Study of QB-CSMA algorithms

Eyal Castiel

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Institut Supérieur de l'Aéronautique et de l'Espace (ISAE)*

Présentée et soutenue le *5 Décembre 2019* **par :**

EYAL CASTIEL

Study of QB-CSMA algorithms, étude des protocoles d'accès CSMA-QB

JURY

COUTIN LAURE	Professeur d'Université	Présidente du jury
GAST NICOLAS	Chargé de Recherche	Examinateur
KRUK ŁUKASZ	Professeur Associé	Rapporteur
MALRIEU FLORENT	Professeur d'Université	Rapporteur
MICLO LAURENT	Directeur de recherche	Directeur de Thèse
ROBERT PHILIPPE	Directeur de Recherche	Examinateur
SIMATOS FLORIAN	Enseignant Chercheur	Directeur de Thèse

École doctorale et spécialité :
MITT : Domaine STIC : Réseaux, Télécoms, Systèmes et Architecture
Unité de Recherche :
DISC (Département Ingénierie des Systemes Complexes)
Directeur(s) de Thèse :
Florian SIMATOS et *Laurent MICLO*
Rapporteurs :
KRUK Łukasz et *MALRIEU Florent*

Contents

0.1 Introduction iii

0.2 Elements techniques, Définitions viii

0.3 Equation de Poisson xiii

0.4 Limites fluides xix

0.5 Charge lourde xxii

0.6 Conclusion, perspectives de recherche xxvi

1 Introduction **1**

1.1 General discussion 1

1.1.1 Motivation 1

1.1.2 Wireless networks modeling 2

1.2 Main algorithms 4

1.2.1 Max-Weight 5

1.2.2 Greedy maximal scheduling, Longest Queue First 6

1.2.3 CSMA-type algorithm 7

1.2.4 Polling 12

1.2.5 Load balancing algorithms 12

1.3 Homogenization and State Space Collapse 13

1.3.1 Stochastic Averaging 13

1.3.2 Poisson equation and Stein's method 17

1.3.3 State Space Collapse, Skorokhod problem 19

1.4 Functional limit theorems 20

1.4.1 Fluid limits 22
1.4.2 Heavy traffic 24

2 Technical elements **27**

2.1 Introduction 27
2.1.1 General notations 27
2.1.2 Model description 30
2.1.3 Localization 33
2.2 Functional analysis 33
2.2.1 General notions 34
2.2.2 Poisson equation 36
2.3 Glauber dynamics for QB-CSMA 38
2.3.1 Historical results 38
2.3.2 Spectral gap of Glauber dynamics for QB-CSMA 39
2.3.3 Influence of threshold function 44

3 Homogenization through Poisson equation **49**

3.1 Main result 49
3.2 Proof of the main result 51
3.2.1 Control in terms of solutions to the Poisson equation 51
3.2.2 Control of solutions to the Poisson equation 55
3.3 Scaled process 57
3.3.1 General considerations 57
3.3.2 Fluid limits in a complete interference graph 59

4 Fluid Limits of QB-CSMA **63**

4.1 Context 63
4.2 Main results and heuristic 64
4.2.1 Limiting process: general interference graph 64
4.2.2 Identification of the limit: Complete interference graph . . . 66

4.2.3 Main result . . . 68
4.2.4 Heuristic, line network . . . 69
4.3 Preliminary . . . 73
4.3.1 Localization and homogenization . . . 75
4.3.2 Proof of preliminary results . . . 76
4.4 General interference graph up to $\tau^0(q^*)$, Theorem 4.5 . . . 80
4.5 Complete interference graph . . . 83
4.5.1 Proof of Theorem 4.6 . . . 83

5 Heavy traffic of QB-CSMA in a complete interference graph 89
5.1 Intuition and discussion . . . 90
5.1.1 Fluid limits in the critical case, State space collapse . . . 90
5.1.2 Idleness in random-access settings . . . 93
5.1.3 Nonstandard behavior . . . 94
5.1.4 Beyond $a < \frac{1}{2}$ . . . 94
5.1.5 Initial state, limiting ODE . . . 95
5.2 Main result . . . 95
5.3 Notation and main steps of the proof . . . 97
5.3.1 General notations . . . 97
5.3.2 Localization, constants . . . 98
5.3.3 Distance to I . . . 98
5.3.4 Main steps . . . 99
5.4 State space collapse . . . 101
5.4.1 Proof of $\mathbb{E}(\Xi^N) \to 0$ . . . 103
5.5 Proof of main result . . . 104
5.5.1 First step: convergence of the sum to S . . . 104
5.5.2 Second step: proof of Theorem 5.4 . . . 107

6 Conclusion 109
6.1 Generalizing the fluid limits result . . . 109

6.1.1 How/when do coordinates of solutions to the ODE reach 0? . 110
6.1.2 Homogenization when touching 0 111
6.2 Generalizing the Heavy traffic result 112
6.2.1 Additional time scales, State space collapse 112
6.2.2 Diffusive time scale 114
6.3 Generalizing the homogenization result 115
6.3.1 Metastability phenomenon 116
6.3.2 Large deviations 116

A Asymptotic approximation of service rates **119**

B Positivity of queue lengths for small times **121**

B.1 General remarks 121
B.2 Main steps of the proof 122
B.3 Proofs using coupling 128

Bibliography **144**

Remerciements

Je tiens pour commencer cette section à remercier tout particulièrement mes direteurs de thèses Florian Simatos et Laurent Miclo. Vous m'avez tous les deux à votre manière fournis un support sans lequel je n'aurai pas réussi à finir ma thèse. Vos encadrements m'ont apporté beaucoup et il me tarde de mettre en action vos enseignements. Je souhaire également remercier les membres de mon jury de thèse: Łukasz Kruk et Florent Malrieu qui ont accepté de rapporter ma thèse. Vos commentaires sur le manuscrit ont contribué à sont amélioration et ont ajouté à sont interet. Je souhaite également remercier la présidente du jury Laure Coutin et les examinateurs Philipe Robert et Nicolas Gast pour leurs questions lors de ma soutenance. Viennent ensuite ma famille, parents, frère, soeur, cousine, cousin, tante, oncle, etc sans qui je n'aurai certainement pas réussi à en arriver là. Vous qui allez être les seuls à lire ce paragraphe de ma thèse (avec mes parents): mes collègues doctorants & doctorantes (vous savez qui vous êtes pas besoin de faire une liste hein...). Je souhaite également remercier mes professeurs (secondaire et supérieurs) qui ont su me donner le gout des mathématiques, et plus préciséments des probabilités. On citera par exemple Sandy Souchet, Thomas Duquesne, Nicolas Fournier, Philipe Robert (bis), Irina Kourkova, Bartek Blasczizyn, Jean-Marc Bardet, M. Bartolli et bien d'autres. Je tiens aussi à remercrier les equipes de l'ISAE et de l'IMT qui m'on acceuillis ainsi que le personel là bas (en particulier Odile). Ils m'ont permis d'entreprendre ma thèse dans de bonnes conditions. Pendant ces années d'études j'ai également rencontré des gens à qui j'ai envie d'adresser des remerciements: par exemple lou, pedro, karine, carmen, mattia, louise et d'autres qui ne se vexerons sûrement pas si je ne les mets pas dans la liste...

There are also people I want to thank in english: first I would like to reiterate my thanks to Łukasz Kruk for his comments on the manuscript and questions at the defense. I stongly believe those help strengthen the value of the thesis. I would also like to thank Sem Borst who hosted me for three month in TU/E, those were some interesting times with Phil Whiting as well. I also want to thank Rajesh Sundaresan and Vivek Borkar for the time they hosted me in Bangalore and Mumbay. Those three visits were quite insightful and I hope our relationships will continue to grow. I would also like to thank Jan-Pieter Dorsman for the advice he gave me. Finally, I want to thank Rami Atar for accepting me as a Post-doc. I am sure this new relationship will flourish to provide beautiful fruits.

I would like to thank Robin H., Isaac A., Melinda S., George M. & O., Joanne R., John T., Dan S., Ursula G., Philip P. & D., Erik H., Bernard W., Arthur C., Pierre B., Christopher P. , Frank H., Jules V., Aldous H., Ernest C., Denis T. and many others...

To conclude, I thank you [Reader] for whatever portion of this manuscript you will read. I hope you find here what you are looking for, feel free to contact me should you have any question on this manuscript.

Synthèse en français

0.1 Introduction

0.1.0.1 Motivations

Dans le monde d'aujourd'hui, les personnes et objets tendent à être de plus en plus connecté à travers les communications sans fil. L'énorme quantité de données échangées sur ces réseaux requiert des algorithmes des communications plus rapides, plus fiables et plus flexibles. Le fait que les utilisateurs d'un réseau sans fil partagent l'air comme moyen de communication crée des difficultées dues à des problèmes d'interférence. Des utilisateurs proches géographiquement ne peuvent utiliser le réseau simultanément sans mélanger leurs messages, les rendant illisibles. Ce genre de problème d'accès à une ressource partagée, populaire en probabilité appliquée depuis au moins 50 ans, a de nombreuses applications des centres d'appels, aux accès Wi-Fi et réseaux pair-à-pair.

De manière générale, un grand nombre de questions associées à des algorithmes de communication peuvent être formuler de la manière suivante: "Comment allouer la ressource partagée?", "Comment mesurer l'efficacité/optimalité dans un algorithme?", "Est-ce que cet algorithme est efficace/optimal?", "Quel est le comportement dans telle condition?", etc... La théorie des files d'attente donne un cadre formel pour répondre à ces questions. Par exemple en établissant des garanties théoriques pour des indicateurs de performances dûment définies. Les utilisateurs du réseau sont modélisés par des files d'attente avec des flux d'arrivés. Chaque file possède un espace d'attente où stocker les requêtes avant de pouvoir les servir. Les files seront aussi appelées nœuds , utilisateurs, terminaux ou serveurs en fonction du contexte.

Laisser une autorité centrale prendre les décisions en fonction de l'état actuel du réseau conduit en général à de meilleures performances si l'information est utilisée de manière adéquate. D'un autre côté, une autorité centrale induit des désavantages. Une attaque sur une autorité centrale peut causer une panne généralisée du réseau. De plus, recueillir des informations exactes sur le réseau entier est un labeur pénible et hasardeux, surtout quand le nombre de terminal est élevé. Plus précisément, dans le cas d'un réseau Wi-Fi pour smartphones, les nœuds entrent et sortent du réseau de manière continue sans heurts, ce qui complique la collecte d'informations car celle-ci peut être incomplète ou datée. D'un autre côté, on

pourrait préférer laisser les utilisateurs prendre les décisions en fonction de leurs environnements dans le réseau. Il est plus facile pour les utilisateurs d'évaluer leurs voisinages géographique respectifs de manière précise. Il s'agit ensuite d'utiliser cette information à des fins d'ordonnancement de manière distribuée, c'est-à-dire en laissant chaque terminal prendre ses décisions en fonction de l'information locale. Cela peut mener à des comportements "gloutons" pour les utilisateurs et de l'inefficacité pour tout le monde. De leurs points de vue, les utilisateurs veulent utiliser le plus de ressource possible pour minimiser leurs temps d'attente. Ce genre de comportement pénalise les autres utilisateurs en limitant leurs accès et peut mener à une utilisation inefficace de la ressource. Le réseau étant plus encombré qu'il ne pourrait l'être, cela entraine des temps d'attente moyens plus longs pour tout le monde. Une question cruciale dans ce domaine de recherche est de trouver un algorithme distribué ayant pour mission de partager la ressource entre les utilisateurs de manière efficace et juste. Les algorithmes distribués sont plus appropriés à la nature changeante des réseaux et sont directement applicable à des réseaux de grande taille.

0.1.0.2 Etat de l'art

Ce travail de rédaction a également été accompagné par une recherche bibliographique dont on rend compte des grandes lignes ici. On distingue trois axes pour exposer nos références: le premier est la littérature reliée aux algorithmes d'accès à une ressource partagée, le second est la littérature orientée vers les problèmes d'homogénéisation, le troisième est les résultats de limites fonctionels types limites fluides et charge lourde.

Pour le premier axe, on se restreint dans cette synthèse à des problème d'ordonnancement sur un graphe d'interférence $G = (V, E)$: les nœuds du graphes sont les utilisateurs du réseau et une arête entre deux nœuds représente l'impossibilité pour eux de transmettre en même temps voir 1.1.2 pour d'autre exemples de problèmes et 0.2.0.2 pour plus de détails sur S l'ensemble des décisions admissibles. On commence par une définition de l'algorithme Max-Weight. Cet algorithme introduit dans [TE92] par Tassiulas et Ephremides a été l'un des premier avec des garanties théoriques sur ses performances. C'est un algorithme à temps discret tel que quand les files d'attente sont dans l'état q, la decision de service est donnée par

$$\sigma \in \operatorname*{argmax}_{\rho \in S} \sum_{v \in V} \rho_v f(Q_v(t)),$$

pour f croissante et S l'ensemble de décisions admissibles. L'étude de l'algorithme Max-Weights est riche et encore vivante près de 30 ans après son début: dans [TE92] les auteurs prouvent l'optimalité de la région de stabilité:

$$\Lambda_{\mathrm{MW}}(S) \coloneqq \left\{\lambda \in \mathbb{R}_+^V,\ (Q(t))_{t \geqslant 0} \text{ avec taux d'arrivé } \lambda \text{ est un processus de Markov ergodique}\right\},$$

en prouvant que cette région peut se réécrire

$$\Lambda^*(S) \coloneqq \{\lambda \in [0,1]^V \mid \exists \rho \in Co(S), \lambda < \rho \text{ coordoné par coordoné }\}.$$

Plus récement, [MS16] établit l'optimalité de cet algorithme au sens du délai à l'équilibre dans un régime de charge lourde.

On mentionne ensuite les algorithmes CSMA encore utilisé aujourd'hui en télécomunication. Dans cet algorithme, chaque nœud a une "fugacité" inhérente qui peut être vu comme un taux d'activation instantané. Ce taux reste fixé. Une fois activé, un nœud reste actif pour une durée exponentielle de parametre $\nu_v \in (0,1)$. Un nœud inactif écoute le canal pour vérifier si il y aurai des problèmes d'interférences avec sa communication. Si il n'y en a pas, il s'active au bout d'une durée exponentielle de paramètre $1-\nu_v$. On présente finalement les algorithme QB-CSMA introduit par Rajagopolan, Shah et Shin dans [RSS09]. On refère à [Yun+12] pour une liste de références sur le sujet. Ces algorithmes ont étés introduits avec l'objectif d'approximer les décisions de service de l'algorithme Max-Weight de manière distribué. La manière dont cela est achevé est astucieuse: d'une manière schématique, on utilise la même procédure que CSMA avec des fugacité données par des fonctions d'activations

$$\Psi_+^v(q) = \frac{e^{f(q_v)}}{1+e^{f(q_v)}}, \quad \text{et } \Psi_-^v(q) = 1-\Psi_+^v(q),$$

avec f paramètre de l'algorithme. Grace à ce choix, la mesure invariante de l'ordonanceur se concentre exponentiellement sur les ensemble stables de poid maximum:

$$\pi^q(\sigma) \propto \exp(\sum_{v\in V} \sigma_v f(q_v)),$$

à une constante de renormalisation près. Il a été prouvé dans [SS12] que cet algorithme a une region de stabilité optimale dès que f croît assez lentement mais on expose dans la Section 1.2.3.3 une conjecture (confirmée par certains résultats commes ceux de [BBL11]) comme quoi f croîssant rapidement peut améliorer le délai, dans une certaine mesure. Se reporter à la Section 1.2.3.3 pour plus de détails. Voir Section 1.2 pour plus une présentation plus fournie de ces algorithmes et d'autres non mentionés ici. L'idée de cette thèse est d'étudier QB-CSMA avec f croissant le plus vite possible en conservant la propriété de stabilité maximum prouvée dans [SS12].

Pour le deuxième axe, on parle prinicpalement d'un principe d'homogénéisation. L'étude d'un système complexe est souvent extremement compliqué à décrire, d'autant plus quand l'espace d'état est grand. Une manière de simplifier l'analyse est de partir du principe qu'une des composante évolue sur une échelle de temps différente des autres. On reprend ici l'exemple du Chapitre 10 de [Gri14]: partons d'un pendule suspendu à une base mobile. Si la base évolue à une vitesse comparable ou plus grande à la vitesse du bout du pendule, les effets d'élan vont rendre l'analyse particulièrement pénible. Si au contraire la base évolue lentement par rapport aux vas et viens du pendule on pourra la considérer comme immobile pour étudier les oscilations de celui ci. L'étude rigoureuse de tels comportements est appellée dans la litérature"homogénéisation" ou "séparation d'échelle de temps". Nous n'avons pas pu appliquer directement les méthodes présentes dans la littérature, et nous avons du coup développé une nouvelle méthode. On mentionne ici la méthode de Freidlin et Wentzell [FW12] se basant sur la théorie des perturbations, la méthode de [PSV77] developée dans [Kur92] basée sur une approche de martingales et la méthode de Luczak et Norris [LN13]. Cette dernière méthode est la plus proche de la notre avec une utilisation de "fonctions correctrices" proche des solutions aux equations de Poissons. Notre méthode est cependant mieux adaptée à des processus de Markov. L'homogénisation s'occupe de modèles variations autours d'un

processus de Markov (Q^N, σ^N) de générateur

$$L^N[f](q,\sigma) = L^{\sigma}_{\mathrm{s},N}[f(\cdot,\sigma)](q) + NL^{q}_{\mathrm{f},N}[f(q,\cdot)](\sigma).$$

Il est séparé en deux parties: $L^q_{\mathrm{f},N}$ agit sur les fonctions de σ mais dépend de la valeur de q et $L^{\sigma}_{\mathrm{s},N}$ agit sur les fonctions de q mais dépend de la valeur de σ. Avec cette hypothèse, on contraint Q et σ à ne pas changer en même temps mais cette hypothèse peut être relachée dans une certaine mesure. Le générateur $L^{\sigma}_{\mathrm{s},N}$ est considéré *lent* quand on le compare à $NL^q_{\mathrm{f},N}$ où les transitions se passent sur une echelle de temps bien plus rapide quand $N \to +\infty$. L'idée générale est que σ évolue si vite que Q n'interagit avec σ qu'à travers $\pi^{N,Q}$ la probabilité invariante de $L^Q_{\mathrm{f},N}$. Le but est de comparer la dynamique du processus de générateur L^N avec celle du processus de Markov de générateur

$$L_{\mathrm{h},N}[f](q) = \pi^{N,q}[L^{\cdot}_{\mathrm{s},N}[f](q)], \text{ quel que soit } q \in \mathbb{N}^V,$$

avec $\pi^{N,q}[f]$ la probabilité invariante de $L^q_{\mathrm{f},N}$. On renvoie le lecteur vers la Section 1.3 pour plus d'informations, une description plus détaillée des méthodes existentes et références sur ce sujet.

Finalement, on aborde le troisième axe avec les théorèmes de limite fonctionelle. L'idée de ces résultats est de fournir un équivalent fonctionels à la loi des grand nombre et au théorème de la limite centrale. Soit $(X_N)_{N\in\mathbb{N}}$ une suite de variable i.i.d. La loi des grands nombres (voire [Kal02], Théorème 3.23) nous dit que tant que $\mathbb{E}[|X_1|] < +\infty$,

$$\frac{1}{N}\sum_{k=1}^{N} X_k \to \mathbb{E}[X_1] \text{ presque sûrement quand } N \to +\infty.$$

De la même manière, si $X(t)$est l'interpolation linéaire de

$$X(K) \coloneqq \sum_{k=1}^{K} X_k.$$

Quel que soit $t > 0$,

$$\frac{X(Nt)}{N} \to t\mathbb{E}[X_1] \text{ presque sûrement quand } N \to +\infty.$$

En fait $(X(Nt)/N)_{t\geqslant 0}$ convergege presque sûrement pour la topologie de la convergence uniforme. On définit la convergence uniforme sur les compactes par le fait que quel que soit $T < +\infty$,

$$\sup_{t\leqslant T}\left|\frac{X(Nt)}{N} - \mathbb{E}[X_1]\,t\right| \to 0 \text{ presque sûrement quand } N \to +\infty.$$

De la même manière, si on prend un échantillon N qui croît et on renormalise la moyenne empirique par N, on a convergence vers une limite deterministe. Ainsi quel que soit $0 < T < +\infty$,

$$\sup_{t\leqslant T}|\frac{X(Nt)}{N}| \to 0 \text{ presque sûrement quand } N \to +\infty.$$

Le fait que la limite soit nulle suggère que l'on devrait prendre plus de termes pour avoir une limite non triviale. Le deuxième résultat important est le théorème de la limite centrale, voire Proposition 5.9 de [Kal02]. Soit $(X_N)_{N\in\mathbb{N}}$ une suite de variables aléatoires i.i.d. de moyenne 0 et variance 1. Alors

$$\frac{1}{N}\sum_{k=1}^{N^2} X_k \Rightarrow \mathcal{N},$$

avec $\mathcal{N}$ loi normale standard et $\Rightarrow$ la convergence en distribution, définie dans la Définition 2.3. Une conséquence directe est que quel que soit $t \geqslant 0$,

$$\frac{X(N^2 t)}{N} \Rightarrow \mathcal{N}',$$

avec $\mathcal{N}'$ loi normale avec moyenne 0 et variance t. Encore une fois, on peut renforcer ce résultat avec une convergence uniforme sur les compactes vers un mouvement brownien:

$$(\frac{X(N^2 t)}{N})_{t\geqslant 0} \to B \text{ en distribution,}$$

avec B un mouvement brownien.

