homogenized generator L_{h}^N acting on functions $f : E^N \to \mathbb{R}$ as

$$\begin{aligned}L_{\mathrm{h}}^N f(q) = N^{a+1} \sum_{v\in V} \lambda_v^N &\left(f\left(q + \frac{e^v}{N}\right) - f(q)\right) \\ &+ N^{a+1} \sum_{v\in V} \pi^{Nq}(v) \mathbb{1}_{q_v>0} \left(f\left(q - \frac{e^v}{N}\right) - f(q)\right).\end{aligned} \tag{3.2}$$

This is the same generator as the generator L_{s}^{σ} of the (scaled) slow process given by (2.1), but where the instantaneous service rate σ_v of node v is replaced by its average value $\pi^q(v)$.

3.1.3. Poisson equation

For any function $g : V_0 \to \mathbb{R}$ and any $q \in E^N$ we denote by $\phi_g^N(q, \cdot)$ the unique solution to the Poisson equation associated with the scaled fast generator $L_{\mathrm{f}}^{N,q}$ and the function g, i.e., $\phi_g^N(q, \cdot)$ is the unique solution with $\pi^{Nq}[\phi_g^N(q, \cdot)] = 0$ to the equation with unknown ϕ

$$L_{\mathrm{f}}^{N,q}\phi = g - \pi^q[g]. \tag{3.3}$$

In the sequel, we will be particularly interested in $\phi_v^N(q, \cdot)$ solution to (3.3) with $g(\sigma) = \sigma_v$ for $v \in V_0$, which therefore satisfies for any $q \in E^N$ and any $\sigma \in V_0$

$$L_{\mathrm{f}}^{N,q}\left(\phi_v^N(q, \cdot)\right)(\sigma) = \sigma_v - \pi^{Nq}(v). \tag{3.4}$$

3.1.4. Initial state, limiting ODE

Recall that we fix throughout an initial state $q^0 \in I \setminus \{0\}$ and we assume that $Q^N(0) \to q^0$. Moreover, we consider $S = (S(t), t \geq 0)$ the solution to the ODE

$$\dot{x} = \mu x^{-a} - s(\gamma)$$

with initial condition $S(0) = s(q^0)$. We also consider $q = (q(t), t \geq 0)$ the $\mathbb{R}^V$-valued function with $s \circ q = S$ and $q(t) \in I$ for all $t \geq 0$, i.e.,

$$q_v(t) = \left(\frac{\lambda_v^\infty}{\mu}\right)^{1/a} S(t),\ t \geq 0,\ v \in V.$$

Note that for any choice of $\gamma \in \mathbb{R}^V$, $S(t)$ is bounded away from zero, i.e., $\inf_{t\geq 0} S(t) > 0$. If $s(\gamma) = 0$, S has an explicit expression:

$$S(t) = \left(\mu(a+1)t + s(q^0)^{a+1}\right)^{1/(a+1)},\ t \geq 0.$$

3.1.5. Localization, constants

Most of the proof of Theorem 2.2 is carried out for a localized process $Q^N(t \wedge T^N)$ with T^N the first time that Q^N significantly departs from q. More precisely, in the rest of the paper we fix some finite time horizon $T > 0$ and we consider the following two constants:

$$M = \min\left(2 \sup_{[0,T]} S, \frac{2}{\inf_{[0,T]} S}, \frac{1}{2}\right) \quad \text{and} \quad m = \frac{1}{M\mu^{1/a}} \min_v \left(\lambda_v^\infty\right)^{1/a}.$$

Here and in the sequel, we will treat as constants all numerical parameters that only depend on a, n, T, λ^∞, q^0 and the sequence (λ^N) as these are fixed throughout the entire paper. Moreover, we will use the letter C to denote positive and finite constants, that only depend on a, n, T, λ^∞, q^0 and (λ^N), and

whose precise value is irrelevant and that may change from line to line. Note in particular that the constants C do not depend on N, so that if $0 \leq u_N \leq Cv_N$ with $v_N \to 0$, then also $u_N \to 0$.

We then define

$$T^N := \inf\left\{t > 0 : \left\|Q^N(t) - q(t)\right\|_1 > \frac{m}{2}\right\},$$

the set $U \subset \mathbb{R}_+^V$

$$U := \left\{q \in \mathbb{R}_+^V : \frac{1}{M} < s(q) < M \quad \text{and} \quad \min_v q_v > m\right\},$$

its intersection U^N with E^N

$$U^N = U \cap E^N,$$

and the exit time of Q^N from U (or U^N):

$$\tau^N := \inf\left\{t \geq 0 : Q^N(t) \notin U\right\}.$$

Because jumps of Q^N are of size $1/N$, at time T^N we have

$$\left\|Q^N(T^N) - q(T^N)\right\| \leq \frac{m}{2} + \frac{1}{N}.$$

The constants m and M have been chosen such that the following result holds. The proof is computational and omitted.

Lemma 3.1. *We have $T^N \leq \tau^N$. In particular, $Q^N(t \wedge T^N) \in U^N$ for all $t \geq 0$.*

3.1.6. Distance to I

In order to control the distance to the invariant manifold I given by (2.4), i.e., to control the state space collapse property, we will use the Kullback-Leibler divergence between λ and $(\pi^q(v), v \in V)$ (note that the latter is not a probability measure). More precisely, for $q \in \mathbb{R}_+^V$ and $N \geq 1$ let

$$d^N(q) = \sum_{v \in V} \lambda_v^\infty \log\left(\frac{\lambda_v^\infty}{\pi^{Nq}(v)}\right).$$

When $N \to \infty$ and $q \in U$ we have $\pi^{Nq}(v) \to \pi_\infty^q(v)$ where

$$\pi_\infty^q(v) = \frac{q_v^a}{\|q\|_a^a}, \ v \in V.$$

We thus introduce

$$d^\infty(q) = \sum_{v \in V} \lambda_v^\infty \log\left(\frac{\lambda_v^\infty}{\pi_\infty^q(v)}\right).$$

which therefore satisfies $d^N(q) \to d^\infty(q)$ as $N \to \infty$. The convergence is actually uniform in $q \in U$, as the next lemma states (the proof is omitted).

Lemma 3.2. *As $N \to \infty$ we have*

$$\sup_{q \in U^N} \left| d^N(q) - d^\infty(q) \right| \to 0.$$

Note that $d^\infty(q) = 0$ if and only if $q \in I$, so d^∞ can indeed be seen as a distance to I. For $x \in I$ we have by definition $x_v = (\lambda^\infty/\mu)^{1/a} s(x)$. The distance d^∞ to I will actually also control the difference between x_v and $(\lambda^\infty/\mu)^{1/a} s(x)$, and the next lemma will prove useful in the sequel.

Lemma 3.3. *For $x \in U$ we have*

$$\left| x_v - \left(\frac{\lambda_v^\infty}{\mu} \right)^{1/a} s(x) \right| \leq C \left[d^\infty(x) \right]^{1/2}, \ v \in V.$$

Proof. Because $y \in [m, M] \to y^{1/a}$ is Lipschitz, for $x \in [m, M]^V$ and $v \in V$ we have

$$\left| x_v - (\lambda_v^\infty)^{1/a} \left\| x \right\|_a \right| \leq C \left| x_v^a - \lambda_v^\infty \left\| x \right\|_a^a \right| \leq C \left| \pi_\infty^x(v) - \lambda_v^\infty \right|$$

and so Pinsker's inequality gives

$$\left| x_v - (\lambda_v^\infty)^{1/a} \left\| x \right\|_a \right| \leq C \left[d^\infty(x) \right]^{1/2}.$$

Thus, since $s(x) = \sum_v x_v$ and $\mu^{1/a} = \sum_v (\lambda_v^\infty)^{1/a}$ we also have

$$\left| s(x) - \mu^{1/a} \left\| x \right\|_a \right| \leq \sum_{v \in V} \left| x_v - (\lambda_v^\infty)^{1/a} \left\| x \right\|_a \right| \leq C \left[d^\infty(x) \right]^{1/2}.$$

Finally, since

$$\left| x_v - \left(\frac{\lambda_v^\infty}{\mu} \right)^{1/a} s(x) \right| \leq \left| x_v - (\lambda_v^\infty)^{1/a} \left\| x \right\|_a \right| + \left| (\lambda_v^\infty)^{1/a} \left\| x \right\|_a - \left(\frac{\lambda_v^\infty}{\mu} \right)^{1/a} s(x) \right|,$$

we obtain the result. □

3.2. Main steps

The proof of Theorem 2.2 has three main steps which are proved in Sections 4–7.

3.2.1. First step: homogenization

The first main step of the proof is the following averaging result: we give the main idea of its proof below, and defer the full proof to Sections 4 and 5. Recall that C denotes a numerical constant allowed to depend on a, n, T, λ^∞, (λ^N) and q^0.