L'idée est d'utiliser le même genre de renormalisation pour des problèmes de file d'attente pour obtenir des approximations de premier et second ordre. Les methodes de limites fluides ont plus de quarantes ans: on cite par exemple [MM79] où les auteurs prouvent les propriété de récurrence/transience de marches aléatoires sur $\mathbb{Z}^2$ et $\mathbb{Z}^3$ par des méthodes de limites fluides. Celles ci ont été utilisé de manière extensive pour répondre à des questions de stabilité depuis. On cite par exemples le chapitre 9 de [Rob03] pour une introduction sur le sujet. Voir également deux articles fondamentaux traitant de questions de stabilité: [RS92] et [Dai95]. Voir Section 1.4.1 pour une description substantielle de ces articles. Les limites fluides de QB-CSMA ont également été étudiées mais dans un cas différent de celui présenté dans cette thèse. Dans [GBW14], les auteurs s'interessent au cas d'activation/deactivations polynomiale dans la taille des files avec une puissance assez grande. Dans ce cas, ils prouvent que pour un certain choix de graphe d'interférence, les limites fluides de QB-CSMA se comportent comme l'algorithme RCA de [FPR10] (et les deux algorithmes ont donc la même région de stabilité). Avec cette procédure, un nœud reste actif jusqu'au moment où il est (presque) vide. Les auteurs de [GBW14] prouvent que pour certains graphes d'interférences, ce type d'algorithme n'atteint pas la région de stabilité optimale. Les résultats des Chapitre 4 et 5 se placent dans le complémentaire de ce résultat: on donnera les limites fluides de QB-CSMA quand la puissance dans les taux polynomiaux est assez petit. Cela change completement le comportement asymptotique qui devient la solution d'une EDO.

L'idée d'évaluer les performance d'un algorithme en étudiant son comportement en présence d'une charge lourde n'est également pas nouvelle. Il y a près de 60 ans, Kingman s'interroge sur l'évolution du délai à l'équilibre lors ce que l'utilisation d'un réseau "simple" tend vers 100% dans les articles [Kin61] et [Kin62]. On renvoie vers [Whi02] pour une revue des résultats standards de charge lourde. Le résultat présenté dans cette thèse est inhabituel ne rentre pas dans la catégorie de résultat standard. On renvoie vers la Section 1.4.2 pour plus de détails et des références vers des résultats de charge lourde non standards.

0.2 Elements techniques, Définitions

Dans cette partie, on présente les définitions discussions et résultats préliminaires du Chapitre 2.

0.2.0.1 Définitions

On commence par des notations utilisées tout au long de cette thèse et synthèse. Ces définitions et d'autres utilisée dans le manuscrit sont définies dans la Section 2.1.1. Soit V un ensemble de $n < \infty$ nœuds . La pseudo-norme L_b usuelle sur $\mathbb{R}^V$ est notée $\|\cdot\|_b$: pour n'importe quel $x \in \mathbb{R}^V$, et $b \geq 0$, elle est définie par $\|x\|_b = (\sum_{v\in V} x_v^b)^{1/b}$. De la même manière, pour $b > 0$ et $q \in \mathbb{R}_+^V$ on utilise $s_b(q)$ pour écrire $\sum_{v\in V} q_v^b$. Soit $G = (V, E)$ un graphe simple non orienté et soit $S(G)$ l'ensemble des ensembles stables de G. La définition formelle de $S(G)$ est donnée dans la prochaien section. On utilise la notation $\mathbf{0}$ pour dénoter l'ensemble stable vide et la configuration où toutes les files sont vides ($\mathbf{0} = 0_{\mathbb{R}^V}$).

Quel que soient $n > 0$, $f : \mathbb{R}^n \to \mathbb{R}$ et $U \subset \mathbb{R}^n$, on note $\|f\|_{\infty,U}$ la norme uniforme sur U: $\sup_{x\in U} \mid f(x) \mid$. Quand $U = \mathbb{R}^n$, on utilise la notation $\|f\|_\infty$. Avec un leger abus de notation, on utilise la même dénomination pour la norme uniforme de vecteurs de $\mathbb{R}^n$: pour $x \in \mathbb{R}^n$, $\|x\|_\infty = \max_n |x_n|$. Pour $g : \mathbb{R} \to U$, On note le temps de sortie de U pour g de la manière suivante:

$$\bar{\tau}^U(f) = \inf\{t > 0,\, g(t) \notin U\}\,.$$

Soit $f : \mathbb{R}^n \to \mathbb{R}$ régulière, quel que soient $i, j \leqslant n$ on note ∂_i sa dérivé partielle selon q_i et ∂_{ij}^2 sa dérivé seconde selon q_i et q_j, i.e.,

$$\partial_i f = \frac{\partial f}{\partial q_i} \quad \text{et} \quad \partial_{ij}^2 f = \frac{\partial^2 f}{\partial q_i q_j}.$$

Tout au long du manuscrit, on utilise la lettre C pour noter une constante positive finie dont la valeur exacte n'influe pas fondamentalement les résultats et dont la valeur peut changer d'une ligne à l'autre. On ne l'autorise pas à dépendre de quantité qui vont évoluer avec la renormalisation (comme N) mais peut dépendre de λ, a, constantes de localisations dans la définition de U, etc...

Pour un processus de Markov à temps continue, l'opérateur crucial est le générateur L^0 muni de son domaine $\mathcal{D}(L^0)$. On définitégalement ici le carré du champs de l'opérateur L^0 dont l'importance est expliquée dans la Proposition 2.1/0.1 redonné plus bas. Soit L^0 le générateur d'un processus de Markov à espace d'état discret

$$L^0[f](x) = \sum_{x'\neq x} L^0(x, x')\,(f(x') - f(x))\,.$$

Le carré du champs de cet opérateur est donné par

$$\Gamma^0[f](x) = L^0[f^2](x) - 2f(x)L^0[f](x).$$

On peut réécrire cette expression par

$$\Gamma^0[f](x) = \sum_{x' \neq x} L^0(x, x') \left(f(x') - f(x)\right)^2 .$$

Proposition 0.1
Soit L^0 générateur de $(l_t)_{t \geqslant 0}$ un processus de Markov non explosif, quelle que soit f telle que f et $f^2 \in \mathcal{D}(L^0)$ le processus

$$M_f(t) = f(l_t) - f(l_0) - \int_0^t L^0[f](l_s)\mathrm{d}s$$

est une martingale locale avec processus croissant prévisible

$$\langle M_f \rangle (t) = \int_0^t \Gamma^0[f](l_s)\mathrm{d}s.$$

Respectivement, $(l_s)_{s \geqslant 0}$ est le seul processus tel que

$$f(l_t) - f(l_0) - \int_0^t L^0[f](l_s)\mathrm{d}s$$

est une martingale quelle que soit f telle que f et $f^2 \in \mathcal{D}(L^0)$.

On définitégalement ici la convergence en probabilité:

Définition 0.2
Soient $(X^N)_{N \in \mathbb{N}}$ et Y des variables aléatoires à valeur dans le même espace métrique $\mathcal{X}$. On dit que X^N converge vers Y en probabilité, noté $X^N \xrightarrow{\mathbb{P}} Y$, si

$$\mathbb{P}\left(|X^N - Y| \geqslant \epsilon\right) \to 0 \text{ quand } N \to +\infty.$$

Soit π une mesure de probabilité et f une fonction sur le même espace de définition $\mathcal{X}$. On utilise la notation $\pi[f]$ pour décrire l'action de π sur f:

$$\pi[f] = \int_{\mathcal{X}} f\mathrm{d}\pi.$$

0.2.0.2 Déscription du model

On présente ici une déscription succinte du modèle étudié dans cette thèse.

Soit V un ensemble fini de n nœuds. Chaque nœud est muni d'une file $M/M/1$ avec politique "premier arrivé premier servi" et vacations dues aux interférences. Les taux d'arrivés sont notés dans un vecteur V-dimensionel λ. Le processus $Q_v(t) \in \mathbb{N}$ compte le nombre de requête en attente au serveur v et au temps t et $\sigma(t) \in \{0,1\}^V$ represente le vecteur d'activité instantané: le serveur v est actif et traite une requête (si il y en a une) à taux 1 quand $\sigma_v(t) = 1$ et attend son tour sinon.

Les nœuds sont placés sur un graphe simple non dirigé $G = (V, E)$. Une arête entre deux nœuds indique qu'ils ne peuvent pas être actifs en même temps. On

utilise le signe $\sim$ pour signifier l'existence d'une arête entre deux nœuds ($v \sim w \Leftrightarrow \{v, w\} \in E$). Une configuration admissible de taux de service est un ensemble stable de G, un élément de $S(G)$ définitici:

$$S(G) \coloneqq \{\sigma \in \{0,1\}^V \mid v \sim w \Rightarrow \sigma_v + \sigma_w \leqslant 1\}.$$

Etant donné l'état de l'ordonnanceur σ, le processus de files d'attente Q évolue comme n files $M/M/1$ indépendantes avec taux d'arrivés λ et taux de départ σ. D'un autre côté, σ évolue également: étant donné le processus de file d'attente Q, un nœud actif v se désactive à taux $\psi_-(Q_v)$ pour une fonction de désactivation ψ_-. De la même manière un nœud inactif s'active à l'aide d'une fonction d'activation ψ_+ quand aucun de ses voisins n'est actif.

D'une manière plus formelle, (Q, σ) est un processus de Markov sur $\mathbb{N}^V \times \{0,1\}^V$ de générateur L qui peut être décomposé en une somme de deux générateurs:

- le générateur L_{s}^{σ} du processus *lent* Q dont la dynamique dépend de σ,
- le générateur L_{f}^{q} du processus *rapide* σ dont la dynamique dépend de Q.

La terminologie *rapide* et *lent* vient de l'homogénéisation, voir Section 1.3.1 pour plus de détails. Le générateur L de (Q, σ) agi sur les fonctions $f : \mathbb{N}^V \times S(G) \to \mathbb{R}$ de la manière suivante:

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

avec

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

et

$$\begin{aligned} L_{\mathrm{f}}^{q}[h](\sigma) = &\sum_{v \in V} \sigma_v \Psi_-(q_v) \left(h(\sigma - e^v) - h(\sigma)\right) \\ &+ \sum_{v \in V} \prod_{w \sim v} (1 - \sigma_w)(1 - \sigma_v) \Psi_+(q_v) \left(h(\sigma + e^v) - h(\sigma)\right) \end{aligned} \tag{2}$$

avec $g : \mathbb{N}^V \to \mathbb{R}$ et $h : S(G) \to \mathbb{R}$ fonctions arbitraires et $e^v \in \{0,1\}^V$ avec des 0 partout sauf à la $v$ième coordonée égale à 1. On peut vérifier que quel que soit $q \in \mathbb{N}^V$, L_{f}^{q} a une unique probabilité invariante π^q. Pour des raison de performances exposées plus en détails Section 1.2.3.3, on choisit

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

avec $a > 0$ paramètre de l'algorithme. Dans ce cas, π^q est donnée par

$$\pi^q(\sigma) = \frac{\prod_{v \in V} (1 + q_v)^{a \sigma_v}}{\sum_{\rho \in S(G)} \prod_{v \in V} (1 + q_v)^{a \rho_v}}, \sigma \in S(G).$$

En plus de π^q, on définit également

$$\bar{\pi}^q(v) = \lim_{N\to+\infty} \pi^{Nq}(\sigma_v = 1).$$

Cette quantité va avoir de l'importance pour les résultats de renormalisations dans les Chapitres 4 et 5. Voir Section 2.1.2 pour plus de détails.

0.2.0.3 Analyse fonctionelle

On présente maintenant les résultats fondamentaux d'analyse fonctionelle utilisé au cours de ce manuscrit. Voir Section 2.2.1 pour plus de détails. Contrairement au manuscrit, on définit ces notions directement pour L^q_{f}.

Définition 0.3
On définitles notions suivantes:

- *La forme de Dirichlet de L^q_{f} définie pour $f, g \in \mathbb{L}^1(\pi^q)$,*

$$\mathcal{E}^q(f,g) = -\left\langle f, L^q_{\mathrm{f}}[g]\right\rangle_{\pi^q}.$$

- *Le trou spectral de L^0, définitpar*

$$\ell^q \coloneqq \inf_{f|\mathrm{Var}_{\pi^q}(f)\neq 0} \frac{\mathcal{E}^q(f,f)}{\mathrm{Var}_{\pi^q}(f)},$$

  avec

$$\mathrm{Var}_{\pi^q}(f) = \sum_{x,y\in S(G)} |f(x)-f(y)|^2 \pi^q(x)\pi^q(y).$$

- *La constante de Log-Sobolev de L^q_{f},définie par*

$$\alpha^q = \inf_{\mathcal{L}^q(f)\neq 0} \frac{\mathcal{E}^q(f,f)}{\mathcal{L}^q(f)},$$

  avec

$$\mathcal{L}^q(f) = \sum_{x\in S(G)} f(x)^2 \log\left(\frac{f(x)^2}{\|f\|_2^2}\right)\pi^q(x).$$

On définit également la distance en variation totale entre deux mesures pour exposer des considérations de temps de mélange.

Définition 0.4
Soient μ et ν deux mesures de probabilités sur $\mathcal{X}$. La variation totale entre μ et ν est donnée par

$$d_{\mathrm{TV}}(\mu,\nu) \coloneqq \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x)-\nu(x)|.$$

La distance en variation totale est lié à une autre pseudo-distance utilisé dans le manuscrit: la divergence de Kullback-Leibler. Voir Section 2.2.1 pour plus de détails sur les relations entre ces notions. La constante de Log-Sobolev apparrait

naturellement dans le raisonnement que l'on va utiliser et il convient donc d'en avoir une bonne approximation. Cette approximation se fait grace au trou spectral et au temps de mélange de la dynamique L_{f}^q.

Définition 0.5
Soit $m_q^x(t)$ la distribution au temps t d'un processus de Markov de générateur L_{f}^q avec condition initiale $x \in S(G)$.

On définit T_{mix} le temps de mélange de L_{f}^q de la manière suivante:

$$T_{\mathrm{mix}}^q \coloneqq \inf\left\{t \geq 0 : \max_{x \in S(G)} d_{\mathrm{TV}}(m_q^x(t), \pi^q) < \frac{1}{2e}\right\}.$$

Cette quantité est aisément estimable et on a de plus en regroupant des résultats standards de processus de Markov la proposition suivante (Proposition 2.11 dans le manuscrit):

Proposition 0.6
- *On a*

$$\frac{\ell^q}{2 - \log(\pi_{\min}^q)} \leqslant \alpha^q \leqslant \frac{\ell^q}{2}.$$

- *De plus,*

$$\frac{1}{\ell^q} - 1 \leqslant T_{\mathrm{mix}}^q \leqslant \frac{1}{\ell^q}\log(\frac{2e}{\pi_{\min}^q}).$$

En particulier

$$\ell^q \leqslant \frac{1}{T_{\mathrm{mix}}^q + 1} \text{ et } \alpha^q \geqslant \frac{1}{(2 - \log(\pi_{\min}^q))(T_{\mathrm{mix}}^q + 1)}.$$

0.2.0.4 Trou spectral/temple de mélange de L_{f}^q

Le trou spectral/temps de mélange de L_{f}^q va avoir une grande influence dans notre résultat. Pour une dynamique avec des taux d'activation/déactivation fixe, cette quantité donne une idée sur le temps que prend le taux de service à chacune des files d'atteindre un équilibre. On refere par exemple à [LY93] pour une borne générale sur le trou spectrale de la dynamique de Glauber sur un graphe cubique. On refere à [RT98] et ses références pour plus de détails sur le trou spectral de L_{f}^q. Qaund l'objectif est d'étudier QB-CSMA, une approche peut être de voir quelle est la dépendence de ℓ^q en $\|q\|_\infty$. Cela peut faire perdre en précision mais borner le trou spetral de cette manière apporte une séparation de l'espace d'état pratique pour l'analyse. Quand le rapport entre taux d'activation et de déactivation est de la forme $\exp(f(q_v))$, la plupart des bornes utilisées sont de la forme

$$\ell^q \geqslant \exp(-\beta f(\|q\|_\infty)).$$

Par exemple dans [SS12], en utilisant l'inégalité de Cheeger, les auteurs obtiennent une borne valable quel que soit le graphe d'interférence:

$$\ell^q \geqslant \exp\left(-2(\Upsilon + 1) f(\|q\|_\infty)\right).$$

avec Υ la taille du plus grand ensemble stable de G. On vera dans les conditions des théorèmes des sections suivantes qu'une meilleur borne sur le trou spectral permet d'obtenir un plus grand choix pour le paramètre a. De manière informelle, on définit $\beta_0(G)$:

$$\beta_0(G) = \inf\{b > 0 \mid \ell^q \geqslant C \|q+1\|_\infty^{-ab} \ \forall q \in \mathbb{N}^V\}.$$

On ommetra le plus souvent la dépendence de β_0 en G. Le but est de trouver une borne pour β_0 aussi petite que possible (dans tous les cas grace au résultat de [SS12], $\beta_0 \leqslant 2(\Upsilon+1)$).

On prouve dans la Section 2.3.2 deux résultats impliquant que $\beta_0 = 1$ est l'ordre de grandeur optimal pour ℓ^q dans le cas d'un graphe complet:

Lemme 0.7
On suppose que G est un graphe d'interférence complet. Il existe $C > 0$ tel que quel que soit $q \in \mathbb{N}^V$,

$$\ell^q \geqslant \frac{c}{\|q+1\|_\infty^a},$$

et

$$T_{\text{mix}}^q \geqslant C \|q+1\|_\infty^a .$$

Quand cela aidera à la présentation, on notera $\beta(G)$ pour insister sur la dépendence de β dans le graphe d'interférence mais elle sera omise dans la plupart des cas. Les preuves de ces deux résultats reposent sur le Lemme 0.6. Voir Section 2.3.2 pour plus de détails.