Proposition 3.4. *If $f : U \to \mathbb{R}$ is continuously differentiable, then for any $v \in V$ we have*

$$\mathbb{E}\left[\sup_{0\le t\le T\wedge T^N}\left|\int_0^t \left(\sigma_v^N(s) - \pi^{NQ^N(s)}(v)\right) f\left(Q^N(s)\right) \mathrm{d}s\right|\right] \le C\,\|f\|_{\infty,U}\,\frac{(\log N)^{3/2}}{N^{1/2}} + C\max_v \|\partial_v f\|_{\infty,U}\,\frac{(\log N)^{3/2}}{N^{1-a}}.$$

The proof of this result has two steps: first, provide a bound in terms of solutions to the Poisson equation (3.3) and then controlling these solutions. These two steps are performed in Sections 4 and 5, respectively and, as far as we know, the bounds that we derive there are new. To see how the Poisson equation arises, let us proceed with the following preliminary computation. We get from (3.4)

$$\sigma_v^N(s) - \pi^{NQ^N(s)}(v) = L_{\mathrm{f}}^{N,Q^N(s)}\left(\phi_v^N(Q^N(s),\cdot)\right)\left(\sigma^N(s)\right).$$

Since f does not depend on σ, this makes it possible to rewrite

$$\int_0^t \left(\sigma_v^N(s) - \pi^{NQ^N(s)}(v)\right) f\left(Q^N(s)\right) \mathrm{d}s = \int_0^t L_{\mathrm{f}}^{N,Q^N(s)}\left(V_v^N(Q^N(s),\cdot)\right)\left(\sigma^N(s)\right) \mathrm{d}s$$

with $V_v^N(q,\sigma) = \phi_v^N(q,\sigma) f(q)$. Making use of the martingale decomposition, we finally rewrite this as

$$\begin{aligned}\int_0^t \left(\sigma_v^N(s) - \pi^{NQ^N(s)}(v)\right) f\left(Q^N(s)\right) \mathrm{d}s &= V_v^N\left(Q^N(t),\sigma^N(t)\right) - V_v^N\left(Q^N(0),\sigma^N(0)\right) \\ &\quad - \int_0^t L_{\mathrm{s}}^{N,\sigma^N(s)}\left(V_v^N(\cdot,\sigma^N(s))\right)\left(Q^N(s)\right) \mathrm{d}s - M_{V_v^N}^N(t). \quad (3.5)\end{aligned}$$

This expression will be the basis for the proof of Proposition 3.4.

3.2.2. Second step: state space collapse

Using the averaging result of Proposition 3.4, the next step is to prove the following state space collapse result.

Proposition 3.5. *As $N \to \infty$ we have*

$$\mathbb{E}\left[\sup_{0\le t\le T\wedge T^N} d^\infty\left(Q^N(t)\right)\right] \to 0.$$

The proof proceeds by controlling the action of the homogenized generator L^N on d^N and then use this result to control $d^\infty \circ Q^N$ thanks to the averaging result of Proposition 3.4.

3.2.3. Third step: full proof

The third step of the proof consists in showing that $Q^N(\cdot \wedge T^N) \Rightarrow q$. The proof proceeds in two steps: first we establish the convergence of the one-dimensional total queue length process $s \circ Q^N(\cdot \wedge T^N) \Rightarrow s \circ q = S$ by using Gronwall's lemma. Together with the state space collapse property of Proposition 3.5, this gives the convergence of the entire n-dimensional process $Q^N(\cdot \wedge T^N)$ stopped at time T^N.

We finally conclude the proof: because the limiting process q does not exit the set U by time T, we prove that with high probability Q^N also stays in U by time T: this implies in particular that $\mathbb{P}(T^N \geq T) \to 1$ which makes it possible to transfer the convergence result from the stopped process $Q^N(\cdot \wedge T^N)$ to the unstopped one Q^N.

4. Control of homogenization in terms of solutions to the Poisson equation

This section provides a first step toward the proof of Proposition 3.4. We first derive a bound in terms of the following constants:

$$\Omega_N \coloneqq \sup_{q \in U^N,\, \|g\|_\infty \leq 1} \left\| \phi_g^N(q, \cdot) \right\|_\infty ,$$

$$B_N \coloneqq \sup_{q \in U^N,\, \|g\|_\infty \leq 1} \max_{i \in V, \sigma \in V_0} \left| \Delta^N_{\pm,i} \phi_g^N(q, \sigma) \right|$$

and

$$\Theta_N = N^{a+1} B_N + N^{1/2} \Omega_N + N^{(a+1)/2} \Omega_N^{3/2} + N^{a+1} \Omega_N B_N^{1/2}$$

Lemma 4.1. *For any $v \in V$ we have*

$$\mathbb{E}\left[\sup_{0 \leq t \leq T \wedge T^N} \left| \int_0^t \left(\sigma_v^N(s) - \pi^{NQ^N(s)}(v) \right) f\left(Q^N(s)\right) \mathrm{d}s \right| \right] \leq C \|f\|_{\infty,U} \Theta_N + C \max_w \|\partial_w f\|_{\infty,U} \left(N^{(a+1)/2} B_N + N^a \Omega_N \right).$$

Recall from the arguments preceding (3.5) that $V_v^N(q,\sigma) = f(q) \phi_v^N(q,\sigma)$: thus for every v we have for $q \in U^N$

$$\left| V_v^N(q,\sigma) \right| \leq \|f\|_{\infty,U} \Omega_N \tag{4.1}$$

and

$$\left| \Delta^N_{\pm,w} V_v^N(q,\sigma) \right| \leq \max_w \|\partial_w f\|_{\infty,U} \frac{\Omega_N}{N} + \|f\|_{\infty,U} B_N. \tag{4.2}$$

We start with two preliminary lemmas.

Lemma 4.2. *We have*

$$\mathbb{E}\left[\int_0^{T \wedge T^N} \sigma_0^N(s) \mathrm{d}s \right] \leq C N^{-a} + C \Omega_N + C N^{a+1} B_N.$$

Proof. Note that

$$\pi^{NQ^N(s)}(0) = \frac{1}{1+\sum_{w\in V}(NQ_w^N(s)+1)^a}$$

and so since $Q_w^N(s) \geq m - 1/N$ for $t \leq T^N$, we have $\pi^{NQ^N(s)}(0) \leq CN^{-a}$ for $s \leq T^N$ and so

$$\mathbb{E}\left[\int_0^{T\wedge T^N} \sigma_0^N(s)\mathrm{d}s\right] \leq CN^{-a} + \mathbb{E}\left[\int_0^{T\wedge T^N} \left(\sigma_0^N(s) - \pi^{NQ^N(s)}(0)\right)\mathrm{d}s\right].$$

Starting from (3.5) with $f = 1$ and taking the mean, we obtain

$$\begin{aligned}\mathbb{E}\left[\int_0^{T\wedge T^N} \left(\sigma_0^N(s) - \pi^{NQ^N(s)}(0)\right)\mathrm{d}s\right] \\ = \mathbb{E}\left[\phi_0^N\left(Q^N(T\wedge T^N), \sigma^N(T\wedge T^N)\right)\right] - \phi_0^N\left(Q^N(0), \sigma^N(0)\right) \\ - \mathbb{E}\left[\int_0^{T\wedge T^N} L_{\mathrm{s}}^{N,\sigma^N(s)}\left(\phi_0^N(\,\cdot\,,\sigma^N(s))\right)\left(Q^N(s)\right)\mathrm{d}s\right].\end{aligned}$$

By definition of $L_{\mathrm{s}}^{N,\sigma}$ we have

$$\begin{aligned}L_{\mathrm{s}}^{N,\sigma^N(s)}\left(\phi_0^N(\,\cdot\,,\sigma^N(s))\right)\left(Q^N(s)\right) = N^{a+1}\sum_{v\in V}\lambda_v^N\Delta_{+,v}^N\phi_0^N(Q^N(s),\sigma^N(s)) \\ + N^{a+1}\sum_{v\in V}\sigma_v^N(s)\mathbb{1}_{Q_v^N(s)>0}\Delta_{-,v}^N\phi_0^N(Q^N(s),\sigma^N(s)).\end{aligned}$$

The result thus follows directly from the definitions of Ω_N and B_N since $Q^N(t\wedge T^N) \in U$ according to Lemma 3.1. □

Lemma 4.3. *We have*

$$\mathbb{E}\left[\sup_{0\leq t\leq T\wedge T^N}\left|M_{V_v^N}^N(t)\right|\right] \leq C\max_w\|\partial_w f\|_{\infty,U}N^{(a+1)/2}B_N + C\|f\|_{\infty,U}\Theta_N.$$

Proof. By Doob's inequality and It's isometry we have

$$\begin{aligned}\mathbb{E}\left[\sup_{t\leq T\wedge T^N}\left(M_{V_v^N}^N(t)\right)^2\right] &\leq 4\mathbb{E}\left[\left(M_{V_v^N}^N(T\wedge T^N)\right)^2\right] \\ &= 4\mathbb{E}\left[\langle M_{V_v^N}^N\rangle(T\wedge T^N)\right] \\ &= 4\mathbb{E}\left[\int_0^{T\wedge T^N}\Gamma^N V_v^N(Q^N(s),\sigma^N(s))\mathrm{d}s\right].\end{aligned}$$

According to (3.1), we have

$$\begin{aligned}\Gamma^N V_v^N(q,\sigma) &= N^{a+1}\sum_{w\in V}\lambda_w^N\left(\Delta_{+,w}^N V_v^N(q,\sigma)\right)^2\\ &+ N^{a+1}\sum_{w\in V}\sigma_w \mathbb{1}q_w>0\left(\Delta_{-,w}^N V_v^N(q,\sigma)\right)^2\\ &+ N^{a+1}\sum_{w\in V}\left(V_v^N(q,0)-V_v^N(q,w)\right)^2\frac{\sigma_w}{1+(Nq_w+1)^a}\\ &+ N^{a+1}\sum_{w\in V}\left(V_v^N(q,0)-V_v^N(q,w)\right)^2\frac{\sigma_0}{1+(Nq_w+1)^{-a}}.\end{aligned}$$

We integrate this quantity over the trajectory (Q^N,σ^N) for $t\le T\wedge T^N$: along this trajectory we bound the terms $\sigma_w^N(s)$ and $1/(1+(NQ_w^N(s)+1)^{-a})$ by one, the terms $1/(1+(NQ_w^N(s)+1)^a)$ by CN^{-a} (because $Q_w^N(s)\ge m-1/N$ for $t\le T^N$) and we use (4.1) and (4.2) to obtain

$$\begin{aligned}&\mathbb{E}\left[\int_0^{T\wedge T^N}\Gamma^N V_v^N(Q^N(s),\sigma^N(s))\mathrm{d}s\right]\\ &\qquad\le CN^{a+1}\left(\max_w\|\partial_w f\|_{\infty,U}\frac{\Omega_N}{N}+B_N\|f\|_{\infty,U}\right)^2\\ &\qquad\quad+C\|f\|_{\infty,U}^2N\Omega_N^2\\ &\qquad\quad+C\|f\|_{\infty,U}^2N^{a+1}\Omega_N^2\mathbb{E}\left[\int_0^{T\wedge T^N}\sigma_0^N(s)\mathrm{d}s\right].\end{aligned}$$

Using $(x+y)^2\le 2x^2+2y^2$ and Lemma 4.2, we therefore obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le t\le T\wedge T^N}M_{V_v^N}^N(t)^2\right]&\le C\max_w\|\partial_w f\|_{\infty,U}^2N^{a+1}B_N^2\\ &+C\|f\|_{\infty,U}^2\left(N^{a+1}B_N^2+N\Omega_N^2+N^{a+1}\Omega_N^3+N^{2a+2}\Omega_N^2B_N\right).\end{aligned}$$

The result then follows by Cauchy-Schwarz and sub-linearity of the square root, and also because

$$N^{(a+1)/2}B_N+N^{1/2}\Omega_N+N^{(a+1)/2}\Omega_N^{3/2}+N^{a+1}\Omega_NB_N^{1/2}\le C\Theta_N.$$

□

Proof of Lemma 4.1. Starting from (3.5), we obtain

$$\begin{aligned}\sup_{0\le t\le T\wedge T^N}\left|\int_0^t\left(\sigma_v^N(s)-\pi^{NQ^N(s)}(v)\right)f\left(Q^N(s)\right)\mathrm{d}s\right| &\le \left|V_v^N(Q^N(0),\sigma^N(0))\right|+\sup_{0\le t\le T\wedge T^N}\left|V_v^N(Q^N(t),\sigma^N(t))\right| \\ &+\sup_{0\le t\le T\wedge T^N}\left|\int_0^t L_{\mathrm{s}}^{N,\sigma^N(s)}\left(V_v^N\left(\cdot,\sigma^N(s)\right)\right)(Q^N(s))\mathrm{d}s\right| \\ &+\sup_{0\le t\le T\wedge T^N}\left|M_{V_v^N}^N(t)\right|.\end{aligned}$$

As $Q^N(t)\in U$ for $t\le T^N$ by Lemma 3.1, similar arguments as in the proof of Lemmas 4.2 and 4.3 give a control on the three first terms in the right-hand side of the previous display, namely

$$\left|V_v^N(Q^N(0),\sigma^N(0))\right|+\sup_{0\le t\le T\wedge T^N}\left|V_v^N(Q^N(t),\sigma^N(t))\right|\le C\|f\|_{\infty,U}\Omega_N$$

and

$$\begin{aligned}\sup_{0\le t\le T\wedge T^N}\left|\int_0^t L_{\mathrm{s}}^{N,\sigma^N(s)}\left(V_v^N\left(\cdot,\sigma^N(s)\right)\right)(Q^N(s))\mathrm{d}s\right| &\\ \le CN^a\Omega_N\max_w\|\partial_w f\|_{\infty,U}+C\|f\|_{\infty,U}N^{a+1}B_N.&\end{aligned}$$

Combining these bounds with the bound of Lemma 4.3 gives the result. □

5. Control of solutions to the Poisson equation

In the previous section we have established a bound on some averaging property in terms of the constants Ω_N and B_N. The goal of this section is to prove the following result which provides a bound on these constants.

Lemma 5.1. *We have the following two bounds:*

$$\Omega_N\le C\frac{(\log N)^{3/2}}{N}\quad\text{and}\quad B_N\le C\frac{(\log N)^3}{N^2}.$$

We will prove this in a series of lemmas. It is more convenient to focus on unscaled quantities. For $q\in\mathbb{N}^V$ let α^q and ℓ^q be the log-Sobolev constant and spectral gap associated with L_{f}^q, respectively, and $\phi_g(q,\cdot)$ the solution to the Poisson equation $L_{\mathrm{f}}^q\varphi=g-\pi^q[\varphi]$.

Lemma 5.2. *For $q\in\mathbb{N}^V$ and $v\in V$ let*

$$\Omega(q)=\frac{(\log(1/\pi^q(0)))^{1/2}\log(1/\pi^q(0)-1)}{\ell^q(1-2\pi^q(0))}$$

and

$$B_v(q) = \frac{\Omega(q)}{q_v^{1-a}} \left(\frac{4\Omega(q)}{q_v^{2a}} + \pi^q(0) \right).$$

Then

$$\|\phi_g(q, \cdot)\|_\infty \leq \|g\|_\infty \Omega(q) \tag{5.1}$$

and

$$\|\phi_g(q \pm e^v, \cdot) - \phi_g(q, \cdot)\|_\infty \leq \|g\|_\infty B_v(q). \tag{5.2}$$

Proof. Let $m^q_{\sigma,t}$ denote the law at time t of the Markov process starting at σ with generator L^q_{f}: then it is well-known that $\phi_g(q, \cdot,)$ is given by

$$\phi_g(q, \sigma) = -\int_0^\infty \left(m^q_{\sigma,t}[g] - \pi^q[g] \right) \mathrm{d}t.$$

This gives

$$\|\phi_g(q, \cdot)\|_\infty \leq 2 \|g\|_\infty \int_0^{+\infty} \left\| m^q_{\sigma,t} - \pi^q \right\|_{\mathrm{TV}} \mathrm{d}t$$

with $\|\cdot\|_{\mathrm{TV}}$ the total variation distance and then

$$\|\phi_g(q, \cdot)\|_\infty \leq 2 \|g\|_\infty \int_0^{+\infty} \left(\frac{1}{2} m^q_{\sigma,t} \left[\varphi^q_{\sigma,t} \right] \right)^{1/2} \mathrm{d}t$$

where $\varphi^q_{\sigma,t} = \log(m^q_{\sigma,t}/\pi^q)$, by Pinsker's inequality. As $\min_\sigma \pi^q(\sigma) = \pi^q(0)$, Theorem 3.6 in [DSC96] gives

$$m^q_{\sigma,t} \left[\varphi^q_{\sigma,t} \right] \leq \log(1/\pi^q(0)) e^{-4\alpha^q t}$$

while Corollary 2.2.10 in [SC97] gives

$$\alpha^q \geq \frac{1 - 2\pi^q(0)}{\log((1 - \pi^q(0))/\pi^q(0))} \ell^q.$$

Gathering the three previous bounds gives the desired bound (5.1) on $\|\phi_g(q, \cdot)\|_\infty$. We now prove (5.2). Fix temporarily $v \in V$, $q \in \mathbb{N}^V$ with $q_v > 0$ and let $\Phi = \phi_g(q - e^v, \cdot) - \phi_g(q, \cdot)$ and $G = L^q_{\mathrm{f}}(\Phi)$. Since $\pi^q(G) = 0$, the first bound (5.1) thus implies

$$\|\phi_g(q - e^v, \cdot) - \phi_g(q, \cdot)\|_\infty \leq \Omega(q) \|G\|_\infty .$$

Since by definition of ϕ_g we have $L^q_{\mathrm{f}}(\phi_g(q, \cdot))(\sigma) = g(\sigma) - \pi^q[g]$ we obtain

$$G(\sigma) = -\left(L^{q-e^v}_{\mathrm{f}} - L^q_{\mathrm{f}} \right) (\phi_g(q - e^v, \cdot))(\sigma) - \sum_{\rho \in V_0} (\pi^{q-e^v}(\rho) - \pi^q(\rho)) g(\rho)$$

and so

$$\|G\|_\infty \leq \left\| \left(L^{q-e^v}_{\mathrm{f}} - L^q_{\mathrm{f}} \right) (\phi_g(q - e^v, \cdot)) \right\|_\infty + \|g\|_\infty \sum_{\rho \in V_0} \left| \pi^{q-e^v}(\rho) - \pi^q(\rho) \right|.$$