0.3 Equation de Poisson

Dans cette section, on présentera les résultats nouveaux relatifs à l'homogénéisations grâce à la solution de l'équation de Poisson. Cette partie reprend les définition et résultats du Chapitre 3 de la thèse. Dans ce chapitre, on obtient une borne explicite pour "l'erreur d'homogénéisation" pour le model définitdans la section 0.2.0.2. Plus précisément on bornera la différence entre

$$L_{\text{f}}^q[h](\sigma) = \sum_{v\in V} \sigma_v \Psi_-(q_v)\left(h(\sigma - e^v) - h(\sigma)\right) + \sum_{v\in V}\prod_{w\sim v}(1-\sigma_w)(1-\sigma_v)\Psi_+(q_v)\left(h(\sigma+e^v)-h(\sigma)\right), \quad (3)$$

et

$$L_{\text{h}}[g](q) = \sum_{v\in V}\lambda_v\left(g(q+e^v)-g(q)\right) + \sum_{v\in V}\pi^q(\sigma_v=1)\mathbb{1}_{q_v>0}\left(g(q-e^v)-g(q)\right). \quad (4)$$

après avoir été intégrés le long d'une trajectoire de (Q,σ). La quantité d'interet pour ce chapitre est

$$\left|\int_0^T \left(L_{\text{s}}^{\sigma(s)} - L_{\text{h}}\right)[g](Q(s))\mathrm{d}s\right|, \quad (5)$$

avec $g : \mathbb{R}_+^V \to \mathbb{R}_+$ régulière et $T < +\infty$. Cette quantité nous permet de comparer la dynamique du processus lent avec celle du processus homogénéisé de générateur L_{h}. On rapelle la définition de $\bar{\tau}^U$:

$$\bar{\tau}^U \coloneqq \bar{\tau}^U(Q^N) = \inf\left\{t > 0,\, Q^N(t) \notin U\right\}.$$

Le but du chapitre est de prouver le Théorème 3.1 réécrit ici:

Théorème 0.1
Soit $U \subset \mathbb{R}_+^V$ tel que $0 < \min_{q\in U} \pi^q(\boldsymbol{0})$. Soit $g : U \to \mathbb{R}$ bornée deux fois différentiable sur U. Alors, pour n'importe quel $T > 1$

$$\mathbb{E}\left[\sup_{t\leqslant T\wedge\bar{\tau}^U}\left|\int_0^t (L_s^{\sigma(s)} - L_h)[g](Q(s))\mathrm{d}s\right|\right] \leqslant CB_0 \max_{v\in V}\|\partial_v g\|_{\infty,U}\, T + C\sqrt{T}\Omega_0\left(\max_{v\in V}\|\partial_v g\|_{\infty,U} + \sqrt{T}\max_{w,v\in V}\left\|\partial^2_{w,v} g\right\|_{\infty,U}\right),$$

avec

$$\Omega_0 = (-\log(\min_{q\in U}\pi^q(\boldsymbol{0})))^{3/2}\frac{1}{\min_{q\in U}\ell^q},$$

et

$$B_0 = \frac{\Omega_0^2}{\min_{v\in V, q\in U} q_v^{1+a}} + \frac{\Omega_0}{\min_{v\in V, q\in U} q_v}.$$

0.3.0.1 Définition

Le principal outil de cette section est l'équation de Poisson, accompagnée de ses solutions. Une définition formelle pour QB-CSMA est donnée dans la Définition 3.2 et réécrite ici:

Définition 0.8
Soit $g : S(G) \to \mathbb{R}$ et $q \in \mathbb{N}^V$ l'unique solution à l'équation de poisson associée au générateur L_{f}^q et fonction g est notée $\phi_g(q,\cdot)$ i.e., $\phi_g(q,\cdot)$ est l'unique solution de l'équation en ϕ

$$L_{\mathrm{f}}^q[\phi] = g - \pi^q[g],\ \pi^q[\phi] = 0. \tag{6}$$

En particulier, $\phi_v(q,\cdot)$ est solution de (6) avec $g(\sigma) = \sigma_v$ pour $v \in V$, qui satisfait pour $q \in \mathbb{N}^V$ et $\eta \in S(G)$

$$L_{\mathrm{f}}^q\left[\phi_v(q,\cdot)\right](\eta) = \eta_v - \pi^q(\sigma_v = 1) \text{ et } \pi^q[\phi_v(q,\cdot)] = 0. \tag{7}$$

0.3.0.2 Elements de preuve

La preuve du Théorème 3.1 se fait en deux étapes: on commence par donner une borne à l'expression en fonction de quantités reliés aux solutions de l'équation de Poisson, puis controler ces solutions séparément. Ces deux étapes sont entreprises dans les sections 3.2.1 et 3.2.2 respectivement. La discussion suivante sert à motiver l'utilisation de solutions à des équations de Poisson. En utilisant (3) et (4), on peut

réécrire la différence des générateurs

$$(L_{\mathrm{s}}^{\sigma(s)} - L_{\mathrm{h}})[g](q) = \sum_{v\in V}(\sigma_v - \pi^q(\sigma_v = 1))\mathbb{1}_{q_v>0}(g(q-e^v) - g(q)).$$

De (7), on obtient

$$\sigma_v - \pi^q(\sigma_v = 1) = L_{\mathrm{f}}^q[\phi_v(q,\cdot)](\sigma).$$

Soit $g \in D(L_{\mathrm{h}})$ fixé, on définit $f_v(q) \coloneqq g(q-e^v) - g(q)$. Puisque f_v ne depends pas de σ, on peu réécrire l'expression précédante

$$(\sigma_v - \pi^q(\sigma_v = 1))\, f_v\,(q) = L_{\mathrm{f}}^q\,[F_v(q,\cdot)]\,(\sigma)$$

avec $F_v(q,\sigma) = \phi_v(q,\sigma)f_v(q)$. Puisque l'intégrale de la trajectoire de (Q,σ) s'arrete avant $\bar\tau^U$ le temps de sortie de U pour Q^N, on a $\mathbb{1}_{Q_v(s)>0} = 1$ quel que soit $v \in V$ et $s \leqslant \bar\tau^U$ car $U \subset (0,+\infty)^V$. Par définition de F_v, en intégrant sur une trajectoire, on obtient

$$\begin{aligned}\int_0^t \left(L_{\mathrm{s}}^{\sigma(s)} - L_{\mathrm{h}}\right)[g](Q(s))\mathrm{d}s &= \sum_{v\in V}\int_0^t \left(\sigma_v(s) - \pi^{Q(s)}(\sigma_v = 1)\right) f_v\,(Q(s))\,\mathrm{d}s \quad (8)\\ &= \sum_{v\in V}\int_0^t L_{\mathrm{f}}^{Q(s)}\,[F_v(Q(s),\cdot)]\,(\sigma(s))\mathrm{d}s.\end{aligned}$$

Pour prouver Théorème 0.1, on bornera chaque

$$\int_0^t \left(\sigma_v(s) - \pi^{Q(s)}(\sigma_v = 1)\right) f\,(Q(s))\,\mathrm{d}s = \int_0^t L^{Q(s)}[\phi_v(Q(s),\cdot)f(Q(s))](\sigma(s))\mathrm{d}s$$

individuellement pour une fonction générique f dans le Lemme 3.3 en utilisant la même décomposition et sommant sur v. L'équation de Poisson nous donne une manière alternative d'écrire $\sigma_v - \pi^q(\sigma_v = 1)$. En utilisant (8) et Proposition 0.1, on obtient l' équation (3.5) réécrite ici:

$$\begin{aligned}\int_0^t L_{\mathrm{f}}^{Q(s)}\,[F_v(Q(s),\cdot)]\,(\sigma(s))\mathrm{d}s &= [F_v(Q(t),\sigma(t)) - F_v(Q(0),\sigma(0))]\\ &\quad - M_{F_v}(t) - \int_0^t L_{\mathrm{s}}^{\sigma(s)}\,[F_v(\cdot,\,\sigma(s)\,)]\,(Q(s))\mathrm{d}s. \quad (9)\end{aligned}$$

C'est avec cette expression qu'on prouvera le Lemme 3.3 dans la section 3.2.1.

0.3.0.3 Lemmes cruciaux

On explique maintenant avec plus de détails les deux étapes de la preuve du Théorème 3.1. On commence par définir les quantités

$$\Omega \coloneqq \sup_{q\in U,\,\|g\|_\infty\leqslant 1}\|\phi_g(q,\cdot)\|_\infty\,,\ \text{et}\ B \coloneqq \sup_{q\in U,\,\|g\|_\infty\leqslant 1}\max_{v\in V,\sigma\in S(G)}\left|\phi_g(q\pm e^v,\sigma) - \phi_g(q,\sigma)\right|.$$

Pour cette section, on considère $v \in V$ fixé. On rapelle que $\bar\tau^U$ est le temps de sortie de U Pour Q et C une constante numérique pouvant dépendre de certains paramètres dont la valeur exacte n'a pas d'importance. La première étape menée

dans la section 3.2.1 est de prouver le lemme suivant:

Lemme 0.9
Quels que soient $v \in V$, horizon fini $T > 0$ et $f : U \to \mathbb{R}_+$ différentiable, on a

$$\mathbb{E}\left[\sup_{0\leqslant t\leqslant T\wedge\tilde{\tau}^U}\left|\int_0^t\left(\sigma_v(s)-\pi^{Q(s)}(\sigma_v=1)\right)f\left(Q(s)\right)\mathrm{d}s\right|\right]$$
$$\leqslant C\Omega\left\|f\right\|_{\infty,U}+C\sqrt{T}\left[\left\|f\right\|_{\infty,U}\left(\Omega+B(1+\sqrt{T})\right)+\max_{v\in V}\left\|\partial_v g\right\|_{\infty,U}\Omega(1+\sqrt{T})\right].$$

La preuve découle de (9): on obtient

$$\int_0^t\left(\sigma_v(s)-\pi^{Q(s)}(\sigma_v=1)\right)f\left(Q(s)\right)\mathrm{d}s=\left[F(Q(t),\sigma^N(t))-F(Q(0),\sigma(0))\right]-M_F(t)$$
$$-\int_0^t L_{\mathrm{s}}^{\sigma(s)}\left[F(Q(s),\,\cdot\,)\right](\sigma(s))\mathrm{d}s, \quad (10)$$

avec $F(q,\sigma)=\phi_v(q,\sigma)f(q)$. Pour les termes en F,

$$|F(q,\sigma)|\leqslant\left\|f\right\|_{\infty,U}\Omega. \quad (11)$$

En utilisant le fait que

$$L_{\mathrm{s}}^{\sigma}[f](q)=\sum_{w\in V}\left[\lambda_w\left(f(q+e^w)-f(q)\right)+\sigma_w\mathbb{1}_{q_w>0}\left(f(q-e^w)-f(q)\right)\right],$$

pour le termes en $L_{\mathrm{s}}^{\sigma}[F]$, on utilise le fait que

$$|F(q\pm e^w,\sigma)-F(q,\sigma)|\leqslant\max_{w\in V}\left\|\partial_w f\right\|_{\infty,U}\Omega+\left\|f\right\|_{\infty,U}B, \quad (12)$$

car $f(q\pm e^w)-f(q)=\int_0^1\partial_w f(q\pm ue^w)\mathrm{d}u$ pour obtenir le résultat.

Le terme de martingale est controlé de manière similaire en utilisant le carré du champs du générateur de (Q,σ), l'inégalité de Doob et l'isométrie d'Itô.

On présente maintenant les bornes obtenues dans la section 3.2.2 pour la deuxième étape de la preuve: elles sont regroupées dans le Lemme 3.6 et retranscrite ici:

Lemme 0.10
Soient $q\in\mathbb{N}^V$ et $v\in V$. On défini

$$\Omega(q)=(-\log(\pi^q(\boldsymbol{0})))^{3/2}\frac{1}{\ell q}$$

et

$$B_v(q)=\frac{\Omega(q)}{q_v+1}+\frac{\Omega(q)^2}{q_v^{1+a}}.$$

Alors,

$$\left\|\phi_g(q,\cdot)\right\|_\infty\leqslant C\left\|g\right\|_\infty\Omega(q)\ \text{ et }\ \max_{\sigma\in S(G)}|\phi_g(q\pm e^v,\sigma)-\phi_g(q,\sigma)|\leqslant C\left\|g\right\|_\infty B_v(q). \quad (13)$$

De plus,

$$\sup_{q\in U}\Omega(q)\leqslant\Omega_0 \;\; et \;\; \sup_{q\in U,v\in V}B_v(q)\leqslant B_0. \tag{14}$$

La borne entre Ω_0 et $\|\phi_g\|_\infty$ est un résultat usuel sur les solutions d'équations de Poissons et découle de l'expression explicite du cas général

$$\phi_g(x)=-\int_0^\infty (m_t^x[g]-\pi[g])\,\mathrm{d}t.$$

La borne entre la dérivé discrete de ϕ_g et B résulte d'une identification entre $\phi_g(q\pm e^v,\cdot)-\phi_g(q,\cdot)$ et la solution d'une autre équation de Poisson.

Plus précisément, on pose

$$H(q,\sigma)\coloneqq L_{\mathrm{f}}^q[\phi_g(q-e^v,\cdot)-\phi_g(q,\cdot)](\sigma),$$

et on obtient que $\phi_g(q-e^v,\cdot)-\phi_g(q,\cdot)$ est la solution de l'équation de Poisson

$$L_{\mathrm{f}}^q[\phi]=H(q,\cdot).$$

Le Théorème 0.1 est une conséquence directe de ces deux résultats.

0.3.0.4 Renormalisation

En dehors d'un processus renormalisé, ce résultat n'est pas extremement efficace. On l'appliquera au processus (Q,σ) renormalisé en temps et en espace de la manière suivante:

$$Q^N(t)=\frac{Q(N^\theta t)}{N},$$

et

$$\sigma^N(t)=\sigma(N^\theta t),$$

avec $\theta\in\{1,1+a\}$. Cette étude est menée dans la Section 3.3 du manuscrit, se rapporter à cette section pour plus de détails. Le cas $\theta=1$ donne une approximation à la loi des grands nombre et le cas $\theta=1+a$ apparait naturellement dans le as d'un graphe complet en charge lourde. On donne une définition des générateurs renormalisés: pour$(q,\sigma)\in\frac{1}{N}\mathbb{N}^V\times S(G)$,

$$L_{\mathrm{s},N}^\sigma[f](q)\coloneqq N^\theta\sum_{v\in V}\lambda_v\left(f\left(q+\frac{e^v}{N}\right)-f(q)\right)+\sigma_v\mathbb{1}_{q_v>0}\left(f\left(q-\frac{e^v}{N}\right)-f(q)\right).$$

Le générateur du processus rapide devient $(q,\sigma)\in\frac{1}{N}\mathbb{N}^V\times S(G)$,

$$\begin{aligned}L_{\mathrm{f},N}^q[h](\sigma)\coloneqq{}& N^\theta\sum_{v\in V}\sigma_v\Psi_-(Nq_v)\left(h(\sigma-e^v)-h(\sigma)\right)\\&+N^\theta\sum_{v\in V}\prod_{w\sim v}(1-\sigma_w)(1-\sigma_v)\Psi_+(Nq_v)\left(h(\sigma+e^v)-h(\sigma)\right),\end{aligned}$$

et le générateur homogénéisé:

$$L_{\mathrm{h},N}[f](q) := N^\theta \sum_{v\in V}\left[\lambda_v\left(f\left(q+\frac{e^v}{N}\right)-f(q)\right)+\pi^{Nq}(\sigma_v=1)\mathbb{1}_{q_v>0}\left(f\left(q-\frac{e^v}{N}\right)-f(q)\right)\right].$$

Ainsi, on considère

$$\begin{aligned}
&\int_0^t \left(L_{\mathrm{s},N}^{\sigma^N(s)}-L_{\mathrm{h},N}\right)[g](Q^N(s))\mathrm{d}s\\
=&N^\theta\int_0^t\sum_{v\in V}\left(\sigma_v^N(s)-\pi^{NQ^N(s)}(\sigma_v=1)\right)\left(g\left(Q^N(s)-\frac{e^v}{N}\right)-g(Q^N(s))\right)\mathrm{d}s\\
=&\int_0^{N^\theta t}\sum_{v\in V}\left(\sigma_v(s)-\pi^{Q(s)}(\sigma_v=1)\right)\left(g\left(Q^N(s)-\frac{e^v}{N}\right)-g(Q^N(s))\right)\mathrm{d}s\\
=&\int_0^{N^\theta t}\left(L_{\mathrm{s}}^{\sigma(s)}-L_{\mathrm{h}}\right)[g^N](Q(s))\mathrm{d}s, \qquad (15)
\end{aligned}$$

avec $g^N(q)=g(\frac{q}{N})$. En utilisant le Théorème 0.1 et (15), on obtient Le corollaire 3.7:

Corollaire 0.11
On suppose que $Q^N(0)\to q^0\in(C_-,C_+)^V$, soit $g:\mathbb{R}_+^V\to\mathbb{R}$ deux fois différentiable. On suppose qu'il existe β et $C\in(0,+\infty)$ tels que quel que soient $q\in U$ et $N\geqslant 1$,

$$\ell^{Nq}\geqslant C\left\|Nq+1\right\|_\infty^{-a\beta}.$$

Alors, quel que soient $\theta>0$ et N assez grand,

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\leqslant T\wedge\bar\tau^U}\left|\int_0^t\left(L_{\mathrm{s},N}^{\sigma^N(s)}-L_{\mathrm{h},N}\right)(g)(Q^N(s))\mathrm{d}s\right|\right]&\leqslant CT\max_v\left\|\partial_v g\right\|_{\infty,U}N^{\theta+a(2\beta-1)-2}\log(N)^3\\
&+CT\max_{v,w\in V}\left\|\partial^2_{v,w}g\right\|_{\infty,U}N^{\theta+a\beta-2}\log(N)^{3/2}\\
&+C\sqrt{T}\max_v\left\|\partial_v g\right\|_{\infty,U}N^{\frac{\theta}{2}+a\beta-1}\log(N)^{3/2}.
\end{aligned}$$

Pour un graphe d'interférence complet et $\theta=1+a$, on obtient

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\leqslant T\wedge\bar\tau^U}\left|\int_0^t\left(L_{\mathrm{s},N}^{\sigma^N(s)}-L_{\mathrm{h},N}\right)(g)(Q^N(s))\mathrm{d}s\right|\right]&\leqslant CT\max_v\left\|\partial_v g\right\|_{\infty,U}N^{a-\frac{1}{2}}\log(N)^3\\
&+CT\max_{v,w\in V}\left\|\partial^2_{v,w}g\right\|_{\infty,U}N^{2a-1}\log(N)^{3/2}.
\end{aligned}$$

La deuxieme partie du résultat sur le graphe complet nécessite un examen plus minutieux de

$$\mathbb{E}\left[\int_0^t\sigma_{\mathbf{0}}^N(t)\mathrm{d}t\right]\sim N^{-a}.$$

0.4 Limites fluides

Dans cette partie, on présente les résultats du Chapitre 4. Dans ce chapitre, on présente deux résultats de limites fluides pour l'algorithme QB-CSMA présenté plus haut. Le premier résultat partant d'une condition initiale positive, stoppé au moment ou une des file touche 0 est une application directe des résultats du chapitre précédent. Le deuxieme résultat s'affranchi de certaines hypothèses réstrictive mais n'est valable que dans le cas d'un graphe d'interférence complet. L'étude complete de ce cas donne des indications sur les obstacles à franchir pour le cas général. La principale difficulté de ce chapitre consiste en gerer les reflections en zero. Cette étude relativement technique est résumé dans le Lemme 4.14 et la preuve formelle est donnée en Appendice B. Le processus d'interet dans ce chapitre est

$$Q^N(t) = \frac{Q(Nt)}{N},\ \sigma^N(t) = \sigma(Nt).$$

On commencera par identifier et analyser le processus limite puis on présentera les résultats principaux du Chapitre 4. On présentera finalement des éléments pour prouver ces résultats.

0.4.0.1 Processus limite

La première partie de ce chapitre consiste en une étude de l'EDO limite dans les deux cas d'étude dans les Sections 4.2.1 et 4.2.2. On résume ici les résultats de ces sections (Lemmes 4.1, 4.3 et 4.4). On reprend la définition de l'équation différentielle limite donnée en (4.2):

$$\begin{cases} f' = g(f) \\ f(0) = q^0 \end{cases}. \tag{16}$$

avec

$$\begin{array}{rll} g : (0,+\infty)^V & \to [-1,1]^V \\ q & \longmapsto \lambda - \overline{\pi}^q \end{array} \tag{17}$$

Lemme 0.12

- *Graphe d'interférence général: Si $q^0 \in (0,+\infty)^V$, il existe une unique solution à (16) définie jusqu'au moment où une coordonée touche 0. On la nome $q^*(\cdot, q^0)$ et on note $T_{\text{ext}}(q^0)$ la borne superieur de son support.*
- *Graphe d'interférence complet: si $q^0 \in [0,+\infty)^V$, il existe une fonction q^* telle que (16) est vérifiée pour tout t tel que $q^*(t) \neq \mathbf{0}$ et de plus $s_1(q^*(t)) = \max(s_1(q^0) + (s_1(\lambda) - 1)t, 0)$. En excluant le temps 0, toutes les coordonées touchent zero au même moment.*

L'ennoncé dans le cas du graphe comlet peut être lourd mais il veut fondamentalement dire quelque chose de simple: si $s_1(\lambda) \leqslant 1$ et $q^*(t) = \mathbf{0}$, alors quel que soit $s \geqslant 0$, $q^*(t+s) = \mathbf{0}$, autrement dit le processus est absorbé en zero. Dans le cas contraire, si $s_1(\lambda) > 1$ ou $q^0 \neq \mathbf{0}$ toutes les coordonées deviennent positives quel que soit $t > 0$ assez petit. Si $s_1(\lambda) < 1$, quelle que soit la condition initiale, le processus limite approche zero en temps fini. Tous les q^*_v restent positifs jusqu'au moment ou $q^*(t) = \mathbf{0}$. La première partie du résultat est une application directe du Théorème de Cauchy-Lipschitz et du résultat d'homogénéisation du chapitre précédent. La question est plus délicate dans le deuxième cas à cause des conditions initiales pos-

siblement nulles et de l'horizon arbitraire. Les preuves de ces résultats ainsi que d'autre lemmes préliminaires sont fournies dans la Section 4.3.