For any function $h : V_0 \to \mathbb{R}$ we have according to (2.2)

$$\left(L_{\mathrm{f}}^{q-e^v} - L_{\mathrm{f}}^q\right)(h)(\sigma) = \sigma_v \left(\Psi_-(q_v - 1) - \Psi_-(q_v)\right)\left(h(\sigma - e^v) - h(\sigma)\right) + \sigma_0 \left(\Psi_+(q_v - 1) - \Psi_+(q_v)\right)\left(h(\sigma + e^v) - h(\sigma)\right)$$

and so since $\Psi_+ + \Psi_- = 1$, this gives

$$\left\| \left(L_{\mathrm{f}}^{q-e^v} - L_{\mathrm{f}}^q\right)(h) \right\|_\infty \leq 4 \|h\|_\infty \left|\Psi_-(q_v - 1) - \Psi_-(q_v)\right|.$$

Therefore, using again the bound (5.1) gives

$$\left\| \left(L_{\mathrm{f}}^{q-e^v} - L_{\mathrm{f}}^q\right)(\phi_g(q - e^v, \cdot)) \right\|_\infty \leq 4\Omega(q) \|g\|_\infty \int_0^1 |\Psi_d'(q_v - u)| \, \mathrm{d}u.$$

Direct calculation yields

$$\Psi_-'(q_v) = -\frac{a}{(q_v + 1)^{1-a}(1 + (q_v + 1)^a)^2}$$

and so $\left|\Psi_-'(q_v - u)\right| \leq \frac{q_v^{a-1}}{(1+q_v^a)^2}$ as long as $u \leq 1$. Moreover, for any $v \in V$ and $w \in V_0$ with $w \neq v$, one can check that

$$|\partial_v \pi^q(w)| = a(q_v + 1)^{a-1} \pi^q(w) \pi^q(0)$$

and

$$|\partial_v \pi^q(v)| = a(q_v + 1)^{a-1} \pi^q(0)(1 - \pi^q(v))$$

so that in any case, $|\partial_v \pi^q(\sigma)| \leq a\pi^q(0)(q_v + 1)^{a-1}$. Gathering the previous bounds gives the result. □

We now prove a lower bound on the spectral gap of L_{f}^q. Related bounds were for instance proved in [SS12] using Cheeger's inequality in a more general setting. However, this method would only lead to $\ell^q \geq C\|q + 1\|_\infty^{-2a}$ which is not sharp enough in our case.

Lemma 5.3. *For any $q \in \mathbb{N}^V$ we have*

$$\ell^q \geq \frac{C}{\|q + 1\|_\infty^a}.$$

Proof. Let $(\eta(t), t \geq 0)$ be a Markov process with generator L_{f}^q and $\mathbb{P}_\sigma^q$ its law started from $\sigma \in V_0$. Let as in the previous proof $m_{\sigma,t}^q$ denote the law of $\eta(t)$ under $\mathbb{P}_\sigma^q$ and define the random times

$$T_{\mathrm{mix}}^q = \inf\left\{ t \geq 0 : \max_{\sigma \in V} \|m_{\sigma,t}^q - \pi^q\|_{\mathrm{TV}} < \frac{1}{2e} \right\}$$

and

$$T_{\mathrm{hit}}^q = \max_{\sigma \in V, A \subset V_0} \pi^q(A) \mathbb{E}_{\sigma^0}^q(T_A)$$

with T_A the hitting time of A for η:

$$T_A = \inf\{t \geq 0 : \eta(t) \in A\}, \ A \subset V_0.$$

Recall that ℓ^q is the spectral gap of L^q_{f}. It is proved in [LP17] that $T^q_{\mathrm{mix}} \geq 1/\ell^q - 1$ (the proof for discrete time extends to continuous time) and in [Ald82] that $T^q_{\mathrm{mix}} \leq c_0 T^q_{\mathrm{hit}}$ for some universal constant c_0. Combining those two results, we get that

$$\ell^q \geq \frac{1}{c_0 T^q_{\mathrm{hit}} + 1}$$

and so in order to prove the desired bound, we only need to prove that $T^q_{\mathrm{hit}} \leq C\|q+1\|^a_\infty$. Since

$$T^q_{\mathrm{hit}} \leq \max_{\sigma^0 \in V_0} \sum_{\sigma \in V_0} \mathbb{E}^q_{\sigma^0}(T_\sigma)$$

this actually reduces to proving that

$$\mathbb{E}^q_\sigma(T_0) \leq C\|q+1\|^a_\infty \ \text{ and } \ \mathbb{E}^q_0(T_\sigma) \leq C\|q+1\|^a_\infty \tag{5.3}$$

for any $\sigma \in V$. Indeed, for $\sigma^0 \neq \sigma \in V$ the process η needs to pass through 0 to go from σ^0 to σ and so the strong Markov property gives

$$\mathbb{E}^q_{\sigma^0}(T_\sigma) = \mathbb{E}^q_{\sigma^0}(T_0) + \mathbb{E}^q_0(T_\sigma).$$

So let us prove (5.3). The bound on $\mathbb{E}^q_\sigma(T_0)$ is obvious since by definition T_0 under $\mathbb{E}^q_\sigma$ is an exponential random variable with parameter $\Psi_-(q_\sigma)$ so that

$$\mathbb{E}^q_\sigma(T_0) = \frac{1}{\Psi_-(q_\sigma)} = 1 + (q_v+1)^a \leq C\|q+1\|^a_\infty.$$

Let us now prove that $\mathbb{E}^q_0(T_\sigma) \leq C\|q+1\|^a_\infty$. Under $\mathbb{P}^q_0$, decompose the trajectory $(\eta(t), 0 \leq t \leq T_\sigma)$ into cycles away from 0: in the a-th cycle, η stays in 0 for a duration X_a, then moves to some $i \in V$ where it stays for a duration Y_a and then comes back to 0. If $A \in \{1, \dots, \}$ denotes the first cycle where η visits σ, we can thus write

$$T_\sigma = \sum_{a=1}^{A-1} (X_a + Y_a) + X_A.$$

Each time η leaves 0, it goes to $i \in V$ with probability

$$p^q_v = \frac{\Psi_+(q_v)}{\sum_{w \in V} \Psi_+(q_v)} = \left(\sum_{w \in V} \frac{1 + (q_v+1)^{-a}}{1 + (q_w+1)^{-a}} \right)^{-1} \geq \frac{1}{2n}.$$

In particular, with a suitable coupling we can write

$$T_\sigma \leq \sum_{a=1}^{G} (X_a + Y_a)$$

with G a geometric random variable with parameter $1/(2n)$ independent from the X_a and Y_a's. Since the $(X_a, a \geq 1)$ and $(Y_a, a \geq 1)$ are two independent sequences of i.i.d. random variables, it follows that

$$\mathbb{E}_0^q(T_\sigma) \leq \mathbb{E}(G)\left[\mathbb{E}_0^q(X_1) + \mathbb{E}_0^q(Y_1)\right].$$

By definition, X_1 under $\mathbb{P}_0^q$ is an exponential random variable with parameter

$$\sum_{v\in V} \Psi_+(q_v) = \sum_{v\in V} \frac{1}{1+(q_v+1)^{-a}} \geq \frac{n}{2}$$

so that $\mathbb{E}_0^q(X_1) \leq 2/n$. Moreover, Y_1 under $\mathbb{P}_0^q$ is distributed as T_0 under $\mathbb{P}_\Sigma^q$ with $\mathbb{P}(\Sigma = v) = p_v^q$ for $v \in V$, so that

$$\mathbb{E}_0^q(Y_1) = \sum_{v\in V} p_v^q \mathbb{E}_v^q(T_0) \leq C\|q+1\|_\infty^a$$

using $\mathbb{E}_v^q(T_0) \leq C\|q+1\|_\infty^a$. Gathering the previous bounds yields the desired result. □

Thanks to Lemmas 5.2 and 5.3 we now provide a proof of Lemma 5.1.

Proof of Lemma 5.1. Let $g : V_0 \to \mathbb{R}$ and $q \in E^N$ given. Recall that $\phi_g^N(q, \cdot,)$ and $\phi_g(Nq, \cdot,)$ are such that

$$L_{\mathrm{f}}^{N,q}\left(\phi_g^N(q, \cdot,)\right) = g - \pi^{Nq}[g] = L_{\mathrm{f}}^{Nq}\left(\phi_g(Nq, \cdot,)\right)$$

and since $L_{\mathrm{f}}^{N,q} = N^{a+1} L_{\mathrm{f}}^{Nq}$, this gives $\phi_g^N(q, \sigma) = N^{-(a+1)}\phi_g(Nq, \sigma)$ by uniqueness. According to (5.1) and (5.2) this gives

$$\left\|\phi_g^N(q, \cdot)\right\|_\infty \leq \|g\|_\infty \frac{\Omega(Nq)}{N^{a+1}} \quad \text{and} \quad \left\|\Delta_{\pm,v}^N \phi_g^N(q, \cdot)\right\|_\infty \leq \|g\|_\infty \frac{B_v(Nq)}{N^{a+1}}$$

and so in order to prove the result, we only have to prove that

$$\Omega(Nq) \leq CN^{a+1}\frac{(\log N)^{3/2}}{N} \quad \text{and} \quad B_v(Nq) \leq CN^{a+1}\frac{(\log N)^3}{N^2}$$

for $q \in U^N$. To do so, note that for $q \in U^N$ we have

$$n(Nm)^a \leq \frac{1}{\pi^{Nq}(0)} \leq 1 + n(NM+1)^a$$

and so our convention makes it possible to write $CN^{-a} \leq \pi^{Nq}(0) \leq CN^{-a}$. Since $\ell^{Nq} \geq CN^{-a}$ by Lemma 5.3, for $q \in U^N$ we obtain the desired bounds on $\Omega(Nq)$ and $B_v(Nq)$. □