Pour faciliter l'exposition, on étend q^* sur $[0, +\infty)$ en disant que q^* arête d'évoluer après $T_{\text{ext}}(q^0)$:

$$q^*(T_{\text{ext}}(q^0) + s) = q^*(T_{\text{ext}}(q^0)) \, \forall s \geqslant 0.$$

0.4.0.2 Résultat principaux

Maintenant que l'on a définit q^*, on peut énoncer les théorèmes principaux de ce chapitre: les Théorèmes 4.5 et 4.6. Soit G le graphe d'interférence fixé. Avec un leger abus de notation, on note β une constante telle que

$$\ell^q \geqslant C \left\| q + 1 \right\|_\infty^{-a\beta}.$$

Techniquement β depend de G mais comme on ne distingue que le cas général et le graphe complet (où $\beta = 1$) on omet cette dépendence.

Théorème 0.2
On suppose que les hypothèses suivantes sont vérifiées:

- $a > 0$ *est tel que* $2a\beta < 1$ *;*
- $Q^N(0) \to q^0$ *pour un* $q^0 \in (0, +\infty)^V$.

Alors $Q^N(\cdot) \xrightarrow{\mathbb{P}} q^*(\cdot, q^0)$ *as* $N \to +\infty$ *uniformément sur les ensembles compacts de* $[0, T_{ext}(q^0))$.

Théorème 0.3
Soient $q^0 \in [0, +\infty)^V$ *et* $\lambda \in \mathbb{R}_+^V$ *fixés. On suppose que les hypothèses suivantes sont vérifiées:*

- $Q^N(0) \to q^0$ *quand* $N \to +\infty$*,*
- G *est un graphe d'interférence complet,*
- $a < \frac{1}{2}$.

Alors $Q^N \xrightarrow{\mathbb{P}} q^*$ *uniformément sur les intervals de temps compacts avec* q^* *characterisée comme dans le Lemme 0.12.*

On note encore une fois la différence entre ces deux résultats: le premier requiert une condition initiale positive et est valable pour le processus stoppé au moment où une des coordonées asymptotique touche zero mais n'a pas d'hypothèse sur le graphe d'interférence (seulement sur a). Le deuxième est valable pour n'importe quelle condition initiale et horizon de temps mais uniquement pour un graphe d'interférence complet.

0.4.0.3 Elements de preuve

Dans les deux cas, la preuve est entreprise avec un argument de localisation expliqué plus en détails dans le chapitre suivant. Pour résumer, la convergence est prouvée pour n'importe quel horizon de temps finis pour un processus stoppé au temps ou il s'echappe d'un tube autour du processus limite. Cela implique que Q^N ne peut jamais s'echaper d'un tube autour du processus limite. On renvoie vers la Section 4.3.1 pour plus de détails. L'idée générale de la preuve est de partir de l'équation

$$\begin{aligned} Q_v^N(t) - q_v^*(t) &= Q^N(0) - q^*(0) + \lambda_v t - \int_0^t \sigma_v(s)\mathrm{d}s - \lambda_v t + \int_0^t \frac{q_v^*(s)^a}{s_a(q^*(s))}\mathrm{d}s + M_v^N(t) \\ &\approx o(1) + \int_0^t \left(\frac{q_v^*(s)^a}{s_a(q^*(s))} - \frac{Q^N(s)^a}{s_a(Q^N(s))} \right) \mathrm{d}s + \int_0^t \left(\pi^{NQ^N(s)}(v) - \sigma_v^N(s) \right) \mathrm{d}s. \end{aligned}$$

Il s'agit ensuite d'utiliser un résultat d'homogénéisation pour que le deuxième terme disparaisse et de réarranger le premier terme pour utiliser le Lemme de Gronwall.

Un autre élément commun au deux cas est la relative compacité de Q^N et le fait que n'importe quel point d'adhérence est une fonction continue. Ceci est la conséquence d'un critère usuel et la preuve est donnée dans la Section 4.3.

0.4.0.4 Cas général

La preuve dans le cas général est une application directe du Corollaire 0.11 et de l'argument de localisation. La preuve repose sur le fait que g est localement Lipschitz et le Lemme de Gronwall ([EK86] Appendix 5) pour prouver que la différence entre le processus des files d'attente et le processus limite ne peut pas croitre significativement. La théorème 0.11 donne

$$\mathbb{E}\left[\int_0^t \left(\pi^{NQ^N(s)}(v) - \sigma_v^N(s) \right) \mathrm{d}s \right] \to 0 \text{ quand } N \to +\infty,$$

et g est lipschitz en dehors de 0. Cela nous permet de prouver la convergence avant qu'une des files touche zero en utilisant le fait que la limite doit être continue. Voir Section 4.4 pour plus de détails.

0.4.0.5 Graphe complet

La preuve dans le cas du graphe complet repose en partie sur les résultats du cas général mais doit en plus fournir un équivalent du Lemme 0.12 pour le processus de file d'attente. Plus particulièrement, il convient de fournir une preuve de l'absorbtion en $\mathbf{0}$ du processus de file d'attente (donné à la fin de la preuve du Théorème 4.6) et le Lemme 4.14 justifiant l'utilisation du résultat d'homogénéisation. On le redonne ici

Lemme 0.13
Si $q^0 \neq \mathbf{0}$ *or* $\sum_{v\in V} \lambda_v > 1$, *alors quel que soit* $\epsilon > 0$ *assez petit, il existe* $t_\epsilon > 0$, *tel que*

$$\mathbb{P}_{q^0}\left(T_-^\epsilon(Q^N) \geqslant t_\epsilon \right) \to 0 \text{ as } N \to +\infty.$$

De plus, $t_\epsilon \to 0$ *as* $\epsilon \to 0$.

L'idée derrière ce lemme est qu'une file peu remplie par rapport au nombre total de requête dans le réseau va en moyenne croître entre deux activations d'un nœud déjà grand car ce nœud reste actif longtemps par rapport au temps d'activation de la file moins remplie. Pour un nombre adéquat d'activation d'une file "remplie" les files "vide" vont avoir crû presque sûrement et le nombre total de requête ne va pas avoir eu le temps de tomber à zero. Le processus limite étant tendu avec un unique points adhérents, cela est suffisant pour prouver la convergence uniforme partant d'une condition initiale dans $[0; +\infty)^V$. Voir Appendices B pour plus de détails sur la preuve du Lemme 0.13 et Section 4.5 pour la preuve du cas graphe complet.

0.5 Charge lourde

Le dernier résultat présenté dans cette thèse concerne le comportement de QB-CSMA en charge lourde avec un graphe d'interférence complet. Le résultat est développé dans le Chapitre 5 est se distingue des résulats de charge lourde usuels de deux manière: la renormalisation n'est pas N^2/N avec une renormalisation en temps le carré de la renormalisation en espace mais N^{1+a}/N avec $a \in (0, 1/2)$ paramètre de QB-CSMA et la limite est déterministe. Dans ce chapitre, le processus d'interet est

$$(Q^N(t), \sigma^N(t))_{t\geqslant 0} \coloneqq (\frac{Q(N^{1+a}t)}{N}, \sigma(N^{1+a}t))_{t\geqslant 0}.$$

Cette distinction rend le résultat plus interessant, d'autant plus que la source de ce comportement inhabituel est identifié: il s'agit de la fraction du temps d'inactivité dans le réseau, inhérente au partage distribué de la ressource. L'ordre de grandeur de cette quantité fait que l'echelle de temps à laquelle le nombre total de requête dans le réseau évolue est N^{1+a} et le fait que $1+a < 2$ implique que les processus de Poisson n'ont pas le temps de s'éloigner de manière significative de leurs interpolations linéaire. Cette dernière observation suggère une limite deterministe sans la garantir mais le résultat d'homogénéisation du Chapitre 3 permet de négliger les fluctuations dûes à l'ordonnanceur.

0.5.0.1 Intuition effondrement de l'espace d'état

A cause du graphe d'interférence complet, au plus une file peut être active à un moment donné dans le temps. Si $s_1(\lambda) < 1$, on peut trouver un algorithme stabilisant le processus des files d'attentes. Si $s_1(\lambda) > 1$, on peut borner le nombre total de paquets dans le réseau par une M/M/1 transiente. Il reste le cas critique quand $s_1(\lambda) = 1$. Pour mieux comprendre pourquoi le processus des files d'attente subit un effondrement de l'espace d'état, on peut regarder q^* la solution de l'EDO de la

section précédente. Soit

$$\begin{aligned} I &\coloneqq \left\{x \in \mathbb{R}_+^V : \forall v \in V,\, \lambda_v = \bar{\pi}^x(v)\right\} \\ &= \left\{x \in \mathbb{R}_+^V : \lambda_v^{-1/a} x_v = \lambda_w^{-1/a} x_w,\ v, w \in V\right\} \\ &= \left\{x \in \mathbb{R}_+^V : x_v = \frac{\lambda_v^{1/a}}{s_{1/a}(\lambda)} s_1(x),\ v \in V\right\}. \end{aligned} \tag{18}$$

Si $q^0 \in I$, par définition de q^*, $q^*(t) = q^0$ quel que soit $t > 0$. Si $q^0 \notin I$, la distance entre q^* et I décroit exponentiellement. Soit

$$d^\infty(q) \coloneqq d_{\mathrm{KL}}(\lambda, \bar{\pi}^q) = \sum_{v \in V} \lambda_v \log\left(\frac{\lambda_v}{\bar{\pi}^q(v)}\right).$$

On remarque que $d^\infty(q) = 0$ si et seulement si $q \in I$, et $d^\infty(q) \geqslant 0$ donc d^∞ peut être vu comme une pseudo-distane à I. On présente maintenant la Proposition 5.1:

Proposition 0.14
Quel que soit $q^0 \in \mathbb{R}_+^V \setminus \{\boldsymbol{0}\}$, si $s_1(\lambda) = 1$ et G est un graphe d'interférence complet,

$$(d^\infty \circ q^*)(t) \leqslant d^\infty(q^*)(0) \exp(-\frac{C'a}{\epsilon} t)$$

La preuve utilise la définition de q^*, les dérivées en q de d^∞ et des relations entre $\|\cdot\|_2$ et d^∞. ce résultat nous dit que dans le cas critique, quel que soit $q^0 \in [0, +\infty)^V$, l'equation aux dérivées partielles de q^* du chapitre précédent a un unique point fixe attractif: l'élément $\widetilde{q}$ de I tel que $s_1(q^0) = s_1(\widetilde{q})$. Ces points vont servir de conditions initiales aux processus d'interet pour cette section mais cette hypothèse peut être relachée grace au Lemme 0.13 et Proposition 0.14. Voir Section 1.3.3 pour une discussion plus fournie et des références sur l'effondrement d'espace d'état. Pour ce model, la discussion dans la Section 5.1.1 apporte plus de détails.

On présente également ici les discussions developées dans les Sections 5.1.2, 5.1.3 et 5.1.4, voir ces sections pour plus de details de références. La fraction du temps que passe le serveur à être inactif ne peut pas être évité en conservant un algorithme distribué. En effet, en se reposant sur l'écoute du canal pour autoriser des activations, il va forcément y avoir une periode d'inactivité entre les activations de deux noeuds. La durée de ces periodes d'inactivité en fonction de la taille des files enttraine une renormalisation différente de celle utilisé habituellement en charge lourde: on regarde l'echelle de temps N/N^{1+a} au lieu de N/N^2 où la renormalisation en temps est le carré de celle en espace. Une métrique de performance impactée par cette inactivité est le délai en charge lourde. Le fait que $1+a < 2$ implique que les processus de Poisson pour les départs et arrivés n'ont pas le temps de s'éloigner de leurs interpolations linéaires pour la topologie uniforme. Les seules sources d'aléatoires sont donc les fluctuations de taux de services dûes à l'ordonnanceur. On montrera grâce au résultat d'homogénéisation que ces fluctuations sont négligeables asymptotiquement et la limite est déterministe.

On suppose qu'il existe une suite ρ^N telle que

$$\frac{1}{\rho_N}\left(\lambda_v - \lambda_v^N\right) \to \gamma_v.$$

avec les mêmes méthodes que celles présentées dans ce chapitre, on peut prouver que la renormalisation "correcte" est $(\rho^N)^{1/a}/(\rho^N)^{1+1/a}$, i.e. on prend $N = (\rho^N)^{-1/a}$. On peut comparer à Max-Weights avec une renomalisation de la taille des files en $(\rho^N)^{-1}$ ou les résultats de [SBB14] qui prouvent une renormalisation en $(\rho^N)^{-2}$ dans un cas d'étude de QB-CSMA différent. Cependant, les auteurs argumentent sur le fait que le délai ne peut pas descendre sous la barre des $(\rho^N)^{-2}$ à cause de la fraction du temps ou aucune file n'est active. Ceci suggère que la condition $a < 1/2$ est optimale pour optenir l'homogénéisation avec une renormalisation charge lourde. On aurait sinon $Q(N^{1+a}t) \approx (\rho^N)^{-1/a}$ avec $a > 1/2$.

0.5.0.2 Résultat principal

Le résultat principal de ce chapitre est le Théorème 5.4:

Théorème 0.4
On suppose que les trois assertions suivantes sont vraies:

- $a < 1/2$;
- $s_1(\lambda) = 1$;
- $Q^N(0) \to q^0$ *pour* $q^0 \in I \setminus \{0\}$.

Alors $Q^N \xrightarrow{\mathbb{P}} \bar{q}$ *uniformément sur les intervales de temps compact, où* $\bar{q}$ *est uniquement définie par:* $\bar{q}(t) \in I$ *quelque soit* $t \geq 0$ *et* $s_1 \circ \bar{q}$ *et l'unique solution de l'EDO* $\dot{x} = \mu x^{-a}$ *avec condition initiale* $x(0) = s_1(q^0)$ *avec* $\mu = s_{1/a}(\lambda)^a$.

Le processus limite $\bar{q}$ a une expression explicite: quel que soient $v \in V$ et $t \geqslant 0$,

$$\bar{q}_v(t) = \frac{\lambda_v^{1/a}}{s_{1/a}(\lambda)} \left(\mu(a+1)t + s_1(q^0)^{a+1}\right)^{1/(a+1)}.$$

Remark 0.15
Ce résultat peut être généralisé au cas presque critique avec les mêmes arguments techniques. Cela donne lieu à une discussion interessante sur la relation entre la distance au bord de la région de stabilité et l'ordre de grandeur de la taille des files mais cette étude n'est pas incluse dans le manuscrit. Voir [Cas+20] pour une étude du cas prèsque critique.

On entreprend l'étude du processus de files d'attente en le stoppant quand il quitte un "tube" autour du processus limite. Cela nous permet de borner les files d'attentes et les borner loins de zero. On introduit

$$C_+ = \max\left(2S(T), \frac{2}{S(0)}, \frac{1}{2}\right) \quad \text{et} \quad C_- = \frac{1}{C_+\mu^{1/a}} \min_v \lambda_v^{1/a}.$$

On définit aussi

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

et le sous ensemble de $\mathbb{R}_+^V$ denommé U:

$$U := \left\{ q \in \mathbb{R}_+^V : \frac{1}{C_+} < s_1(q) < C_+ \quad \text{et} \quad \min_i q_i > C_- \right\},$$

et le temps de sortie de U pour Q^N:

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

Les constantes C_- et C_+ sont définies de telle sorte que le lemme suivant est vrais. Voir Section 5.3.2 pour plus de détails.

Lemme 0.16
Presque sûrement, $T^N < \bar{\tau}^U$. En particulier, $Q^N(t \wedge T^N) \in U$ quel que soit $t \geq 0$.

0.5.0.3 Heuristique et effondrement de l'espace d'état

L'idée pour la preuve du résultat de ce chapitre est de considerer l'évolution du nombre total de requête dans le réseau. Cette évolution est donnée dans l' équation (5.6):

$$s_1(Q^N(t)) - S(t) = s_1(Q^N(0)) - S(0) + \int_0^t \left(N^a \sigma_0^N - \frac{\mu}{S(s)^a} \right) \mathrm{d}s + \sum_{v \in V} \int_0^t \sigma_v^N(s) \mathbb{1}_{Q_v^N(s)=0} \mathrm{d}s + M_{s_1}^N(t). \quad (19)$$

Pour mieux comprendre l'effondrement de l'espace d'état, il convient d'enlever et ajouter le terme homogénéisé:

$$s_1(Q^N(t)) - S(t) = s_1(Q^N(0)) - S(0) + \int_0^t \left(N^a \sigma_0^N - \frac{1}{N^{-a} + s_a(Q^N(s))} \right) \mathrm{d}s + \int_0^t \left(\frac{1}{N^{-a} + s_a(Q^N(s))} - \frac{\mu}{S(s)^a} \right) \mathrm{d}s + \sum_{v \in V} \int_0^t \sigma_v^N(s) \mathbb{1}_{Q_v^N(s)=0} \mathrm{d}s + M_{s_1}^N(t). \quad (20)$$

Le terme avec $\mathbb{1}_{Q_v^N(s)=0}$ disparait quand on stoppe le processus avant $\bar{\tau}^U$. Chacun des termes entre parenthèses est géré séparément. Le premier terme est l'erreur d'homogénéisation pour le temps d'inactivité. Cette quantité est asymptotiquement négligable grace aux résultats du Chapitre 3 (Théoreme 3.1/0.1). Grace à l'effondrement de l'espace d'état, on peut réécrire le deuxième terme comme une fonction de $|S(s) - s_1(Q^N(s))|$. L'inégalité de Gronwall nous permet ensuite de prouver la convergence de $s_1(Q^N)$ vers S. On expose ici les résultats principaux pour la preuve du Théorème 5.4, voir Section 5.3.4 pour plus de détails.

Première étape: Homogénéisation

Proposition 0.17

Soit $f : U \to \mathbb{R}$ deux fois différentiable et soit $T > 1$. Alors

$$\mathbb{E}\left[\sup_{t\leqslant T\wedge\bar{\tau}^U}\left|\int_0^t\left(L_{\mathrm{s},N}^{\sigma^N(s)}-L_{\mathrm{h},N}\right)(g)(Q^N(s))\mathrm{d}s\right|\right]\leqslant C\max_v\|\partial_v g\|_{\infty,U}N^{a-\frac{1}{2}}\log(N)^{3/2}$$
$$+C\max_{v,w\in V}\left\|\partial^2_{v,w}g\right\|_{\infty,U}N^{2a-1}\log(N)^{3/2}.$$

Proof. ce résultat est une répétition du Corollaire 3.7. □

Deuxième étape: Effondrement de l'espace d'état

En utilisant le résultat d'homogénéisation, on prouve le résultat suivant:

Proposition 0.18
Quand $N \to \infty$,

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

La preuve se fait en deux étapes: On controle d'abord l'action du générateur homogénéisé sur d^∞ puis on utilise ce résultat pour controler $d^\infty \circ Q^N$ grace au résultat précédent. Ce résultat est prouvé dans la Section 5.4, la preuve repose sur l'équation (5.4):

$$d^\infty(Q^N(t))=d^\infty(Q^N(0))+\int_0^t L^N_{\mathrm{h}}[d^\infty](Q^N(s))\mathrm{d}s$$
$$+\int_0^t(L^N-L^N_{\mathrm{h}})[d^\infty](Q^N(s),\sigma^N(s))\mathrm{d}s+M^N_{d^\infty}(t). \quad (21)$$

Troisième étape: Preuve principale Pour conclure cette preuve, on utilise l'équation (5.6)/(19) et le fait que grâce à l'inégalité de Pinsker

$$|q_v^a-\lambda_v s_a(q)^a|^2\leq Cd^\infty(q)$$

pour obtenir dans un premier temps la convergence de $s_1(Q^N)$ puis obtenir la convergence des files individuelles car I est 1-dimensionelle. Cette preuve est menée dans la Section 5.5.2

0.6 Conclusion, perspectives de recherche

Dans la conclusion de ce manuscrit, on présente plusieurs pistes pour des recherches futures. Ces pistes de recherches peuvent être organisé en 3 axes principaux: généralisation des résultats de limite fluide, généralisation des résultats de charge lourde et généralisation des résultats d'homogénéisations.