Proof of Proposition 3.4. The proof of Proposition 3.4 now follows readily from Lemmas 4.1 and 5.1 since for $a < 1/2$ one readily checks that $\Theta_N \leq C(\log N)^{3/2}/N^{1/2}$ and

$$N^{(a+1)/2} B_N + N^a \Omega_N \leq C\frac{(\log N)^{3/2}}{N^{1-a}}.$$

□

6. State space collapse

In this section we prove Proposition 3.5 through a series of lemmas. In view of Lemma 3.2, it is enough to prove the result with d^N instead of d^∞, i.e., to prove that

$$\mathbb{E}\left[\sup_{0\le t\le T\wedge T^N} d^N(Q^N(t))\right]\to 0.$$

Starting from the semimartingale decomposition of $d^N\circ Q^N$ and then adding and subtracting L^N_{h}, we obtain

$$\begin{aligned} d^N(Q^N(t)) = d^N(Q^N(0)) &+ \int_0^t L^N_{\mathrm{h}} d^N(Q^N(s))\mathrm{d}s \\ &+ \int_0^t (L^N - L^N_{\mathrm{h}}) d^N(Q^N(s),\sigma^N(s))\mathrm{d}s + M^N_{d^N}(t) \end{aligned}$$

where, in order to give sense to $L^N d^N$ we consider $d^N(q,\sigma)=d^N(q)$. Taking the supremum and the expectation and using that $Q^N(t)\in U^N$ for all $t\le T^N$, this leads to

$$\mathbb{E}\left[\sup_{0\le t\le T\wedge T^N} d^N(Q^N(t))\right] \le d^N(Q^N(0)) + T\sup_{q\in U^N} L^N_{\mathrm{h}} d^N(q) + I + II \tag{6.1}$$

with

$$I = \mathbb{E}\left[\sup_{0\le t\le T\wedge T^N}\left|\int_0^{t\wedge T^N}(L^N - L^N_{\mathrm{h}}) d^N(Q^N(s),\sigma^N(s))\mathrm{d}s\right|\right]$$

and

$$II = \mathbb{E}\left[\sup_{0\le t\le T\wedge T^N}\left|M^N_{d^N}(t)\right|\right].$$

The first term $d^N(Q^N(0))$ in the right-hand side of (6.1) vanishes because $Q^N(0)\to q^0\in I$ (and because of Lemma 3.2). The next three lemmas show that $\sup_{q\in U^N} L^N_{\mathrm{h}} d^N(q)\to 0$ and that the terms I and II also vanish.

Lemma 6.1. *For N large enough, we have for any $q\in E^N\cap U$*

$$L^N_{\mathrm{h}} d^N(q) \le CN^{-a} + CN^{-(1-a)}.$$

Proof. Let $q\in U^N$ and for each $v\in V$, let $\zeta^v_\pm = (\zeta^v_{\pm,w}, w\in V)$ such that

$$d^N\left(q\pm\frac{e^v}{N}\right) = d^N(q)\pm\frac{1}{N}\partial_v d^N(q) + \frac{1}{2N^2}\partial^2_{v,v} d^N(\zeta^v_\pm).$$

Note that $\zeta^v_{\pm,w} = q_w$ if $w \neq v$ and $|\zeta^v_{\pm,v} - q_v| \leq 1/N$. Then, recalling that $\Delta^N_{\pm,v} d^N(q) = d^N\left(q \pm e^v/N\right) - d^N(q)$, we obtain

$$\begin{aligned} L^N_{\mathrm{h}} d^N(q) &= N^{a+1} \sum_{v\in V} \lambda^N_v \Delta^N_{+,v} d^N(q) + N^{a+1} \sum_{v\in V} \pi^{Nq}(v) \Delta^N_{-,v} d^N(q) \\ &= N^{a+1} \sum_{v\in V} \left(\lambda^\infty_v - N^{-a}\gamma_v\right) \left(\frac{1}{N}\partial_v d^N(q) + \frac{1}{2N^2}\partial^2_{v,v} d^N(\zeta^v_+)\right) \\ &\quad + N^{a+1} \sum_{v\in V} \pi^{Nq}(v) \left(-\frac{1}{N}\partial_v d^N(q) + \frac{1}{2N^2}\partial^2_{v,v} d^N(\zeta^v_-)\right) \\ &= A + B \end{aligned}$$

with

$$A = N^a \sum_{v\in V} \left(\lambda^\infty_v - \pi^{Nq}(v)\right) \partial_v d^N(q) - \sum_{v\in V} \gamma_v \partial_v d^N(q)$$

and

$$B = \frac{1}{2N^{1-a}} \sum_{v\in V} \left(\left(\lambda^\infty_v - N^{-a}\gamma_v\right) \partial^2_{v,v} d^N(\zeta^v_+) + \pi^{Nq}(v) \partial^2_{v,v} d^N(\zeta^v_-)\right).$$

We now show that $A \leq CN^{-a}$ and $B \leq CN^{-(1-a)}$, which will give the result. Let us start with controlling A. Let in the sequel $\delta^N_v(q) = \lambda^\infty_v - \pi^{Nq}(v)$. Then it may be checked through elementary algebra that

$$\partial_v d^N(q) = -\frac{a\delta^N_v(q)}{q_v + \frac{1}{N}}$$

which leads to the relation

$$A = -aN^a \sum_{v\in V} \frac{\delta^N_v(q)^2}{q_v + \frac{1}{N}} - a \sum_{v\in V} \frac{\gamma_v \delta^N_v(q)}{q_v + \frac{1}{N}}.$$

Since $q \in U^N$, and in particular $m \leq q_v \leq M$ for every v, we obtain by using the equivalence of the L_1 and L_2 norms that

$$A \leq -c_1 N^a \left\|\delta^N(q)\right\|_2^2 + c_2 \left\|\delta^N(q)\right\|_2$$

for some positive constants c_1 and c_2 that only depend on n, m, M and γ. It is readily checked that the supremum of the function $x \mapsto c_2 x - c_1 N^a x^2$ is equal to $c_2^2/(4c_1 N^a)$, which gives $A \leq CN^{-a}$ as desired.

Let us now control B. Computing the second derivative of d^N gives

$$\partial^2_{v,v} d^N(q) = \frac{a\left(\delta^N_v(q) + a\pi^{Nq}(v)(1 - \pi^{Nq}(v))\right)}{(q_v + \frac{1}{N})^2}.$$

In particular, since $q \in U^N$ and $\|\zeta^v_\pm - q\|_\infty \leq 1/N$, we have

$$\left|\partial^2_{v,v} d^N\left(\zeta^N_\pm\right)\right| \leq \frac{1}{m^2}$$

and so

$$B \leq \frac{1}{2mN^{-(1-a)}} \left(2 + N^{-a} s(\gamma)\right)$$

which gives $B \leq CN^{-(1-a)}$ for N large enough, as desired. □

Lemma 6.2. *We have*

$$\mathbb{E}\left[\sup_{0\leq t\leq T\wedge T^N} \left| \int_0^{t\wedge T^N} (L^N - L^N_{\mathrm{h}}) d^N(Q^N(s), \sigma^N(s)) \mathrm{d}s \right| \right] \to 0.$$

Proof. Since

$$\begin{aligned}(L^N - L^N_{\mathrm{h}}) d^N(q,\sigma) &= (L^{N,\sigma}_{\mathrm{s}} - L^N_{\mathrm{h}}) d^N(q) \\ &= N^{a+1} \sum_{v\in V} \left(\sigma_v - \pi^{Nq}(v)\right) \mathbb{1}_{q_v>0} \Delta^N_{-,v} d^N(q)\end{aligned}$$

Proposition 3.4 gives

$$\begin{aligned}\mathbb{E}\Bigg[\sup_{0\leq t\leq T\wedge T^N} \Bigg| \int_0^{t\wedge T^N} (L^N - L^N_{\mathrm{h}}) d^N(Q^N(s), \sigma^N(s)) \mathrm{d}s \Bigg| \Bigg] \\ \leq CN^{a+1} \max_{v\in V} \left\|\Delta^N_{-,v} d^N\right\|_{\infty,U} \frac{(\log N)^{3/2}}{N^{1/2}} \\ + CN^{a+1} \max_{v,w} \left\|\partial_w \Delta^N_{-,v} d^N\right\|_{\infty,U} \frac{(\log N)^{3/2}}{N^{1-a}}.\end{aligned}$$

For $q \in U^N$ and $w \in V$ one can check that $\left|\Delta^N_{-,v} d^N(q)\right|, \left|\partial_w \Delta^N_{-,v} d^N(q)\right| \leq \frac{C}{N}$ so that

$$\mathbb{E}\left[\sup_{0\leq t\leq T\wedge T^N} \left| \int_0^{t\wedge T^N} (L^N - L^N_{\mathrm{h}}) d^N(Q^N(s), \sigma^N(s)) \mathrm{d}s \right| \right] \leq C \frac{(\log N)^{3/2}}{N^{1/2-a}}.$$

Thus for $a < 1/2$ this bound indeed vanishes, which proves the result. □

Lemma 6.3. *We have*

$$\mathbb{E}\left[\sup_{0\leq t\leq T\wedge T^N} \left| M^N_{d^N}(t) \right| \right] \leq \frac{C}{N^{(1-a)/2}}.$$

Proof. Proceeding as in the proof of Lemma 4.3 we obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{0\leq t\leq T\wedge T^N} M^N_{d^N}(t)^2 \right] \leq 4N^{a+1} \mathbb{E}\left[\int_0^{T\wedge T^N} \sum_{v\in V} (\Delta^N_{+,v} d^N(Q^N(s))^2 \mathrm{d}s\right] \\ + 4N^{a+1} \mathbb{E}\left[\int_0^{T\wedge T^N} \sum_{v\in V}^{n} (\Delta^N_{-,v} d^N(Q^N(s))^2 \mathrm{d}s\right].\end{aligned}$$

The result then follows from the same Taylor expansion as in the proof of Lemma 6.1. □

7. Proof of main result

To prove Theorem 2.2, we will establish its equivalent for the stopped process $Q^N(\cdot \wedge T^N)$ using Gronwall's lemma. We then transfer the result on the stopped process to Q^N using Lemma 3.1.