0.6.0.1 Généralisation des résultats de limites fluide

On identifie deux approches pour généraliser le résultat de limites fluides: on peut exhiber une classe de graphe $\mathcal{G}$ ayant la propriété suivante, analogue du Lemme 0.13 pour le processus limite. On note $\tau^0 = \bar{\tau}^{(\mathcal{R}_+^*)^n}$.

Définition 0.19
Soit $\mathcal{G}$ l'ensemble des graphes d'interférences tel que quels que soient $G \in \mathcal{G}$, et $q^0, \lambda \in (0,+\infty)^V$,

$$\tau^0(q^G(\cdot, q^0)) = \tau^0(s_1 \circ q^G(\cdot, q^0)).$$

Une fois $\mathcal{G}$ identifié, il suffit de justifier du comportement des files proche de $\mathbf{0}$ pour prouver le résultat suivant:

Conjecture 0.20
Soient $q^0 \in (0,+\infty)^V$, $\lambda \in \mathbb{R}_+^V$. On suppose que

- *la condition initiale $Q^N(0)$ converge vers q^0 as $N \to +\infty$,*
- *$G \in \mathcal{G}$,*
- *et le trou spectral est tel que $2a\beta(G) < 1$.*

Alors

$$\frac{Q(N\cdot)}{N} \xrightarrow{\mathbb{P}} q^G(\cdot, q^0)$$

uniformément sur les compacts $N \to +\infty$, avec q^G solution de (16).

Pour prouver cette conjecture, il conviendrait de prouver que si $\lambda \in (1+\epsilon)\Lambda^*(S)$, le processus de file d'attente ne touche jamais $\mathbf{0}$ et qu'il est absorbé par cet état en temps fini dès que $\lambda \in (1-\epsilon)\Lambda^*(S)$. On pourrait également inclure des conditions initiales nulles avec un équivalent du Lemme 4.14 (Lemme 0.13) pour un graphe d'interférence non complet. Lemme 4.4 (deuxième énoncé du Lemme 0.12) nous dit que $\mathcal{G}$ contient le graphe d'interférence complet quel que soit le nombre de nœuds. Dans certain graphes, la propriété

$$\tau^0(q^G(\cdot, q^0)) = \tau^0(s_1 \circ q^G(\cdot, q^0))$$

peut dépendre des taux d'arrivé ou de l'état initial. Par exemple, pour 4 nœuds sur un carré. on peut activer simultanément deux coins oposé du carré mais pas deux consécutifs. Les files 1 et 3 ont le même taux de service au premier ordre. Si $q_1^0 \approx q_3^0$ et $\lambda_1 < \lambda_3$ ou $\lambda_1 \approx \lambda_3$ et $q_3^0 > q_1^0$, il est raisonnable de s'attendre à voir la file 1 toucher zero avant la file 3.

Pour cette raison on pourrai s'interesser au résultat d'homogénéisation lors ce qu'une des files touche zero. Les résultats des Chapitres 4 et 5 peuvent être légèrement amélioré: on explique dans la Remarque 4.8 que l'on peut obtenir un résultat d'homogénéisation jusqu'au moment on $Q(Nt)$ devient plus petit que $N^{\max(a\beta+1/2, \frac{2a\beta}{2+a})}$. Ce résulat n'est pas suffisant pour résoudre des cas simples comme celui présenté dans la Section 4.2.4 ou la somme des exposants doit être

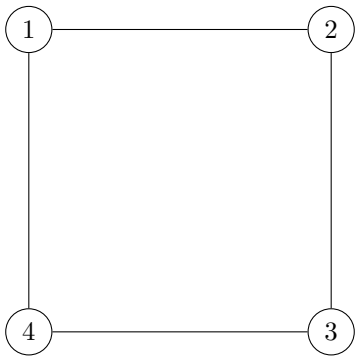

Figure 1: Graphe carré avec 4 nœuds Figure 6.1

1. Il peut être interessant d'améliorer ce résultat pour inclure des situations ou certaines files sont peu remplies. Ces situations apparraissent naturellement quand les nœds appartiennent à des ensembles stables maximaux de taille différentes.

0.6.0.2 Généralisation du résultat en charge lourde

On peut améliorer le Théorème 0.4 pour avoir une limite dans le cas presque critique plus habituel en charge lourde: on considère des taux d'arrivé λ^N tels que il existe $\lambda \in \partial\Lambda^*(S)$ ou ∂ donne le bord de l'ensemble $\Lambda^*(S)$. Plus précisément, on suppose qu'il existe une suite ρ^N telle que

$$\frac{1}{\rho_N}\left(\lambda_v - \lambda_v^N\right) \to \gamma_v.$$

En fonction du rapport entre ρ^N et N trois comportements peuvent être observé, l'étude est menée dans [Cas+20]. Quand $\rho^N \approx N^{-a}$, la dynamique limite a un unique point stable $\bar{q} \in \mathbb{R}_+^V$. Ce point est de plus attractif. Si $Q^N(0) \to \bar{q}$, le processus limite est constant et cela suggère de regarder à une echelle de temps plus rapide comme pour le théorème de la limite centrale. Dans le cas du graphe complet, il semble que l'echelle de temps suivante est N^2/N où on devrait commencer à observer des deviations entre les processus de Poisson et leurs interpolations linéaire. Notre résultat d'homogénéisation ne peut pas directement s'appliquer dans ce cas on s'attend au résultat suivant:

Conjecture 0.21
On suppose que $Q^N(0) \to \bar{q}$, G est un graphe d'interférence complet et $a < 1/2$.

Alors $\bar{Q}^N$ converge vers un processus gaussien avec une dérive et une variance qui dépendent de l'état du processus limite.

Voir section 6.2.2 pour plus de détails.

0.6.0.3 Résultat d'homogénéisation

On peut reformuler les résultats du Chapitre 3 pour avoir un résultat dans un cardre plus général: soit L le générateur d'un processus de Markov irréductible tel que quel que soit $q \in \mathbb{N}^V$ et $\sigma \in S^0$ avec S^0 finie,

$$L[f](q,\sigma) = L_{\mathrm{f}}^q[f(q,\cdot)](\sigma) + L_{\mathrm{s}}^\sigma[f(\cdot,\sigma)](q).$$

On suppose que quel que soit $q \in \mathbb{N}^V$, L_{f}^q a une unique mesure de probabilité invariante π^q . On peut reformuler Théorème 0.1 à l'aides des quantités suivantes:

- La constante de Log-sobolev de L_{f}^q:

$$\alpha^q,$$

- La dérivé en q des taux de saut pour σ:

$$d_l(q) \coloneqq \max_{v\in V,\sigma,\sigma'\in S^0} \partial_v L_{\mathrm{s}}^q(\sigma,\sigma');$$

- et les dérivé en q de la mesure invariante de L_{f}^q

$$d_\pi(q) \coloneqq \max_{v\in V,\sigma\in S^0} \partial_v \pi^q(\sigma).$$

Dans ce cadre légèrement plus général, on obtient la borne suivante:

$$\begin{aligned}\mathbb{E}\left[\int_0^{t\wedge\bar{\tau}^U}\left(L_{\mathrm{s}}^{\sigma(s)} - \pi^{Q(s)}[L_{\mathrm{s}}^{\sigma(s)}]\right)[f](Q(s))\mathrm{d}s\right] &\leqslant Ct\max_{v\in V}\|\partial_v g\|_{\infty,U}\|\alpha^q\|_{\infty,U}\|d_\pi\|_{\infty,U}\\ &+ Ct\max_{v\in V}\|\partial_v g\|_{\infty,U}\|\alpha^q\|_{\infty,U}^2\|d_l\|_{\infty,U}\\ &+ C\sqrt{t}\|\alpha^q\|_{\infty,U}\left(\max_{v\in V}\|\partial_v g\|_{\infty,U} + \sqrt{t}\max_{v,w\in V}\left\|\partial_{v,w}^2 g\right\|_{\infty,U}\right).\end{aligned}$$

Un objectif pour mes futures recherches seraient d'étudier d'autres modèles que QB-CSMA ayant un phénomène d'homogénéisation et d'investiguer les interactions avec d'autres spécificité du modèle, par exemple à travers une étude des phénomènes de méta-stabilité ou de grande déviations. Voir Sections 6.3.1 et 6.3.2 pour plus de détails.

Chapter 1

Introduction

Contents

1.1 General discussion 1
1.1.1 Motivation 1
1.1.2 Wireless networks modeling 2
1.2 Main algorithms 4
1.2.1 Max-Weight 5
1.2.2 Greedy maximal scheduling, Longest Queue First 6
1.2.3 CSMA-type algorithm 7
1.2.4 Polling 12
1.2.5 Load balancing algorithms 12
1.3 Homogenization and State Space Collapse 13
1.3.1 Stochastic Averaging 13
1.3.2 Poisson equation and Stein's method 17
1.3.3 State Space Collapse, Skorokhod problem 19
1.4 Functional limit theorems 20
1.4.1 Fluid limits 22
1.4.2 Heavy traffic 24

1.1 General discussion

1.1.1 Motivation

People and object tend to be more and more connected through wireless communications. The huge amount of data exchanged on wireless networks requires faster, more reliable and more flexible communication schemes. Users of a wireless network share the air as common resource for communications, which creates some difficulties. If users that are close geographically try to access the wireless network on the same bandwidth, their request will get mixed and interference will occur. From applications in peer to peer networks and Wi-Fi access to call centers, this type of

resource sharing problem has been a central part of research in applied probability for at least 50 years.

In a broad sense, many questions associated with communication algorithm can be stated as "How to allocate the shared communication resource?", "How to measure efficiency in communication networks?", "Is this algorithm efficient/optimal (in what sense?)?", "How does it behave under such conditions?", and so on. Queuing theory provides a methodological framework to be able to deal with this kind of questions. For instance establishing theoretical guaranty for duly defined performance metrics of communication schemes or model and simulate the behavior of a real communication network. Users of a real network are modeled as queues with incoming jobs. Each queue has buffer area to store incoming jobs before being able to serve them. From a terminology standpoint, queues will be refered to as nodes, users or server depending on the context. Similarly, jobs will also be refered to as request.

Having a central authority making decisions taking into account the state for the whole network can usually lead to more efficient decisions and good performance if the information is used to its full extent. However, having a central authority allocating the resource also induces some drawbacks. An attack on a central authority could cause a major breakdown of the system and gathering information about the whole network can be challenging. Nodes like smartphones are required to seamlessly enter or exit the network which makes gathering information both time consuming because of the number of nodes and unreliable because of information that may be incomplete or dated. On the opposite of a central authority, one could prefer to let users take decisions based on their surroundings in the network. In addition to the reasons provided previously, it seems more reasonable to rely on individual users to gather information in their surroundings and make decisions using this information but without any coordination, this could lead to greedy behavior by individual stations and inefficiency for everyone. From an individual point of view, users want to be able to use as much of the resource possible to minimize the amount of time they have to wait before completion of their jobs. However this type of behavior penalizes other users because it limits their access to the resource. This could prevent the resource to be efficiently distributed. A major question is to find a fully distributed algorithm designed to efficiently and fairly share the resource between users. This type of distributed algorithm is more appropriated with the constant changing nature of wireless networks and easier to scale as the size of networks grows.

1.1.2 Wireless networks modeling

Markov processes provide a rigorous mathematical framework to study the performance of different types of algorithms. In addition, queuing theory provides a rigorous framework to study congestion phenomenon. A qualitative approach is to consider the stability of the queuing process. Formally, a Markov process is said to be stable if its distribution is tight. Stability of the queueing process gives a first qualitative idea: the number of jobs waiting in the network is closely related to the waiting time in equilibrium through Little's Law, stating that the average number of customers in a system is equal to the product of the average arrival rate λ and the average time spent in the system. Informaly, if the total number of jobs in the

system diverges, so does the mean waiting time jobs in the system. See for instance [LG08] Chapter 12 Section 2. If it exists, the steady state average of the process of queue lengths gives a good approximation of the queueing process when run for a long time.

One major challenge in wireless communication is interference. There are two main ways to mathematically model interference: Boolean interference graphs and Signal to Interference + Noise Ratio (SINR).

Gilbert proposed a graph modeling for telephone central offices as an application for his paper on random graphs in [Gil59]. Nodes of the network are offices and links represent a communication route between offices. His result can be applied to compute the probability of the existence of a route between each office and launched the study of percolation in random graphs. A Boolean interference graph is a more advanced interference situation. Consider a set of users with communication links between them. A user can only communicate with one other user at any given time. The line graph of the network is the graph whose edges are users of the network and nodes are communication links between users. This line graph is used as interference graph. Each station can be in contact with at most one other station so the scheduling decision needs to be an independent set of the interference graph or equivalently a matching of the network graph. Sampling independent sets/ matching is an integral part of randomized scheduling. This example comes from peer-to-peer networks but the technique is not limited to it. The interference graph is not necessarily the line graph of a network. As soon as there is some interference, the messages concerned are lost and cannot be recovered by the recipient. This is not the case for the SINR model.

The intuition behind SINR can be found in [Gil61] with the introduction of stochastic geometry and percolation. The model studied in [Gil61] is constructed with a Poisson process on a plane. Points of the Poisson process that are at a distance less than $R > 0$ are joined by an edge. He studies the existence of an infinite component and the probability of a typical node being in this component. This model can be used for a communication network where stations can transmit with a range R. SINR was formalized more recently in [DBT03] and [Dou+06]. See [BB09] for an overview of the results on SINR models. With the SINR modeling, each node has a signal strength attached and the quality of the signal degrades over the distance it travels. When trying to transmit, each user emits its message on the same channel. Depending on how close, neighboring users will also receive the message as interference. If the ratio between the signal strength and the signal and noise coming from all the other stations is above a threshold then the message passes through. Otherwise, the destination is not able to decipher the message. This modeling is more realistic and it is in fact a lot more complicated to study than the previous one. In this manuscript we will only consider the interference graph case.

Another distinction is also made between single-hop and multi-hop. In single-hop once a request is treated by a server it leaves the system. In multi-hop, once it is treated by a server it may need to jump a certain amount of time to other servers. Similarly, all of those models can be discriminated with respect to the size of their waiting area or the service policy at each queue. Each queue can have a finite or infinite buffer for the number of requests that can be in standby waiting to get service. The service policy gives the order in which jobs are processed in the

waiting line of each queue.

In practice any distribution could be used for inter-arrival times and service requirement. In this manuscript, we will only consider exponential interarrivals and service times: the size of each job will be an exponential variable and the time between new jobs arrival is exponential. With jobs of exponential distribution and Poisson arrivals, the process of queue lengths is a finite dimensional Markov process. If the job distribution is not exponential, it is still possible to define a Markov process but this one will be measure valued (infinite dimensional, which is impractical to study). We mention here the existence of multi-type networks where different types of job may arrive in the network. Jobs of different types can have another size distribution, mean, have to folow a specific route in the network, ... In some cases, only the mean, variance and higher order moments of the job size distribution influence performance metrics. Insensitivity in the job size distribution have been introduced in [BP02] for a load balancing situation. Insensitivity results present a particular interest for applications by allowing network designer to dimension their networks without the need to know in advance the precise distribution of the lengths of transmissions. We mention as well [Ven+10] for an example on CSMA networks. They establish insensitivity of the stationary distribution in the distribution of job size and back-off times. Insensitivity results are very important for applicability because it removes the need for exponential job size to apply bounds proved on the Markov process. The exponential case is usually easier to study than the general case.

Servers can use different service disciplines governing the order in which they process the request in their queue: priority, First Come First Serve (FCFS), processor sharing (PS) for instance. For FCFS jobs are served in the order of their arrival at each queue. For PS the service capacity is evenly distributed all jobs. We also mention Head of Line PS where each server with a non-empty queue gets an equal fraction of the service capacity and servers process one job at a time in a FCFS fashion. For the priority queue, for each server, establish a list of priority rank for different type of jobs. Customer of a higher class will always get priority over lower classes. If a job with a rank higher than the one currently being served, the lower rank interrupts its service to let the high priority job be processed. Then the server resumes the job where it was interrupted.

1.2 Main algorithms

In the context of queueing theory, the question “How to allocate the shared communication resource?” can be declined in multiple subjects of research. How to allocate the resource can mean different things depending on the context. For instance, there can be a fixed number of servers to allocate for a fixed number of queues sharing the service capacity. This is called a polling system. A second resource sharing model is to consider an interference graph. In this situation, the question becomes for which queue do we allow transmission at any given time, knowing that activating one node prevents all neighbors from transmitting. This is called a scheduling problem. Next we can also think about a centralized system with a fixed number of queues where jobs arrive at a central dispatcher which is tasked with finding the best possible route for each job. This is called a routing problem. An important

example of scheduling problem is that of a switched network. A switched network is an array of $N \times K$ queues. The scheduling constraint is that at most one queue can be active on any column and any row of the matrix. This type of constaints come from physical switches in telephone network.

To study the performance of those algorithms, a Markovian framework can be useful by providing rigorous performance metrics. The question "How to measure efficiency in communication networks?" can be specified further into qualitative result and quantitative ones. Heuristically to measure performance, once could for instance measure the time it takes for a new job entering the system to get service. This is called delay. Delay can usually be related to the number of jobs present at each of the queues. One possible qualitative performance metric is the recurrence/transience of dynamics related. For instance if the queuing process is a positive recurrent Markov process, the total number of jobs in the system remains in a compact set of $\mathbb{N}$ with high probability even for large times. From a quantitative standpoint, since the process of queue lengths is Markov, if it is stable (otherwise said positive recurrent), computing the steady state delay gives an indication on the waiting time users experience at each queue when the network has already been running a long time. Heavy traffic approximations can also be seen as a third strain of performance results: they give an approximation of the process of queue lengths when the network is on the verge of overloading.

1.2.1 Max-Weight

An important algorithm in the field of scheduling is the celebrated Max-Weight algorithm from the seminal paper [TE92]. It was introduced for constrained queueing network in a multi-hop multi-type setting. We explain the procedure here on a simple scheduling example. We have V a finite set of queues and S be the set of possible service decisions. Elements of S are seen as elements of $V^{\{0,1\}}$ as configurations of admissible service rates. Let Let $f : \mathbb{N} \to \mathbb{R}_+$ increasing, with some additional technical assumptions not detailed here. The f-Max-Weight algorithms is a discrete time algorithm that chooses a service decision $\sigma \in S$ such that

$$\sigma \in \operatorname*{argmax}_{\rho \in S} \sum_{v \in V} \rho_v f(Q_v(t)), \tag{1.1}$$

where ties between service decisions of maximum weights being cut uniformly at random. In general, solving this optimization problem is complex even approximately. The appeal of this procedure is that it provides good performance in practice and theoretical guaranty, the drawbacks are its centralized nature and computational complexity. The stability region of Max-Weight can be defined as

$$\Lambda_{\mathrm{MW}}(S) \coloneqq \left\{\lambda \in \mathbb{R}_+^V,\ (Q(t))_{t \geqslant 0} \text{ with arrival rates } \lambda \text{ is an ergodic Markov process}\right\}.$$

We call capacity region the set of $\lambda \in \mathbb{R}_+^V$ such that there exists $\rho \in Co(S)$, with $\lambda < \rho$ component-wise and $Co(S)$ the convex hull of S, with elements of S seen as elements of $\mathbb{R}_+^V$. This capacity region is the largest possible stability region in the sense that if λ is not in this set, for any scheduling algorithm, the norm of process of queue lengths diverges to infinity as $t \to +\infty$. Let

$$\Lambda^*(S) \coloneqq \{\lambda \in [0,1]^V \mid \exists \rho \in Co(S), \lambda < \rho \text{ component-wise } \}$$

be the capacity region. For any $\lambda \in \Lambda^*(S)$, the authors of [TE92] proved that the process of queue lengths is stable. For any $\lambda \in \Lambda^*$, they presented a Lyapunov function whose mean is decreasing over time, meaning that $\Lambda_{\mathrm{MW}}(S) = \Lambda^*(S)$. An algorithm with this stability region is said to be throughput optimal.

1.2.2 Greedy maximal scheduling, Longest Queue First

Instead of choosing independent sets with maximum weights, there are multiple algorithms that schedule maximal independent sets. Maximal independent sets are sets of nodes such that any added node makes the independent set condition fail. This is much easier to implement and yields some decent performance. In general this type of algorithm is still centralized but only requires solving a simple optimization problem (comparing the size of two or more queues).

A natural algorithm to provide a simple approximation of Max-Weight is the LQF (Longest Queue First) algorithm introduced in [McK95]. To find the schedule selected by the algorithm schedule the queue with the longest backlog. Remove all servers that interfere with the selected server and schedule again the server with the longest backlog. Repeat the operation until the set of possible nodes to add is empty. In [DW06], Dimakis and Walrand provided some sufficient conditions for the optimality of the stability region of this algorithm for general service distributions using fluid limits as defined in Section 1.4.1. Their necessary condition involves "complete local pooling", a structural property of the interference graph. In fact, the fraction of the optimal stability region achieved by LQF is equal to the local pooling factor defined in [JLS09] and [LBX11]. For any subgraph $G' = (V', E')$ of G, let $M(G')$ be a $0-1$ matrix with $\mid V' \mid$ rows and whose columns are maximal stable sets of G. The pooling factor of G' is defined as

$$\sigma^0(G') \coloneqq \sup\{\sigma > 0 \mid \exists \alpha \in [0,1]^{|V|}, \sigma e \leqslant \alpha M(G) \leqslant e\}$$

with e the vector of size cardinality of the sets of stable sets of G with 1 for every coordinate and α seen as a line vector. The local pooling factor of G is

$$\sigma^*(G) \coloneqq \min\{\sigma^0(S),\ S \text{ subgraph of } G\}.$$

A dual representation for the definition of $\sigma^0(G')$ is given by

$$\begin{aligned} &\max_{x, w \geqslant 0} w \\ &\text{subject to} \quad \max_{\sigma \text{independent set of } G'} \langle x, \sigma \rangle \leqslant 1, \\ &\qquad\qquad \min_{\sigma \text{independent set of } G'} \langle x, \sigma \rangle \geqslant w \end{aligned}$$

The local pooling factor essentially measures how evenly can the resource be distributed. For any scheduling policy, let $\gamma^*(G)$ be defined as

$$\gamma^*(G) \coloneqq \sup\{\gamma \mid \text{ Queue lengths are stable for any } \lambda \in \gamma\Lambda^*(G)\}.$$

The result in [JLS09] states that for LQF, $\gamma^*(G) = \sigma^*(G)$. Later, [Bir+12] provided a simple characterization of network graphs such that $\sigma^*(G) = 1$. This entails that LQF is throughput optimal on those graphs. Essentially, the condition states that if G contains at most one cycle of size 5 or 7, then LQF is throughput optimal.

1.2.3 CSMA-type algorithm

1.2.3.1 Classical CSMA

The celebrated CSMA algorithm is a classical algorithm for scheduling queues in a wireless network. Its implementation is simple and it has been extensively studied in the literature. The acronym CSMA stands for Carrier-Sense-Multiple-Access. This indicates that servers have a carrier sensing capability enabling them to listen to the communication channel, sensing when an interfering user is occupying it, and multiple access means that multiple servers may be active at the same time. The carrier sensing ability enables the scheduler to completely avoid collision of messages in the continuous time case. In discrete time, collisions can occur but the only possibility is if two or more queues start transmiting at the same time. Without this kind of assumptions, the performance of scheduling algorithms is quite bad. In [Lie+09], the authors presented some Back-of-the-Envelope computation to estimate the quality of approximation between the idealized mathematical model and real wireless networks.

We mention for instance the ALOHA protocol [Abr70] where nodes enter a back-off period refraining from transmiting after collisions. ALOHA only collects information about collisions after they occur. With ALOHA, users try to use the network whenever they want. If there is no collision, they are free to send the next job whenever they want. If there is a collision, nodes are forced to enter a random back-off period where they refrain from competition. For any positive arrival rates, the queue lengths are not stable when operating ALOHA. It is possible to refine this algorithm. For instance, the more collisions happen in a row, the longer the back-off periods can become. This is a way to diminish the pressure on the channel when the network is heavily loaded. In a landmark paper, Aldous proved in [Ald87] that even with this refinement of the algorithm, its stability region is empty. Using the carrier sensing ability provides a non-negligible advantage to CSMA-type algorithms.

For CSMA, the schedule follows a Markovian dynamic with rates independent from the process of queue lengths. The schedule actually follows a single site update dynamic for the hard-core model: the Glauber dynamic described below. Each node v has a fixed "fugacity" ν_v. The term fugacity is borrowed from statictical physics and can be seen as an activation rate. When a node is active, it will remain that way for a duration that is exponential with parameter $\frac{1}{1+\nu_v}$. When a server is inactive, it checks if some of its neighbors are active with its carrier sensing ability. If none of them are active, the server runs an exponential clock with parameter $\frac{\nu_v}{1+\nu_v}$ to determine when to activate. When a neighbor of v becomes active, v's clock freezes and resumes when the channel is clear. All interfering queues compete for activation. This is done to avoid any collision between messages. The long term throughput at each node is given by the invariant measure of the dynamic of the schedule: a Glauber dynamic on the stable sets of G ; each node updates its state at a fix rate, conditioned in staying a stable set of the interference graph.

One of the feature that distinguishes CSMA-type algorithm from ALOHA is that because of the carrier sensing ability, collisions can only happen if two competing servers activate at the same time. For the continuous Markov chains, this does not happen with probability 1. For discrete time, interference can only happen at the start of an active period for a server.

If the arrival rate are in the capacity region and known, it is possible to design the fugacities such that the long term throughput of each queue is greater than the average input and the process is stable. The long term throughput of the algorithm is given by the invariant measure of the Glauber dynamics. This measure indicates the fraction of the time each schedule is chosen. With a vector of fugacities ν, the invariant measure is given for any $\sigma \in S(G)$ by

$$\pi^{\nu}(\sigma) = \frac{\exp\left(\sum_{v \in V} \sigma_v \log(\nu_v)\right)}{Z^{\nu}},$$

with

$$Z^{\nu} = \sum_{\rho \in S(G)} \exp(\sum_{v \in V} \rho_v \log(\nu_v)).$$

For the cross product of elements $q, \sigma \in \mathbb{R}^V$, we will use the notation

$$\langle \sigma, q \rangle \coloneqq \sum_{v \in V} \sigma_v q_v.$$

If $\lambda \in (1-\epsilon)\Lambda^*(S(G))$, it is possible to find ν such that $\pi_\nu(\sigma_v = 1) \geqslant \lambda_v$ for all $v \in V$ by Proposition 1 of [JW10]. If the arrival rates are not known, designing fugacities in advance to handle any configuration of arrival rates is not possible. Assign weight $\log(\nu_v)$ to node v for each $v \in V$. As $\max_v \nu_v \to +\infty$ this invariant measure concentrates on independent sets of largest weights. Many algorithms where the fugacities evolve over time have been designed since then in order to be able to use this type of algorithm in situations where the arrival rates and/or the interference graph is not known. The idea behind those is either to find fugacities such that the throughput rate is greater than the input rate at every queue or to approximate the service decisions of a known good algorithm such as Max-Weight.

In addition to their insensitivity result, the authors of [Ven+10] also prove conditions on the arrival rates such that CSMA with given fixed activation rates is stable. Their result is quite different from the usual question which is: with given arrival rates, find activation rate such that the process of queue lengths is stable. Their question is much more challenging as soon as the interference graph is not complete.

1.2.3.2 Rate-based adaptive CSMA

One approach to provide for an "adaptive" CSMA is the rate based CSMA algorithm introduced in [JW10]. The idea behind it is to update fugacities in the classical CSMA algorithm so that the long term throughput matches the input rate. The idea is to estimate the arrival rates on the fly. In the first article, they assume a technical assumption, which is proven in the later technical report [Liu+08]. The proof relies on the study of stochastic sub-gradient algorithms modulated by a Markov chain. Essentially, the idea is to match the time average of the time spent at each node with their arrival rates. The key is to gradually improve the invariant measure of CSMA by running it with fixed parameter for some time, and then update by solving an optimization problem with the empirical arrival rates during that period. If the parameters of CSMA are changed slowly enough, the schedule will reach equilibrium and the empirical arrival rates will be close to their mean

before the next update happens.

If $\lambda \in \Lambda^*(G)$, for any $\sigma \in S(G)$ there exists $\bar{p}_\sigma \in (0,1)$ not necessarily unique such that such that

$$\lambda_v = \sum_{\sigma \in S(G)} \bar{p}_\sigma \sigma_v.$$

For any configuration of fugacities ν, recall the long term fraction of the time schedule σ is chosen given in the previous discussion

$$\pi^\nu(\sigma) = \frac{\exp(\langle \sigma, \log(\nu) \rangle)}{Z^\nu}.$$

The goal of the algorithm is to maximize the "log-likelihood function":

$$F(\nu, \lambda) \coloneqq \sum_{\sigma \in S(G)} \bar{p}_\sigma \log\left(\pi^\nu(\sigma)\right).$$

If the function $F(\nu, \lambda)$ is maximized in ν^0, then $\lambda_v \leqslant \sum_{\rho \in S(G)} \rho_v \pi^{\nu^0}(\rho)$ for all $v \in V$. They prove that this optimization problem has a solution with $\nu^* < +\infty$ component-wise as soon as $\lambda \in \Lambda^*(G)$.

1.2.3.3 QB-CSMA

The next algorithm presented is the one that is studied throughout this manuscript: Queue-Based CSMA (QB-CSMA). The general idea behind Queue-based CSMA is similar to the classical CSMA: nodes decide in a distributed fashion who is permitted to transmit. The main difference is that fugacities evolve over time and actually depend on the number of jobs at each of the queue. In practice, this means that the fugacities from classical CSMA become $\nu_v \to \exp(f(q_v))$.

One of the key feature of QB-CSMA is that it is throughput optimal if f increases slowly enough. To prove this result, the usual method is to establish a fully coupled stochastic averaging principle where the schedule averages in such a way that its service decisions are approximate solutions to the Max-Weight problem (1.1). The literature on optimal CSMA algorithms is rich and the interested reader is for instance referred to the thorough survey by Yun et al. [Yun+12] for more details. In this manuscript 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 deactivation rates Ψ_+^v and Ψ_-^v being adapted as a function of queue lengths. Rajagopalan, Shah and Shin study in particular the case where

$$\Psi_+^v(q) = \frac{e^{W_v(q)}}{1 + e^{W_v(q)}}, \quad \text{and } \Psi_-^v(q) = 1 - \Psi_+^v(q),$$

with

$$W_v(q) = \max(f(q_v), h(f(\max_w q_w))), \tag{1.2}$$

for some functions f and h. 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. In essence, for any $q \in \mathbb{N}^V$, they use

fugacities $\nu_v(q) = \exp(W_v(q))$, which gives an invariant measure

$$\pi^q(\sigma) = \frac{\exp(\langle \sigma, W(q)\rangle)}{Z^q},$$

with Z^q the normalizing constant. The main property of this probability measure is proved in Lemma 2.19. It states that for any $0 < \epsilon < 1$, there exists $q^* \in \mathbb{R}_+$ such that for any $q \in \mathbb{R}_+^V$ with $\max_v q_v > q^*$,

$$\pi^q[\langle \cdot, \log(W(q))\rangle] \geqslant (1-\epsilon) \max_{\rho \in S(G)} \langle \rho, \log(W(q))\rangle .$$

If the schedule at time t is effectively distributed close to $\pi^{Q(t)}$, it can serve as an approximation of the decisions of the Max-Weight algorithm for the weights $\log(W)$. See Section 2.3.3 for a discussion about the function h and its role. In [SS12], the authors proved a time scale separation for $h(x) = \sqrt{x}$ as long as for any $\delta \in (0,1)$,

$$\lim_{x \to +\infty} \exp(f(x)) f'(f^{-1}(\delta f(x))) = 0.$$

These algorithms also use some information on the current maximum queue length and this can hurt the performance of the algorithm. Results of [GBW14] can use $h(x) = 0$ and suggest that if f grows polynomially, then it is only throughput-optimal for some interference graph, depending on the relation between the graph topology and the exponent of the polynomial growth of f. Under the assumption that the fluid limit (as defined in Section 1.4.1) converges and under a time scale separation assumption, the authors of [GBW14] prove that QB-CSMA is throughput optimal. A rigorous justification for their assumption is given in Chapter 4 on a complete interference graph.

The intuition for seeking fast-increasing functions f is that the fraction of the time no queue is active decreases faster with $\|q\|_\infty$ when f grows fast. More formally, a folklore result has it that delay is improved with faster increasing functions f, an intuition which is backed up by results in [BBL11]. They consider fixed arrival rates λ_v and service rates μ_v. The activation and deactivation rates may depend on the state of the network. Let μ and ν be two vectors in $\mathbb{R}_+^V$. The deactivation rate of node v with queue size q_v, is given by $g_v(q_v) = \mu_v \Psi_v^-(q_v)$ and activation rate $f_v(q_v) = \nu_v \Psi_+^v(q_v)$. Let f^{-1} and g^{-1} be the inverse applications of f and g. In essence, their result states that if the $f_v = f$ is a concave positive and strictly increasing function and $\Psi_-^v(q) = 1$, for any clique $\mathcal{N}$ of G,

$$\mathbb{E}\left[\sum_{v \in \mathcal{N}} Q_v\right] \geqslant \lambda_{\mathcal{N}} \frac{\sum_{v \in \mathcal{N}} \frac{\lambda_v}{\mu_v^2}}{1+\rho_{\mathcal{N}}} + |\mathcal{N}|\, f^{-1}\left(\frac{\lambda_{\mathcal{N}}}{1-\rho_{\mathcal{N}}}\right),$$

with $\lambda_{\mathcal{N}} = \sum_{v \in \mathcal{N}} \lambda_v$ and $\rho_{\mathcal{N}} = \sum_{v \in \mathcal{N}} \rho_v$ with $\rho_v = \frac{\lambda_v}{\mu_v}$. There is a symmetric result if $\Psi_+^v(q) = 1$ and $g_v = g$ is a decreasing convex function:

$$\mathbb{E}\left[\sum_{v \in \mathcal{N}} \rho_v Q_v\right] \geqslant \rho_{\mathcal{N}} g^{-1}\left(\frac{(1-\rho_{\mathcal{N}})\nu_{\mathcal{N}}}{\rho_{\mathcal{N}}}\right),$$

with $\nu_{\mathcal{N}} = \sum_{v \in \mathcal{N}} \nu_v$. In some interference graphs, having an activation ratio increasing too fast can be nocuous to performance: this is the center of the result from [GBW14] stating that if the deactivation rate decreases too fast queues stay

active until they are almost empty.

Polynomial activation and deactivation functions should therefore achieve the optimal trade-off between throughput and delay for this class of algorithms, because it is not possible to achieve a time scale separation if the activation function increase faster than polynomially, see Section 2.3.3 for a more detailed discussion on the matter. 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.

1.2.3.4 Q-CSMA

One final refinement of CSMA mentionned here is the Q-CSMA introduced by J. Ni, B. Tan and R. Srikant in [NTS09] more or less simultaneously to Rate based and QB-CSMA. The idea for this algorithm is also to distributively approximate the decisions of the Max-Weight algorithm by updating multiple nodes at a time with probabilities that depend on their queue lengths. They provide a proof of throughput optimality under a time scale separation assumption using the same proof method and Lyapunov function as for Max-Weight. They also indicate a proof method following the lines of Lemma 12 of [RSS08] to prove the time scale separation assumption, similarly to [SS12]. This is done by a careful consideration of the evolution speed of the target invariant measure and the mixing time of the dynamic with fixed fugacities.

The procedure for Q-CSMA has multiple steps and multiple degrees of refinement in terms of applicability. We present here the simplest version. The procedure operates in discrete time.

- At the beginning of time slot k, select an independent set m^k.
- For all v such that $m_v^k = 1$, if no queue neighbor of v was active in the time slot $k-1$, set $\sigma_v(k) = 1$ with probability $p_v(Q_v(k-1))$ and $\sigma_v(k) = 0$ with probability $1 - p_v(Q_v(k-1))$. If node v has some neighbors that were active in time slot $k-1$, let $\sigma_v(k) = 0$.
- Nodes that were not present in m^k do not change their status.

Servers that are active process jobs at unit rate. One of the drawbacks of this method is the need to generate a new independent set m^k at each time slot in order to decide which node to update. This dynamic is a multi-site generalization of the classical Glauber dynamics used for CSMA, with adaptative rates. They use the same type of activation rates as CSMA but the fugacities actually depend on the state of the network in a fashion similar to QB-CSMA. For activation and deactivation rates, they use the same Ψ_+ and Ψ_- as QB-CSMA and result in the same invariant measure. If the queue lengths were fixed, the multi-site dynamic reaches equilibrium faster than the single site dynamic and most of what has been proved for QB-CSMA could realistically be adapted for Q-CSMA. The need to generate an additional independent set for the update decision creates additional communication overhead between nodes and the need for a central supervisor overseeing

the update procedure. Communication between nodes is usually handled with a "control" time slot dedicated to this exchange of information. More information to be transferred means more time allocated to this slot and longer queue sizes.

1.2.4 Polling

When the interference graph is a complete graph, at most one queue can be active at any given time. A scheduling algorithm is then the answer to two questions: "How long does a server stay active?" and "In which order do servers activate?". This is called a polling model. There are multiple popular answers to the first question: exhaustive service (the server only releases the channel when it is empty), gated service (the server releases the channel when all the jobs that were present when it started serving have been processed), fixed service (server processes a fixed number of jobs before releasing), threshold service (the server provides service until its backlog becomes smaller than a fixed threshold), the server may process a random number of requests, etc ... When a queue deactivates, there may or may not be some idling time before the next queue activates. In the literature there is a distinction between polling models with and without switch-over times. For the second question, there are also multiple answers: the next active queue can be random (with a distribution that may depend on the current queue lengths) or fixed (Round-robin with the server visiting each queue one time in each cycle or more complex route).

On a complete interference graph, CSMA and QB-CSMA are in fact a polling systems with switchover time and random activation duration. This equivalence has been exploited to use results for polling systems where the server only processes one job before moving to the next queue and a probabilistic routing policy. It has been used to analize CSMA algorithms where nodes deactivate at a fixed (non-queue-based) rate, see for instance [Cec+16] and [Dor+15]. We will go in greater details about this comparison in Chapter 5, that focuses solely on the case of the complete interference graph with a heavy traffic result.

1.2.5 Load balancing algorithms

Load balancing algorithms answer a simple question. All jobs arrive at a central dispatcher. That dispatcher sends each job to a queue where it will eventually receive some service. The problem is to decide where to send each new job. A naive way to balance the load would be to send jobs to a server uniformly at random. This random balancing is very simple and easy to implement but not efficient at all because this does nothing to prevent situations where some server may spend some time without processing any job and waste service capacity while another one has a lot of job waiting for service. Depending on the amount of communication between nodes, different algorithms have been proposed. The ideal situation is when the state of the network is fully known by an outside observer. In order to maximize the use of all servers, a solution might be to always send jobs to the server that has shortest queue. This is known as JSQ (Join the Shortest Queue). This algorithm requires an exact knowledge of the state of every server's queue length at the time of each arrival. In situation where this is not possible, the Power of d algorithm was designed to limit the required amount of communication overhead and knowledge

on the state of the network. Each time a new job enters the network, the dispatcher selects d servers and probes their queue sizes. Then it routes the incoming job to the shortest of the d queues. If there are n queues, Power of n is JSQ and power of 1 is random routing. Between those two policies there is a large range of load balancing algorithms. In addition, one can mention the celebrated JIQ (Join the Idle Queue) where servers that are empty send a token to the dispatcher signaling their availability. The dispatcher sends jobs in priority to idle queues and needs to make a decision if no server is idle, for instance using a Power of d algorithm. The benefit from this procedure is that it significantly diminishes the communication overhead needed to run the algorithm. Most of the gain in terms of delay is obtained when changing from random routing to Power of 2. See for instance [Mit92],[VDK96] and [Mit01] where the authors establish that the distribution of the waiting time has tails that decreases like $\pi(Q \geqslant k) \approx \lambda^{\frac{d^k-d}{d-1}}$ for the power of d algorithm when the arrival rate is $\lambda < 1$ and $d \geqslant 2$. In comparison, this double exponential decay is much faster than the decay for random routing: in this case all queues evolve independently and have a geometric invariant measure. The rate of decay for the tail distribution of the waiting time is exponential with random routing. Bramson established optimality of the stability region for JSQ in a general setting (general service distribution, multiple types of jobs, ...) in [Bra11]. He then studies a model with a service distribution such that JSQ is not as efficient as one could think in terms of workload.