7.1. First step: $s \circ Q^N(\cdot \wedge T^N) \Rightarrow S$

The first step is to prove that $s \circ Q^N(\cdot \wedge T^N) \Rightarrow S$ uniformly on $[0,T]$, which we do now. Starting from the definition of $L_{\mathrm{s}}^{N,\sigma}$ and using $s(\lambda^N) = 1 - N^{-a}s(\gamma)$ by (2.3) and $\sum_{v\in V}\sigma_v = 1-\sigma_0$ we obtain

$$L_{\mathrm{s}}^{N,\sigma}s(q) = N^a s(\lambda^N) - N^a \sum_{v\in V}\sigma_v \mathbb{1}_{q_v>0} = N^a\sigma_0 + N^a\sum_{v\in V}\sigma_v \mathbb{1}_{q_v=0} - s(\gamma).$$

The semimartingale decomposition of $s\circ Q^N$ and the fact that

$$S(t) = S(0) + \mu\int_0^t S(s)^{-a}\mathrm{d}s - s(\gamma)t$$

by definition of S then leads to

$$s(Q^N(t)) - S(t) = s(Q^N(0)) - S(0) + \int_0^t \left(N^a\sigma_0^N - \frac{\mu}{S(s)^a}\right)\mathrm{d}s + \sum_{v\in V}\int_0^t \sigma_v^N(s)\mathbb{1}_{Q_v^N(s)=0}\mathrm{d}s + M_s^N(t). \quad (7.1)$$

Define

$$\varepsilon^N(t) = s(Q^N(0)) - S(0) + \eta^N(t) + e^N(t) + h^N(t) + M_s^N(t)$$

where

$$\eta^N(t) = \int_0^{t\wedge T^N}\left(\frac{1}{N^{-a} + \|Q^N(s)+1/N\|_a^a} - \frac{1}{\|Q^N(s)\|_a^a}\right)\mathrm{d}s,$$

$$e^N(t) = \int_0^{t\wedge T^N}\left(\frac{1}{\|Q^N(s)\|_a^a} - \frac{\mu}{s(Q^N(s))^a}\right)\mathrm{d}s$$

and

$$h^N(t) = N^a\int_0^{t\wedge T^N}\left(\sigma_0^N(s) - \pi^{NQ^N(s)}(0)\right)\mathrm{d}s.$$

Since $Q_v^N(s) > 0$ for $t < T^N$, starting from (7.1) and plugging in the above expressions, we obtain

$$s(Q^N(t\wedge T^N)) - S(t) = \varepsilon^N(t) + \mu\int_0^t\left(\frac{1}{s(Q^N(s))^a} - \frac{1}{S(s)^a}\right)\mathrm{d}s.$$

Since $x \in [m, M] \mapsto x^{-a}$ is Lipschitz and all queue lengths are in $[m, M]$ before time T^N, we finally obtain

$$\left|s(Q^N(t \wedge T^N)) - S(t)\right| \leq \left|\varepsilon^N(t)\right| + C \int_0^t \left|s(Q^N(s \wedge T^N)) - S(s)\right| \mathrm{d}s$$

and Gronwall's lemma implies

$$\sup_{0 \leq t \leq T} \left|s(Q^N(t \wedge T^N)) - S(t)\right| \leq \left(\left|s(Q^N(0)) - S(0)\right| + \bar{\eta}^N + \bar{e}^N + \bar{h}^N + \sup_{0 \leq t \leq T \wedge T^N} \left|M_s^N(t)\right|\right) e^{CT}$$

with

$$\bar{\eta}^N = \int_0^{T \wedge T^N} \left| \frac{1}{N^{-a} + \|Q^N(s) + 1/N\|_a^a} - \frac{1}{\|Q^N(s)\|_a^a} \right| \mathrm{d}s,$$

$$\bar{e}^N = \int_0^{T \wedge T^N} \left| \frac{1}{\|Q^N(s)\|_a^a} - \frac{\mu}{s(Q^N(s))^a} \right| \mathrm{d}s$$

and

$$\bar{h}^N = N^a \sup_{0 \leq t \leq T \wedge T^N} \left| \int_0^t \left(\sigma_0^N(s) - \pi^{NQ^N(s)}(0)\right) \mathrm{d}s \right|.$$

By assumption we have $s(Q^N(0)) \to S(0)$ and so in order to prove the desired result $s \circ Q^N(\cdot \wedge T^N) \Rightarrow S$ on $[0, T]$, we only have to prove that $\bar{\eta}^N, \bar{e}^N, \bar{h}^N$ and the martingale term vanish. The fact that $\bar{\eta}^N \Rightarrow 0$ comes directly from the fact that $Q^N(s) \in U$ for $s \leq T \wedge T^N$. The martingale term is handled with the exact same arguments as the previous martingale terms in Lemmas 4.3 and 6.3, the proof is omitted. The next two lemmas show that the last two terms $\bar{\epsilon}^N$ and $\bar{h}^N$ also vanish.

Lemma 7.1. *We have $\bar{e}^N \Rightarrow 0$.*

Proof. Since the process is stopped at T^N and so all coordinates considered are bounded away from 0, all the functions considered are Lipschitz and so according to Lemma 3.3 we have

$$\left| Q_v^N(s)^a - \frac{\lambda_v^\infty}{\mu} s(Q^N(s)) \right| \leq C d^\infty(Q^N(s))^{a/2}.$$

Using the triangular inequality and the fact that $s(\lambda^\infty) = 1$, we thus obtain

$$\left| \left\|Q^N(s)\right\|_a^a - \frac{1}{\mu} s(Q^N(s)) \right| \leq C d^\infty(Q^N(s))^{a/2}.$$

The convergence $\bar{e}^N \Rightarrow 0$ follows therefore readily from Proposition 3.5 which implies that $d^\infty(Q^N(s)) \Rightarrow 0$ uniformly in $s \leq T \wedge T^N$. □

Lemma 7.2. *We have* $\mathbb{E}(\bar{h}^N) \to 0$.

Proof. Since $\sigma_0 = \sum_{v\in V} \sigma_v$ and $\pi^q(0) = \sum_{v\in V} \pi^q(v)$ we have

$$\bar{h}^N \leq N^a \sum_{v\in V} \sup_{0\leq t\leq T\wedge T^N} \left| \int_0^t \left(\sigma_v^N(s) - \pi^{NQ^N(s)}(v)\right) \mathrm{d}s \right|$$

and so Proposition 4.1 with $f(q) = q$ implies that

$$\mathbb{E}(\bar{h}^N) \leq C \frac{(\log N)^{3/2}}{N^{1/2-a}}.$$

As $a < 1/2$ we have the result. □

7.2. Second step: proof of Theorem 2.2

We now conclude the proof of Theorem 2.2, so we have to control $Q_v^N(t) - q_v(t)$. The idea is to combine the convergence $s \circ Q^N(\cdot \wedge T^N) \Rightarrow s \circ q = S$ of the previous step, together with the state space collapse property $d^\infty \circ Q^N \Rightarrow 0$ of Proposition 3.5. Since $q(t) \in I$, we have $q_v(t) = (\lambda_v^\infty/\mu)^{1/a} S$ and so this leads us to write

$$\sup_{0\leq t\leq T\wedge T^N} \left|Q_v^N(t) - q_v(t)\right| \leq \sup_{0\leq t\leq T\wedge T^N} \left| Q_v^N(t) - \left(\frac{\lambda_v^\infty}{\mu}\right)^{1/a} s(Q^N(t)) \right| + \sup_{0\leq t\leq T\wedge T^N} \left|s(Q^N(t)) - S(t)\right|.$$

The second term vanishes by the first step, and so the first term vanishes as a consequence of the state space collapse property (combine Proposition 3.5 and Lemma 3.3). Thus we have proved that $Q^N(\cdot \wedge T^N) \Rightarrow q$.