1.3 Homogenization and State Space Collapse

1.3.1 Stochastic Averaging

The study of a complex interacting system is often untractable. That is especially true when the state space of the process is huge, for instance the space of configurations on a large set. One way to simplify such problems is to assume that parts of this interacting system evolve on a faster time scales than others. In many physical systems, this type of behavior can significantly simplify the analysis. We begin with a simple example with a two stage pendulum, taken from Chapter 10 of [Gri14]. To explain the time scale separation phenomenon, they talk about the simple example of a pendulum whose support is oscillating very slowly. If the support goes slow enough compared to the speed of the pendulum, the system will behave in a simple way where the base can be considered fixed when considering the oscillations of the pendulum but over long period of times, the basis of the pendulum will have moved.

In a Markovian setting, assume that two components of a Markov process evolve on different time scales, and given the slow process the fast component is Markov. We can expect the fast process to reach a steady state equilibrium before the slow process has time to evolve significantly. We will say that one process is fast compared to the other if the mixing time of the fast process given the slow one is "much" smaller than the time it takes the slow process to evolve significantly. If in turn the value of the fast process influences the dynamic of the slow one, this is called a fully coupled stochastic system and stochastic averaging is also called fully coupled stochastic averaging principle. For each value of the slow process, there is a corresponding invariant measure for the fast process. Over long time scales, the

slow process only interacts with the fast process through its steady state average corresponding to the current value of the slow process. In this section, we are going to explore the classical ways of proving that such a behavior occurs. We are going to focus on three approaches: a dynamical system approach by Freidlin and Wentzell in [FW12], a martingale approach developed by Papanicolaou, Stroock and Varadhan [PSV77] and adapted by Kurtz [Kur92], and a corrector function approach developped by Luczak and Norris in [LN13]. The idea of such a behavior was already present in [RSS09] and [SS12] introducing the algorithm that we study in this manuscript.

The first approach presented is the dynamical system framework developed in [FW12]. The randomized process is seen as a small perturbation of the homogenized process and the authors make use of perturbation theory to prove convergence. Second, in [PSV77] the authors develop a general martingale approach that was later reformulated in [Kur92] and applied to loss networks in [HK94]. Third, Luczak and Norris [LN13] also developed a new method which they applied to a variant of the supermarket model. The method in [LN13] allows for a better identification of the limiting process and is closest to the one we developed. Using Poisson equations, we obtain explicit bounds on the distance between the process and the homogenized version. Rigorous proofs of stochastic averaging principles were established for polling systems [CPR95],[CPR98] and [Jen10], for models of distributed hash tables [FR14], for the X model [PW13] and the supermarket model in [LN13].

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 a timescale separation indeed occurs. The type of problems described in the previous paragraph where such situation can arise are usually variation around the theme of a two components Markov process (Q^N, σ^N) such that its generator is given by

$$L^N[f](q,\sigma) = L^{\sigma}_{\mathrm{s},N}[f(\cdot,\sigma)](q) + NL^{q}_{\mathrm{f},N}[f(q,\cdot)](\sigma). \tag{1.3}$$

It is separated in two distinct parts: $L^q_{\mathrm{f},N}$ acts on functions of σ but depends on the value of q and $L^{\sigma}_{\mathrm{s},N}$ acts on functions of q but depends on the value of σ. With this decomposition we assume that Q and σ cannot jump at the same time but this assumption can be weakened. The generator $L^{\sigma}_{\mathrm{s},N}$ is considered *slow* when compared to $NL^q_{\mathrm{f},N}$ where transitions occur on a much faster time scale when $N \to +\infty$. The main idea is that σ evolves so fast that Q only interacts with σ through $\pi^{N,Q}$ the invariant measure of $L^Q_{\mathrm{f},N}$. Heuristically, the goal is to compare the original process with the Markov process of generator given by

$$L_{\mathrm{h},N}[f](q) = \pi^{N,q}[L^{\cdot}_{\mathrm{s},N}[f](q)], \text{ for any } q \in \mathbb{N}^V,$$

with $\pi^{N,q}[f]$ being the integration of the function f with respect to the measure $\pi^{N,q}$. See Section 2.1.1 for more details on notations. The generator $L_{\mathrm{h},N}$ is the homogenized version of L^{σ}_{h} where the transition rates of the slows process are averaged with respect to the invariant measure of the slow process.

Dynamical system approach:

The random setup in [FW12] is taylored for situations where the process of interest is such that

$$(\frac{\mathrm{d}Q^\epsilon}{\mathrm{d}t})(t) = \epsilon b(Q^\epsilon(t), \sigma(t)),$$

with σ a random process, and do a time change $X^\epsilon(t) = Q^\epsilon(\frac{t}{\epsilon})$, which gives

$$(\frac{\mathrm{d}X^\epsilon}{\mathrm{d}t})(t) = b(X^\epsilon(t), \sigma(\frac{t}{\epsilon})).$$

They assume the existence of a function $\bar{b}$ such that for any $\delta > 0$, and any q,

$$\lim_{T\to+\infty} \mathbb{P}\left(\left|\frac{1}{T}\int_t^{t+T} b(q, \sigma(s))\mathrm{d}s - \bar{b}(q)\right| \geqslant \delta\right) = 0. \tag{1.4}$$

If for instance b is bounded, X^ϵ evolves much slower than $\sigma(\frac{\cdot}{\epsilon})$ under suitable conditions. Under suitable assumptions, the homogenization result states that X^ϵ converges to $\bar{q}$ the solution to the ODE problem $f' = \bar{b}(f)$. This result is a direct consequence of the theory of perturbation in dynamical systems. They also prove a second order approximation for normal variations around the limit and a large deviation principle. Condition (1.4) is obvious if σ is an ergodic Markov process but the goal is to have this kind of result for interacting systems. The method has been generalized in [Ver00] and [Ver13] to include different cases of interactions between slow and fast components. Let W_0, W_1, W_2 and W_3 be independent Brownian motions. In [Ver00], the author proves a large deviation principle under suitable conditions on the functions f_0, f_1, f_2, f_3, f_4 and f_5 for systems of the form

$$\begin{cases} \mathrm{d}Q^\epsilon(t) = f_0(Q^\epsilon(t), \sigma^\epsilon(t))\mathrm{d}t + \epsilon(f_1(Q^\epsilon(t), \sigma^\epsilon(t))\mathrm{d}W_1(t) + f_2(Q^\epsilon(t), \sigma^\epsilon(t))\mathrm{d}W_2(t) \\ \mathrm{d}\sigma^\epsilon(t) = \epsilon^{-2} f_3(Q^\epsilon(t), \sigma\epsilon(t))\mathrm{d}t + \epsilon^{-1}(f_4(\sigma^\epsilon(t))\mathrm{d}W_1(t) + f_5(\sigma^\epsilon(t))\mathrm{d}W_3(t). \end{cases}$$

In [Ver13], he proves the same kind of theorem for systems of the form

$$\begin{cases} \mathrm{d}Q^\epsilon(t) = f_0(Q^\epsilon(t), \sigma^\epsilon(t))\mathrm{d}t \\ \mathrm{d}\sigma^\epsilon(t) = \epsilon^{-2} f_1(Q^\epsilon(t), \sigma^\epsilon(t))\mathrm{d}t + \epsilon^{-1} f_2(Q^\epsilon(t), \sigma^\epsilon(t))\mathrm{d}W_0(t). \end{cases}$$

Martingale approach:

Most of the previous works, in particular [FR14], [HK94] and [PW13], rely on the machinery from [Kur92]. The setup is based on some martingale problems: Kurtz assumes the existence of two operator A and B such that the processes

$$f(X_N(t)) - \int_0^t A[f](X_N(s), \sigma_N(s))\mathrm{d}s$$

and

$$g(\sigma_N(t)) - \int_0^t \beta_N B[g](X_N(s), \sigma_N(s))\mathrm{d}s + \delta_g^N(t),$$

with $\beta_N \to +\infty$ and $\beta_N \delta_g^N \to 0$ are martingales for a set of function dense in the space of continuous functions. He considers the occupancy measure Σ^N for the fast process σ^N. It is a random measure which takes values in $\mathcal{P}(\mathbb{R}_+ \times S(G))$ the space

of probability measures on $\mathbb{R}_+ \times S(G)$. It is given by

$$\Sigma^N((0,t)\times A) = \int_0^t \mathbb{1}_{\sigma^N(s)\in A}\mathrm{d}s,$$

and they manage to prove that (Q^N, Σ^N) is tight so they simply need to identify the limit. To do that they rely on martingale arguments. Assume the existence of the limiting generators given by $L_{\mathrm{f}}^{\infty,q} \coloneqq \lim_{N\to+\infty} L_{\mathrm{f},N}^{q}$ and $L_{\mathrm{s}}^{\infty,\sigma} \coloneqq \lim_{N\to+\infty} L_{\mathrm{s},N}^{\sigma}$. Notice that any limiting point $(\bar{q}, \Sigma)$ of (Q^N, Σ^N) must be such that for any f, g bounded, such that $L[f]$ and $L[g]$ are bounded, the processes

$$f(\bar{q}(t)) - \int_0^t \int L_{\mathrm{s}}^{\infty,\sigma}[f](\bar{q}(s))\Sigma(\mathrm{d}s, \mathrm{d}\sigma),$$

and

$$\int_0^t \int L_{\mathrm{f}}^{\infty,\bar{q}(s)}[g](\sigma)\Sigma(\mathrm{d}s, \mathrm{d}\sigma)$$

are martingales. This is true because their prelimit equivalent are. Because of their assumption on the time scales of the two processes, the martingale component of the slow process must converge 0. The second martingale is continuous and has finite variation as an integral of a bounded function, so it must be constant to its initial value: 0. As it happens, the first marginal of Σ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}_+$. There exists a measure valued process γ such that

$$\Sigma((0,t)\times A) = \int_0^t \gamma_s(A)\mathrm{d}s.$$

Because of this decomposition and their assumptions, they are able to identify γ_s as the unique invariant measure of $L_{\mathrm{f}}^{\infty,\bar{q}(s)}$: for any $t \geqslant 0$, and g in the domain of $L_{\mathrm{f}}^{\infty,q}$ for all q,

$$\int_0^t \int L_{\mathrm{f}}^{\infty,\bar{q}(s)}[g](\sigma)\gamma_s(\mathrm{d}\sigma)\mathrm{d}s = 0$$

Thus, for any g in the domain of $L_{\mathrm{f}}^{\infty,q}$ for all q,

$$\int L_{\mathrm{f}}^{\infty,\bar{q}(s)}[g](\sigma)\gamma_s(\mathrm{d}\sigma) = 0$$

for almost every $s \geqslant 0$, and so if $D(L_{\mathrm{f}}^{\infty,q})$ is dense in the space of continuous bounded functions, γ_s is the invariant measure of $L_{\mathrm{f}}^{\infty,\bar{q}(s)}$ so we write $\gamma_{\bar{q}(s)}$ instead to highlight the dependency in q. Once Σ has been identified, the first martingale gives an identification of $\bar{q}$ in terms of Markov processes: $\bar{q}$ must be a Markov process of generator

$$\bar{L}[f](q) \coloneqq \int L_{\mathrm{s}}^{\infty,\sigma}[f](q)\gamma_q(\mathrm{d}\sigma).$$

Corrector function approach:

The starting point for the method in [LN13] is to extract slow coordinates of the form $X \coloneqq x(\xi) = (x_i(\xi))_{i=1,\dots,n}$ and a fast component $Y \coloneqq y(\xi)$ from a Markov process $(\xi(t))_{t\geqslant 0}$. Let's define two accessory processes and state spaces: $\xi(t) \in \mathcal{X}$ and $Y(t) \in I$ both state space of countable. Let us define the projections on the slow and fast variables $x : \mathcal{X} \to \mathbb{R}^n$ and $y : \mathcal{X} \to I$. If L^0 is the generator of ξ, the

key is to assume the approximations

$$L^0[x](\xi) \approx b(x(\xi), y(\xi)),$$

and for any $\xi^0 \in \mathcal{X}$ and $y' \in I$ such that $y' \neq y(\xi^0)$,

$$\sum_{\xi' \in \mathcal{X},\, y(\xi')=y'} L^0(\xi^0, \xi') \approx g(x(\xi^0), y(\xi^0), y')$$

hold for a suitable b and g. The function g must be such that for any $x^0 \in \mathbb{R}^n$, the operator given by $(g(x^0, y_1, y_2))_{y_1,\, y_2 \in I}$ is a Markovian generator having a unique invariant distribution π^{x^0}. In other words, we have to be able to approximate the transition rates for $Y(t)$ using only the values of $X(t)$ and $Y(t)$. They define the average drift for the slow process:

$$\bar{b}(x^0) = \pi^{x^0}[b(x^0, \cdot)].$$

Using "corrector function", they prove explicit bounds on the difference between the slow process and $\bar{q}$ the solution to the ODE

$$\begin{cases} \dot{f} & = \bar{b}(f) \\ f(0) & = x^0 \end{cases}$$

The general idea of the method is that solutions to Poisson equations provide a "corrector function" giving an alternate description for

$$b(x^0, y^0) - \bar{b}(x^0).$$

In order to use a Gronwall inequality, they assume that $\bar{b}$ is a Lipschitz function. We will see in Chapter 5 an example where this condition fails.

1.3.2 Poisson equation and Stein's method

The problem of Poisson equations and their solutions isn't recent and the literature on the subject is dense. We refer the reader to Section 2.2.2 for a formal definition of Poisson equations and the explicit formula for their solutions. If L^0 is the generator of a Markov process of invariant measure π and g a function such that g is integrable with respect to π, a Poisson equation is the equation with unknown ϕ,

$$L^0[\phi](x) = g(x) - \pi[g],\ \pi[\phi] = 0.$$

Since the seminal paper [Nev71] and book [Rev84] outlining Poisson equations and their role in the classical potential theory for Markov processes much has been done. See for instance [PV01], [PV03] and [PV05] for a detailed account of its application for diffusion approximations in a SPDE context using solutions to Poisson equations. They have already been used to prove homogenization in discrete state space as well, see [PSV77] and [LN13]. The general idea in these papers is to use the solutions to Poisson equations as "corrector functions". These functions control the difference between the transition rates of the slow process and their homogenized versions. One of the contribution of this thesis in Chapter 3 is to provide bounds on the regularity of solutions to Poisson equations when the generator depends on a parameter. To the best of our knowledge few such bounds exist in our setting, see for instance

[LN13] where the authors proved similar bounds.

One of the most common use of the Poisson equation today would be the celebrated Stein's method. Introduced by Stein in [Ste72], it was first used to bound the distance between the empirical measure for a sum of dependent random variables and a Gaussian distribution. Let W be a random variable of the distribution of interest. The general idea is to analyze the difference between the averages of $g(W)$ and $g(N)$ with N a gaussian variable by finding a suitable function f such that

$$g(W) - \mathbb{E}\left[g(N)\right] = f'(W) - Wf(W)$$

and then use the structure of W to estimate $\mathbb{E}\left[f'(W) - Wf(W)\right]$. By replacing f with F' (and f' by F'') this equation can be interpreted as a Poisson equation

$$g(W) - c = \mathcal{A}[F](W),$$

with $\mathcal{A}$ the generator of an Ornstein-Uhlenbeck process given by

$$\mathcal{A}[f](x) = f''(x) - xf'(x).$$

The method was later improved to be able to handle different distributions. One of the first refinement was by Chen [Che75], it enables Poisson approximations to be obtained finding a functional application $g \mapsto f_g$ such that it solves another Poisson equation in g, given by

$$wf(w) - \lambda f(w+1) = g(w) - c,$$

and even convergence rates as exposed in [Bar88]. Later diffusion approximations were also considered by Barbour [Bar90]. Apart from some *ad hoc* tricks for sum of dependent random variables, the common point of these methods is the use of a "Stein's kernel" which is simply a Markov generator which has the target probability as invariant measure. Let $\mathcal{H}$ be a functional space. We define the $d_{\mathcal{H}}$ distance between two probability measures as

$$d_{\mathcal{H}}(\mu,\nu) = \sup_{h\in\mathcal{H}} \left| \int_X h(x)\mu(\mathrm{d}x) - \int_X h(x)\nu(\mathrm{d}x) \right|.$$

By choosing $\mathcal{H}$, we can obtain some usual distances: for instance by taking $\mathcal{H}$ to be the space of 1-Lipschitz functions, we get the Wasserstein distance and by taking $\mathcal{H}$ to be indicators of Borelian sets, we get the total variation distance. The idea is then to write the distance between two probability measures as the supremum over a suitable class of functions of the mean of Stein's kernel applied to the solution to the Poisson equation. Let us define M_1 a random variable of distribution μ. Let $\mathcal{A}$ be a Markov generator of invariant measure ν and for any suitable f, let ϕ_f be the solution to the Poisson equation

$$f(x) - \nu[f] = \mathcal{A}(\phi_f)(x).$$

We get

$$d_{\mathcal{H}}(\mu,\nu) \leqslant \sup_{h\in\mathcal{H}} \mathbb{E}\left[\mathcal{A}[\phi_h](M_1)\right].$$

Once this is done, the method requires to use the specific structure of M_1 and the generator $\mathcal{A}$ to obtain some bounds on the error between the two distributions. See

[Ros11] for more details and a survey of the method in different contexts. Choosing the right Markovian dynamic is not always easy and is of independent interest. The interpretation of Stein's method in terms of infinitesimal generator enabled the method to be used for an even wider range of distribution using solutions to the respective Poisson equations.

Stein's method has also been used in the context of queueing networks. Three of he first papers to use this method were [Gur14], [BDF16] and [BD17]. In those papers, they establish some explicit error bounds for the difference between the diffusion approximations and their steady state average for the Erlang-A, Erlang-C models and M/Ph/n+M queues. In [BDF16], the authors discuss the Erlang-A and Erland-C models and prove that the rate of convergence to a limiting diffusion process scales like $\frac{1}{\sqrt{\rho}}$ when the load of the network (as will be defined in 1.4.1) $\rho \to 1$. We mention as well [GW19] that provides explicit bounds for the heavy traffic approximation of a single queue. In an M/G/1 queue with arrival rate λ and arbitrary job size distribution S, let's call $\rho = \lambda \mathbb{E}[S]$ and W the stationary waiting time. It is well known (see for instance [Kin62]) that

$$(1-\rho)W \to \frac{\mathbb{E}\left[S^2\right]}{2\mathbb{E}[S]} Z \text{ as } \rho \to 1$$

in probability with Z an exponential variable with parameter 1. Their bound pertains to d_{W} the Wasserstein distance between μ and ν the distributions of $\frac{2\mathbb{E}[S](1-\rho)}{\mathbb{E}[S^2]\rho} W$ and Z respectively. They prove that

$$d_{\mathrm{W}}(\mu,\nu) \leqslant 4\frac{4\mathbb{E}\left[S^3\right]\mathbb{E}[S](1-\rho)}{3(\mathbb{E}[S^2])^2\rho}.$$

They also manage to prove that

$$(1-\rho) \leqslant d_{\mathrm{W}}(\widetilde{\mu},\nu) \leqslant (1-\rho)(1+4\frac{\mathbb{E}\left[S^3\right]\mathbb{E}[S]}{(\mathbb{E}[S^2])^2}),$$

with $\widetilde{\mu}$ being the distribution of $\frac{2\mathbb{E}[S](1-\rho)}{\mathbb{E}[S^2]} W$. It means that the $O(1-\rho)$ scaling is in fact optimal. They provide two different proofs for this result both using Stein's method: a coupling approach and a generator approach. See [GW19] and the references therein for more details and applications of Stein's method.