Let us now remove the localization and prove that $Q^N \Rightarrow q$. In order to do so, it is enough to show that $\mathbb{P}(T^N \geq T) \to 1$. By definition of T^N, we have

$$\left\|Q^N(T^N) - q(T^N)\right\|_1 \geq \frac{m}{2}.$$

Since $T^N \wedge T = T^N$ in the event $\{T^N \leq T\}$, this entails

$$\mathbb{P}\left(T^N \leq T\right) \leq \mathbb{P}\left(\left\|Q^N(T^N \wedge T) - q(T^N \wedge T)\right\|_1 \geq \frac{m}{2}\right).$$

Since we have proved that $Q^N(\cdot \wedge T^N) \Rightarrow q$ uniformly on $[0, T]$, the previous probability vanishes. This concludes the proof of Theorem 2.2.

8. Extensions and directions for future research

8.1. Beyond $a < \frac{1}{2}$

Proposition 3.4 shows that the averaging approximation (2.5) holds for $a < 1$. This is in line with Lemma 5.3 which shows that the mixing time of the fast

process is of the order of N^a: since the typical time scale of the slow process is N, the condition $a < 1$ reflects that the fast process evolves much faster than the slow process, which is the condition expected for homogenization to hold.

However, our condition in Theorem 2.2 is the more stringent condition $a < 1/2$. To see why this condition appears, consider the following semimartingale decomposition of Q^N:

$$Q_v^N(t) - Q_v^N(0) = N^a \int_0^t \left(\lambda_v^\infty - \pi^{NQ^N(s)}(v)\right) \mathrm{d}s + (\text{martingale term}) + N^a \int_0^t \left(\sigma_v^N(s) - \pi^{NQ^N(s)}(v)\right) \mathrm{d}s.$$

The martingale term can be shown to vanish for $a < 1$, but we see that in order for the first term to also vanish we would need to show that the integral on the second line is $o(N^{-a})$: Proposition 3.4 shows that this term is $O(1/N^{1/2} + 1/N^{1-a})$ and so although it is $o(1)$ for $a < 1$, in order to have it $o(N^{-a})$ we need to assume that $a < 1/2$. Whether Theorem 2.2 continues to hold for $1/2 < a < 1$ constitutes in our view an interesting open problem, which also testifies to the difficulty of proving fully coupled stochastic averaging principles even in seemingly simple cases.

8.2. Two other scalings

Keeping $\varepsilon > 0$ as the distance to the stability region, we now discuss what happens on different space scales than the scale $N = \varepsilon^{-1/a}$ studied so far. To be more precise, we continue to consider arrival rates λ given by

$$\lambda = \lambda^\infty - \varepsilon\gamma$$

with $s(\lambda^\infty) = 1$, but now we consider the queue length process on the space scale $N = \varepsilon^{-1/a'}$ with $a' > 0$. Let N^b be the general time scale, and so consider the scaled processes

$$Q^N(t) = \frac{1}{N} Q(N^b t) \quad \text{and} \quad \sigma^N(t) = \sigma(N^b t), \ t \geq 0.$$

Assume for a moment that the stochastic averaging principle and state space collapse continue to hold: thus, determining the asymptotic behavior of Q^N reduces (at least informally) to understanding the asymptotic behavior of $s \circ Q^N$ under the homogenized dynamic. With the considered scaling, the homogenized generator is given by

$$L_{\mathrm{h}}^N f(q) = N^b \sum_{v \in V} \lambda_v^N \left(f\left(q + \frac{e^v}{N}\right) - f(q) \right) + N^b \sum_{v \in V} \pi^{Nq}(v) \mathbb{1}(q_v > 0) \left(f\left(q - \frac{e^v}{N}\right) - f(q) \right)$$

and so for $q > 0$ we have

$$\begin{aligned} L_{\mathrm{h}}^N s(q) &= N^{b-1} \sum_{v \in V} \lambda_v^N - N^{b-1} \sum_{v \in V} \pi^{Nq}(v) \\ &= N^{b-1} \left(1 - s(\gamma) N^{-a'}\right) - N^{b-1} \left(1 - \pi^{Nq}(0)\right). \end{aligned}$$

We have

$$\pi^{Nq}(0) = \frac{1}{1 + \sum_{v \in V} (Nq_v + 1)^a} \approx \frac{1}{N^a \left\| q \right\|_a^a}$$

and since the state space collapse assumption entails $\|q\|_a^a = s(q)^a/\mu$, we obtain

$$L_{\mathrm{h}}^N s(q) \approx -N^{b-1-a'} s(\gamma) + N^{b-1-a} \mu s(q)^{-a}.$$

We see that except when $a = a'$, which is the case studied so far, we cannot have both terms contributing in the limit: one dominates the other. In some sense, the space scale $N = \varepsilon^{-1/a}$ is the only one where we see in the limit at the same time the influence of the idleness induced by the distributed scheduling and the asymptotic drift term arising from the pre-limit processes being near-critical. More precisely, two cases arise:

Case $a' < a$**:** this is the space scale on which the near-criticality assumption dominates, the idleness induced by the distributed scheduling has no impact. In this case, the right time-scale is $b = 1 + a'$ and $Q^N \Rightarrow q$ with $q(t) \in I$ for all $t \geq 0$, and $s \circ q$ solution to $\dot{x} = -s(\gamma)\mathbb{1}(x > 0)$, i.e., $s(q(t)) = (s(q(0)) - s(\gamma)t)_+$;

Case $a' > a$**:** this is the reversed case: one only sees the idleness induced by the distributed scheduling, the limit is the same as the one obtained in Theorem 2.2 with $\gamma = 0$. In this case, the right time-scale is $b = 1 + a$ and $Q^N \Rightarrow q$ with $q(t) \in I$ for all $t \geq 0$ and $s \circ q$ solution to $\dot{x} = \mu x^{-a}$.

Except for controlling $s \circ Q^N$ after (potentially) hitting 0 in the case $a' < a$, these results can be established by making appropriate changes in the arguments developed above for $a' = a$. Actually, only minor changes are needed along the way. When $a' < a$ and $s(\gamma) > 0$, which is the only case where $s \circ q$ hits 0 in finite time, namely $s(q(0))/s(\gamma)$, $s \circ Q^N$ can be controlled after time $s(q(0))/s(\gamma)$ by coupling arguments that will be developed in [Cas].

8.3. Interchange of limits

Heavy traffic results are often investigated as a means to establish convergence of stationary distributions according to the well-known interchange of limits argument presented schematically in Figure 1. In our case, Q^N admits a stationary distribution $Q^N(\infty)$ in the subcritical case $s(\gamma) > 0$, and in this case we have $Q^N \Rightarrow q$ with $q(t) \in I$ for all $t \geq 0$ and $s \circ q$ solution to $\dot{x} = \mu x^{-a} - s(\gamma)$. Although we do not know how to solve this equation explicitly, it is readily seen that, when $s(\gamma) > 0$, the solution to this ODE converges to $\beta := (\mu/s(\gamma))^{1/a}$

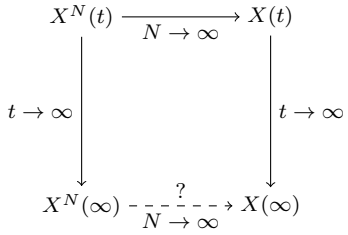

FIG 1. *Illustration of the interchange of limits argument: if $X^N \Rightarrow_N X$ and $X(t) \Rightarrow_t X(\infty)$, then provided technical assumptions (typically, tightness of $(X^N(\infty))$) we have $X^N(\infty) \Rightarrow_N X(\infty)$.*

as $t \to \infty$. At first, the interchange of limits of arguments seems therefore to suggest that $Q^N(\infty) \Rightarrow \beta$. The result being deterministic, this would be a rather unusual heavy traffic result. However, we do not know whether this reasoning applies because of the following argument.

If we start at $Q^N(0)$ with $s(Q^N(0)) \to \beta$, then $Q^N \Rightarrow q$ with q a constant function, which suggests to look at Q on a faster time scale than N^{1+a}. Indeed, it is thus conceivable that Q scaled differently converges to a non-constant, possibly random, limit. More precisely, the fact that $Q^N \Rightarrow q$ with q a constant function opens the possibility that $\widetilde{Q}^N \Rightarrow X$ with X a diffusion process, where

$$\widetilde{Q}^N(t) = \frac{1}{N} Q(N^b t),\ t \geq 0,$$

for some $b > 1 + a$. In this case, the interchange of limits would suggest that $Q^N(\infty) \Rightarrow X(\infty)$, provided X has a stationary distribution.

This argument is plausible because this is typically what happens when considering the near-critical case. For instance, a near-critical $M/M/1$ queue in the fluid regime converges to a constant function, but in the diffusive scale it converges to a positive recurrent Brownian motion. If one were only to consider the fluid limit and naively apply the interchange of limits principle, one would be led to conclude that the stationary distribution converges to a constant, which is not the case. Thus, the asymptotic behavior of Q^N stationary distribution constitutes in our view an intriguing open question.

8.4. Beyond a complete interference graph

What makes the case of a complete interference graph tractable is that all queues are of the same order of magnitude and remain away from 0 at all times, i.e., all coordinates are scaled by N and the limiting process q satisfies $\inf_{t \geq 0} q_v(t) > 0$ for every $v \in V$. We believe that the techniques developed in the present paper can be applied beyond the case of the complete interference graph as long as this property holds. For instance, they should be applicable to a square interference graph with four nodes $1, 2, 3, 4$ and edges $(1, 2)$, $(2, 3)$, $(3, 4)$ and $(1, 4)$ and equal

arrival rates $\lambda_v = 1/4$ at all nodes. Note that in this case, the stability condition is $\max(\lambda_1, \lambda_3) + \max(\lambda_2, \lambda_4) < 1$ so all λ's equal to $1/2$ is indeed critical.

However, the fact that all queues remain positive at all times now depends on the underlying interference graph and also on the arrival rates. If we take the above square interference graph with $\lambda_1 = \lambda_2 = \lambda_3 = 1/2$ but $\lambda_4 < 1/2$, then we believe that queue 4 will remain at 0. In this case, our techniques can no longer apply, especially the localization arguments that need all queue lengths to be bounded away from 0. We believe that in such cases, subtle behavior can arise and calls for new ideas.

To give a flavor of the kind of possible new behavior, consider three nodes on a line: the interference graph has three nodes $1, 2, 3$ and two edges $(1, 2)$ and $(2, 3)$. In this case, the 'outer' nodes 1 and 3 compete against the 'middle' node 2 to access the channel. The two maximal independent sets are $\{1, 3\}$ and $\{2\}$ with respective weight, for the Glauber dynamics,

$$\pi^q(\{1,3\}) = \frac{(q_1 q_3)^a}{1 + q_1^a + q_2^a + q_3^a + (q_1 q_3)^a}$$

and

$$\pi^q(\{2\}) = \frac{q_2^a}{1 + q_1^a + q_2^a + q_3^a + (q_1 q_3)^a}.$$

In the critical and symmetric case $\lambda_1 = \lambda_2 = \lambda_3 = 1/2$, we must have $\pi^q(\{1,3\}) = \pi^q(\{2\}) = 1/2$ in order for service to match arrivals, which imposes $q_1 q_3 = q_2$. This relation imposes a constraint on the product $q_1 q_3$ but not on the individual queues q_1 and q_3. In particular, if q_2 is of the order of N, then q_1 and q_3 will be much smaller, say $\sqrt{N}$ each. Different queues may thus live on different space scales, which suggests the necessity for a multiscale analysis.

References

[AC19] R. Atar and A. Cohen. Serve the shortest queue and Walsh brownian motion. *Ann. Appl. Probab.*, 29(1):613–651, 2019.

[Ald82] D. J. Aldous. Some inequalities for reversible Markov chains. *J. London Math. Soc. (2)*, 25(3):564–576, 1982.

[BBvL11] N. Bouman, S. Borst, and J. van Leeuwaarden. Achievable delay performance in CSMA networks. In *Proc. 49th Annual Allerton Conference*, pages 384–391, September 2011.

[BW14] M. A. A. Boon and E. M. M. Winands. Heavy traffic analysis of k-limited polling systems. *Probab. Engrg. Inform. Sci.*, 28(4):451–471, 2014.

[Cas] E. Castiel. Fluid limits for queue-based CSMA algorithms on the complete graph. In preparation.

[CBvW16] F. Cecchi, S.C. Borst, J.S.H. van Leeuwaarden, and P.A. Whiting. Mean-field limits for large-scale random-access networks. *arXiv*, 11 2016.

[CPR95] E. G. Coffman, Jr., A. A. Puhalskii, and M. I. Reiman. Polling systems with zero switchover times: a heavy traffic averaging principle. *Ann. Appl. Probab.*, 5(3):681–719, 1995.

[CPR98] E. G. Coffman, Jr., A. A. Puhalskii, and M. I. Reiman. Polling systems in heavy traffic: a Bessel process limit. *Math. Oper. Res.*, 23(2):257–304, 1998.

[Dob68] R. L. Dobrushin. The problem of uniqueness of a gibbsian random field and the problem of phase transitions. *Functional Anal. Appl.*, 2:302–312, 1968.

[DBBV15] Jan-Pieter L. Dorsman, Sem C. Borst, Onno J. Boxma, and Maria Vlasiou. Markovian polling systems with an application to wireless random-access networks. *Perform. Eval.*, 85–86:33–51, 2015.

[DSC96] P. Diaconis and L. Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Appl. Probab.*, 6(3):695–750, 1996.

[FPR10] M. Feuillet, A. Proutiere and P. Robert. Random capture algorithms: fluid limits and stability. In *Proc. Information Theory and Applications Workshop*, 1–4, 2010.

[FR14] M. Feuillet and P. Robert. A scaling analysis of a transient stochastic network. *Adv. in Appl. Probab.*, 46(2):516–535, 2014.

[FW84] M. I. Freidlin and A. D. Wentzell. *Random perturbations of dynamical systems*, volume 260 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, 1984.

[GBW14] J. Ghaderi, S. C. Borst, and P. A. Whiting. Queue-based random-access algorithms: fluid limits and stability issues. *Stoch. Syst.*, 4(1):81–156, 2014.

[GS10] J. Ghaderi and R. Srikant. On the design of efficient CSMA algorithms for wireless networks. In *proceedings of IEEE Conference on Decision and Control (CDC)*, pages 954–959, 2010.

[Har95] J. M. Harrison. Balanced fluid models of multiclass queueing networks: a heavy traffic conjecture. *Institute for Mathematics and Its Applications*, 71:1, 1995.

[HR81] J. M. Harrison and M. Reiman. Reflected Brownian motion on an orthant. *Ann. Probab.*, 9(2):302–308, 1981.

[HW96] J. M. Harrison and R. Williams. A multiclass closed queueing network with unconventional heavy traffic behavior. *Ann. Appl. Probab.*, 6(1):1–47, 1996.

[HK94] P. J. Hunt and T. G. Kurtz. Large loss networks. *Stochastic Process. Appl.*, 53(2):363–378, 1994.

[Jen10] O. B. Jennings. Averaging principles for a diffusion-scaled, heavy traffic polling station with K job classes. *Math. Oper. Res.*, 35(3):669–703, 2010.

[JW08] L. Jiang and J. Walrand. A distributed CSMA algorithm for throughput and utility maximization in wireless networks. In: *Proc. Allerton '08 Conf.*, 2008.

[Kur92] T. G. Kurtz. Averaging for martingale problems and stochastic

approximation. In *Applied stochastic analysis (New Brunswick, NJ, 1991)*, volume 177 of *Lect. Notes Control Inf. Sci.*, pages 186–209. Springer, Berlin, 1992.

[Kru11] Ł. Kruk. An open queueing network with asymptotically stable fluid model and unconventional heavy traffic behavior. *Math. Oper. Res.*, 36(3):538–551, 2011.

[LN13] M. J. Luczak and J. R. Norris. Averaging over fast variables in the fluid limit for Markov chains: application to the supermarket model with memory. *Ann. Appl. Probab.*, 23(3):957–986, 2013.

[LP17] D. A. Levin and Y. Peres. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2017.

[Puh15] A. L. Puha. Diffusion limits for shortest remaining processing time queues under nonstandard spatial scaling. *Ann. Appl. Probab.*, 25(6):3381–3404, 2015.

[PW13] O. Perry and W. Whitt. A fluid limit for an overloaded X model via a stochastic averaging principle. *Math. Oper. Res.*, 38(2):294–349, 2013.

[RSS09] S. Rajagopalan, D. Shah, and J. Shin. Network adiabatic theorem: An efficient randomized protocol for contention resolution. In *proceedings of SIGMETRICS/Performance*, volume 37, pages 133–144, 2009.

[Rei84] M.I. Reiman. Open queueing networks in heavy traffic. *Math. Oper. Res.*, 9 (3), pages 441–458, 1984.

[Rei05] M.I. Reiman. Some diffusion approximations with state space collapse. In: F. Baccelli, G. Fayolle (eds.), *Modelling and Performance Evaluation Methodology*, Lecture Notes in Control and Information Sciences, Vol. 60, Springer, pages 207–240, 2005.

[SBB14] F. Simatos, N. Bouman, and S. C. Borst. Lingering issues in distributed scheduling. *Queueing Syst.*, 77(2):243–273, 2014.

[SC97] L. Saloff-Coste. Lectures on finite Markov chains. In *Lectures on probability theory and statistics (Saint-Flour)*, volume 1665 of *Lecture Notes in Math.*, pages 301–413. Springer, Berlin, 1997.

[SS12] D. Shah and J. Shin. Randomized scheduling algorithm for queueing networks. *Ann. Appl. Probab.*, 22(1):128–171, 2012.

[SST11] D. Shah, J. Shin, and P. Tetali. Medium access using queues. In: *Proc. FOCS 2011 Conf.*, 2011.

[Sto04] A. L. Stolyar. Maxweight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. *Ann. Appl. Probab.*, 14(1):1–53, 2004.

[TE90] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. In *Proc. CDC '90*, volume 4, pages 2130–2132, 1990.

[vdBS94] J. van den Berg and J. E. Steif. Percolation and the hard-core lattice gas model. *Stochastic Process. Appl*, 49(2):179–19, 1994.

[vdM07] R. D. van der Mei. Towards a unifying theory on branching-type

polling systems in heavy traffic. *Queueing Syst.*, 57(1):29–46, 2007.

[YYSE12] S.-Y. Yun, Y. Yi, J. Shin, and D. Y. Eun. Optimal CSMA: A survey. In *Proc. ICCS '12*, pages 199–204, Nov 2012.