1.3.3 State Space Collapse, Skorokhod problem

The term "State Space Collapse" comes from the works of Reiman [Rei84a] and [Rei84b]. In both those papers, the author proves convergence of the total number of jobs in the network to a reflected Brownian motion. The first paper considers an open network comprised of two queues where jobs flow randomly from one server to another before eventually leaving, the second considers priority queues, a load balancing situation implementing JSQ, and networks with one 'bottleneck' station. In all those examples the multidimensional process of queue lengths converges to a one dimensional process. When the limiting process is random it is driven by a 1-dimensional Brownian motion. In [Bra98] and [Wil98], a technical condition (resource pooling) that ensures that when the network is heavily loaded, the state space of an open multiclass Jackson network collapses to a one dimensional manifold.

The resource pooling condition ensures that only one node is critically loaded. For Max-Weight scheduling, [SW12] proved a collapse of the state space even when the resource pooling condition fails, but the manifold needs not be 1-dimensional. In [KW12], Kang and Williams made progress towards a heavy traffic result for Max-Weight without resource pooling by assuming that all stations are heavily loaded. They consider a $N \times N$ switched network (with N^2 communication links) and prove that the process of queue lengths lives on a manifold of dimension $2N - 1$. Before we go in further details about the proof method for convergence, the general idea behind the state space collapse is that there is a lower dimensional manifold that is attractive for the sample paths of the limiting fluid limits process.

In [Bra98], [Wil98], [SW12] and [KW12], the authors characterize the limiting process using a Skorokhod problem, introduced in [Sko61a] and [Sko61b] in dimension 1. The goal was to study a diffusion with different reflection property when reaching the boundary. It was later generalized to a multidimensional process in [Tan79]. Let $d \in \mathbb{N}$ and P a $d \times d$ matrix, with spectral gap less than 1. If Y is a càdlàg function with values in $\mathbb{R}^d$ such that $Y(0) \geqslant 0$, there exists a unique couple of functions $X_Y = (X_{Y,i}, i = 1, \dots, d)$ and $R_y = (R_{Y,i}, i = 1, \dots, d)$ càdlàg functions such that for $t \geqslant 0$ and $1 \leqslant i \leqslant d$,

- $X_Y(t) = Y(t) + (I - P)R_Y(t)$;
- $X_{Y,i}(t) \geqslant 0$ and $R_{Y,i}$ is non-decreasing with $R_{Y,i}(0) = 0$.
- (X, R) satisfies the reflection condition

$$\int_0^{+\infty} X_{Y,i}(s)\mathrm{d}R_{Y,i}(s) = 0.$$

For any Y, there exists a unique explicit solution (X_Y, R_Y). Moreover, the application $Y \mapsto (X_Y, R_Y)$ solutions to the Skorokhod problem is continuous for the topology of uniform convergence. See Theorem 1 of [HR81]. If a sequence of processes are solutions to a sequence of Skorokhod problem and the processes converge, any limiting point must be the unique solution to the limiting Skorohod problem.

More recently, the authors of [MS16] proved a State Space collapse in steady state for Max-Weight on a switched network. In steady state, the drift of certain functions must be null, which can be seen as constraints on the queue lengths. This method has been introduced in [ES12]. This state space collapse ensures that the queue lengths are always in a cone not touching 0. The main result of [MS16] is that Max-Weight is delay optimal in steady state.

1.4 Functional limit theorems

Functional limit theorems are functional versions of laws of large numbers and the central limit theorems. Let $(X_N)_{N\in\mathbb{N}}$ be a sequence of independent identically distributed random variables. The law of large numbers (see [Kal02], Theorem

3.23) states that as long as $\mathbb{E}[|X_1|] < +\infty$,

$$\frac{1}{N}\sum_{k=1}^{N} X_k \to \mathbb{E}[X_1] \text{ almost surely as } N \to +\infty.$$

Similarly, define $X(t)$: the linear interpolation of

$$X(K) \coloneqq \sum_{k=1}^{K} X_k.$$

For any $t > 0$,

$$\frac{X(Nt)}{N} \to t\mathbb{E}[X_1] \text{ almost surely as } N \to +\infty.$$

In fact, the convergence is almost sure for the topology of uniform convergence over compact time sets. Almost sure uniform convergence over compact time set means that for any $T < +\infty$,

$$\sup_{t \leqslant T} \left| \frac{X(Nt)}{N} - \mathbb{E}[X_1]\, t \right| \to 0 \text{ almost surely as } N \to +\infty.$$

Similarly to the law of large numbers, if we take a sample that grows with N and rescale the cumulative samples by the same parameter N, it converges to a deterministic limit. A direct consequence is that if we assume that $\mathbb{E}[X_1] = 0$, for any $0 < T < +\infty$,

$$\sup_{t \leqslant T} \left| \frac{X(Nt)}{N} \right| \to 0 \text{ almost surely as } N \to +\infty.$$

The fact that the limiting process is constant suggests that we need to look further in time to see variation around the initial value. There is a second important limit theorem called central limit theorem, see Proposition 5.9 of [Kal02]. If $(X_N)_{N \in \mathbb{N}}$ is a sequence of i.i.d. random variables with 0 mean and variance 1,

$$\frac{1}{N}\sum_{k=1}^{N^2} X_k \Rightarrow \mathcal{N},$$

with $\mathcal{N}$ a standard Gaussian random variable and $\Rightarrow$ the convergence in distribution, defined in 2.3. A direct consequence of the central limit theorem is that for any $t \geqslant 0$,

$$\frac{X(N^2 t)}{N} \Rightarrow \mathcal{N}',$$

with $\mathcal{N}'$ a Gaussian variable with 0 mean and variance t. Once again, the result is uniform in a functional space: as processes, uniformly over compact time sets,

$$\left(\frac{X(N^2 t)}{N}\right)_{t \geqslant 0} \to B \text{ in distribution},$$

with B a standard Brownian motion.

The kind of results that we prove in this manuscript is similar to those behaviors.

In Chapter 4, we will prove an equivalent to the law of large number for QB-CSMA. In the context of queueing networks, this is called fluid limits. In Chapter 5, we will prove result similar to a central limit theorem but uncommon in two different ways.

1.4.1 Fluid limits

When considering stability of queueing networks, in terms of tightness of its distribution, it is common to consider a rescaling in time and space of the process and let the scaling parameter go to $+\infty$. When the scaling is the same in time and space, the limiting points (if they exist) are called "fluid limits". When unique, the fluid limit can serve as a first order approximation of the process for large N and stability of the original process can be deduced from the behavior of the fluid limit. See Chapter 9 of [Rob03] for an introduction on the subject and some examples. This method is more than 40 years old: it has been introduced in [MM79] where the authors use it to establish the transient/recurrent behavior of random walks on $\mathbb{Z}^2$ and $\mathbb{Z}^3$. We next discuss further the fundamental papers [RS92] and [Dai95] and their main question.

For the type of communication schemes we introduced, we can define the load of each node as the ratio between the average number of jobs entering the queue and the average service rate. A natural necessary condition for stability for this kind of network is that the load of each node is smaller than 1, meaning that there is on average less work arriving in the system than it is able to process. If this condition is not met, there is a queue where the amount of work arriving is greater than the service capacity of this queue and thus it will increase endlessly. Whether or not this condition is also sufficient in a given communication scheme still is a central question in queueing theory.

To illustrate this notion of load, let's present Jackson Networks, first introduced in [Jac57]. Consider n queues and K types of jobs each type of job has a different size (for instance all have an exponential distribution but with a different parameter). Jobs of type k first get served in queue $s(k)$ and then follow a Markov chain on the set of servers. Once a job is served in queue i, it becomes a job of type j with probability $p_{i,j}$ and leaves the network with probability $1 - \sum_j p_{i,j}$. Jobs of type j are served in station $s(j)$. The matrix $(p_{i,j})_{i,j=1,\dots,n}$ is sub-Markovian. This model is a multiclass extension of Jackson networks introduced, where queues and types are in bijection. Assume all service rates are given by $\mu > 0$. The load of a node is defined as

$$\rho_v i = \frac{\lambda_i p_{i,i} + \sum_{j \neq i} \lambda_j p_{v,k}}{\mu},$$

the ratio between ingoing and outgoing traffic at node v. If $\rho_v < 1$ for all $v \in V$, the process of queue lengths is stable.

In [RS92], the authors study a network of two queues in tandem. Jobs arrive at each of the queues, and flow from one queue to the other after a service. Jobs enter the other queue waiting line when they complete a service in their queue of origin. Once they complete their second service, jobs leave the system. They study two service policies: FCFS and a priority queue, both with exponential job sizes. In the priority queue, a server will only serve jobs that have already been served by

the other queue when it does not have any of its own request waiting. If we number the queues 1 and 2, queue i receives jobs of type i at rate λ_i. Those jobs will flow from queue i to queue $j \neq i$ when they complete a service in queue i. When a type i job is processed at server k, it takes a time that has an exponential distribution with parameter $\nu_{i,k}$. For $(i,j) = (1,2)$ or $(2,1)$, they define the load of each node as the mean amount of work they receive:

$$\rho_i = \lambda_i \nu_{i,i} + \lambda_j \nu_{j,i}.$$

The authors prove that $\rho_i < 1$ for all i is sufficient for stability when using the FCFS service discipline. More surprisingly, they also give a simple example where this condition is not sufficient with the priority queue. If type 1 jobs have an absolute priority over type 2 jobs when being served by queue 1, they construct an example where $\rho_1 \vee \rho_2 < 1$ but the fluid limits diverge as $t \to +\infty$. Bramson in [Bra94] provided one of the first simple FCFS example where $\rho < 1$ is not enough for stability. He studies a model that is similar to the priority queue: jobs that enter at one of two stations i and j. A job arriving at station i completes an exponential service at this station. Once this service is done, it has to complete a fixed number of exponential service in the other queue before receiving a final service at its original queue. If the amount of time a job of type i has to spend in station j is very large, heuristically, the time it takes for jobs of type i to leave station j is huge compared to the time it takes a job of type j to leave the same station, which resembles the behavior of the priority queue. Since the server processes requests in a FCFS fashion, it will have time to fully complete a service of the other type only if none of its own jobs are pending because each time the replicating job finishes, it goes back to the end of the queue to await for another service a large number of time.

In [Dai95], the author used a fluid limit to give a general necessary condition for the stability of multiclass multihop extensuin of Jackson networks. The result of Dai states that if the fluid limit reaches 0 in finite time and stays absorbed, the process is stable. He applies this criterion to prove that $\rho_i < 1$ for all i is sufficient for stability in specific cases. To handle possible reflections, the authors of [DM95] use a multidimensional Skorokhod problem. The examples are a general Jackson network, a single queue with multiple types of jobs or a single type Jackson network with K nodes and routing matrix $p_{k,k+1} = 1$ for $k < K$ and 0 everywhere else. See also [DM95] where the authors derive the limits of moments of queue lengths.

The main result of [GBW14] is a fluid limit approximation. In QB-CSMA with polynomial activation rates, when the exponent of the polynomial is too big, queues are unable to release the channel until they are almost empty. This behavior is similar to the so-called "Random Capture Algorithm" (RCA) introduced in [FPR10]. They introduce it for a polling system where only one queue can be active at any given time. With the RCA scheduling policy, an active server remains active until it has no job to process, with server activating either in a random or a fixed order. This procedure can be extended to general interference graphs. To generalize it, simply assign an activation rate. For queues with no active neighbors in the interference graph let an exponential clock run and the first one to tick is allowed to activate until it is empty. The main result of [FPR10] is throughput optimality of RCA for K-partite interference graphs and a description of the fluid limit. Fluid limits serve both as a qualitative and quantitative approximation for queue lengths. Absorption of the fluid limit in 0 implies stability. In addition, the delay before a

new job gets some service is related to the number of jobs in the system through Little's law, stating that the average number of customers in a system is equal to the product of the average arrival rate and the average time spent in the system.

The fluid approximations of QB-CSMA with polynomial activation/deactivation ratio from [GBW14] has some interesting properties. In fact, they prove that if the exponent is large enough, on the fluid scale, queues cannot deactivate before being (almost) empty. In fact, the authors of [GBW14] prove that in a specific graph topology, QB-CSMA with polynomial activation/deactivation ratio with parameter $a > 1$ and RCA of [FPR10] have the same fluid limits (and thus the same stability region for instance). We expect this to be true in any graph topology: if $a > 1$ and the queue is of order N, if the queues were fixed, the deactivation time would have been of order N^a. The real queue lengths decreases linearly during that time. The back-off times are exponentially distributed. The memory-less property ensures that there is no early deactivation until the queue is (almost) empty. Essentially, when the queue is large, the time it takes for a queue to deactivate will remain larger than the time for the queue to be almost empty with high probability. Using this fluid limits arguments, they prove that in some graph topology, RCA creates some inefficiency and is not throughput optimal. The results of Chapter 4 lie in the complementary of [GBW14] : consider a small enough so that the schedule can rapidly alternate between activity states and averages around a steady state equilibrium. The authors of [GBW14] proved that in any graph topology under a homogenization hypothesis and if the fluid limits converge, the queueing process is stable.

1.4.2 Heavy traffic

When implementing real life network, administrators usually want to minimize costs. It is an incentive to responsively design the number of servers, service capacity, etc... It is then relevant to evaluate the performance of an algorithm when the network is as loaded as it can be while remaining stable, corresponding to an optimal use of the resource. This is the central question behind heavy traffic analysis of queueing networks. Another way to look at it is to consider a dimensioning problem as in [HW81]. The authors answer the questions of the type: "how many server should there be in a load balancing problem as the arrival rate goes to inifinity. The result in [HW81] states that if the arrival rate is $N \times \lambda$ the total number of servers should greater than $N + c\sqrt{N}$ to maintain the queue lengths stable. In a broad sense, this type of questioning is close to 60 years old. Some of the first mentions come from a series of papers by Kingman: [Kin61], [Kin62]. These seminal papers mainly focus on the performances in steady state and prove results on the asymptotic of the steady state delay as the load goes to 1. Their proof methods rely on Laplace transforms. The general idea about heavy traffic result is to take a sequence of networks under a specific algorithm and let the arrival rates λ^N converge to a λ on the border of the stability region. To ensure that the process of queue lengths is stable for $N < +\infty$, λ^N approach λ from inside the stability region. Usually when at criticality, some key element of the fluid limit remains constant and we need to go at a faster time scale to see fluctuations around the limit. The critical behavior of most queueing networks that we know of is akin to a functional CLT: the time scale usually needs to be the square of the space scale and the limit is a diffusion, typically a reflected Brownian motion. The most common is a scaling N in time

and $\sqrt{N}$ in space. With λ^N arrival rates and μ^N departure rates, the usual heavy traffic assumption is,

$$\forall v,\ \sqrt{N}\left(\lambda_v - \lambda_v^N\right) \to c_v \in \mathbb{R} \text{ as } N \to +\infty,$$

with λ be on the border of the stability region. All coordinates must converge to criticality at the same rate and the vector $(c_v)_{v\in V}$ usually gives an indication about the drift of the limiting process. Sometimes, it is assumed that $c_v > 0$ to ensure that λ^N stays inside the stability region for all $N \in \mathbb{N}$.

As we already mentioned the heavy traffic result is a functional form of the central limit theorem. Instead of considering the limit of the steady state distribution, a popular approach is to prove functional limit theorems. In the same way that fluid limits serve as a first order approximation, the heavy traffic limit provides a long term approximation when the fluid limits are constant. In [Rei84a], Reiman proved that the process of queue lengths converges to a reflected Brownian motion with drift c in an open Jackson network as $N \to +\infty$ under the heavy traffic assumption mentioned above and a mild moment assumption on inter-arrival and service distributions. The proof is based on a multidimensional Skorohod problem and *ad-hoc* computations. In [Bra98]; Bramson studied a switched multiclass multihop network in heavy traffic. Jobs of different types have the same service requirement when at the same station. He considers several service disciplines: FCFS, Head of line processor sharing (HLPPS with servers sharing the resource evenly but each of them proceeding in a FCFS fashion serving only the first job of each type) and static priority queues (certain type of jobs have a strict priority over other types when served by a given queue). With those service disciplines, Reiman gives a full description of the fluid limit and proves that the process of queue lengths collapses on a lower dimensional manifold for large time horizons. Finally he establishes a Semi-martingale Reflected Brownian Motion as a heavy traffic limit. An additional assumption of "complete resource pooling" ensures that the state space collapse is one dimensional. Bramson in [Bra98] proves that the line of the state space collapse must be attractive for any fluid sample path. Almost simultaneously with the paper by Bramson, Williams published another paper [Wil98] on the subject of heavy traffic for an open multiclass Jackson network. Under the assumption that a state space collapse holds on a 1 dimensional manifold, she proves some new diffusion approximations for Kelly networks; queueing models with service rates depending on queue lengths with a specific balance property ensuring a product form for the steady state average of queue lengths.

In [MS16], the authors prove a state space collapse for the steady state distribution of queue lengths. This property helps them prove asymptotic optimality of Max-Weight from a delay standpoint for a switched network. See [Whi02] for an overview of heavy traffic results in a standard setting. The heavy traffic result in this thesis falls outside of this category of conventional results with both an unusual scaling and limit. One of the first example of unconventional heavy traffic result was published in [HW96]. The authors studied a closed system of two queues with a particular priority scheme. Because of their particular scheme, over a time frame of size N, some coordinate are of size N while others are of size $\sqrt{N}$ leading to a "mixed" scaling. The limit they obtained is constructed from a Brownian motion but is not the usual reflected Brownian semi-martingale. Some other recent works investigate single-server queues with nonstandard heavy traffic limits. For instance, Atar and Cohen [AC19] study a multiclass single-server queue which, subject to the

usual CLT scaling, converges to a Walsh Brownian motion. Another example is Puha [Puh14], who studies the SRPT (Shortes Remaining Processing Time) policy: there the scaling is nonstandard but the limiting diffusion is conventional, i.e., the heavy traffic limit is a reflected Brownian motion.

Chapter 2

Technical elements

Contents

2.1 Introduction 27
2.1.1 General notations 27
2.1.2 Model description 30
2.1.3 Localization 33
2.2 Functional analysis 33
2.2.1 General notions 34
2.2.2 Poisson equation 36
2.3 Glauber dynamics for QB-CSMA 38
2.3.1 Historical results 38
2.3.2 Spectral gap of Glauber dynamics for QB-CSMA 39
2.3.3 Influence of threshold function 44

2.1 Introduction

2.1.1 General notations

First, let's gather some notations that will be used throughout the manuscript. Introduce V, a finite set of n nodes. The usual L_b-pseudonorm on $\mathbb{R}_+^V$ is denoted by $\|\cdot\|_b$ for $b > 0$: for any $x \in \mathbb{R}_+^V$ and $b > 0$ it is defined by $\|x\|_b = (\sum_{v\in V} x_v^b)^{1/b}$. Similarly, for any $b > 0$ and $q \in \mathbb{R}_+^V$ we use $s_b(q)$ to denote $\sum_{v\in V} q_v^b$. Let $G = (V, E)$ be a graph and let $S(G)$ be the set of stable sets of G.

For any $n > 0$, $f : \mathbb{R}^n \to \mathbb{R}$, if $U \subset \mathbb{R}^n$, we write $\|f\|_{U,\infty} \coloneqq \sup_{x\in U}|f(x)|$ and $\|f\|_\infty \coloneqq \sup|f|$. With a slight abuse in notation it is possible to use the infinity norm for vectors and write the supremum norm $\|q\|_\infty = \max_v|q_v|$ for $q \in \mathbb{R}^n$.

Whenever $f : \mathbb{R}^n \to \mathbb{R}$ is smooth enough, for any $i, j \leqslant n$ we denote by ∂_i its

partial derivative along q_i and ∂^2_{ij} its second-order derivative along q_i and q_j, i.e.,

$$\partial_i f = \frac{\partial f}{\partial q_i} \text{ and } \partial^2_{ij} f = \frac{\partial^2 f}{\partial q_i q_j}.$$

We will alto consider the discrete partial derivatives $\Delta^N_{\pm,i} f$ for a function $f : \mathbb{R}^n \to \mathbb{R}$, given by

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

with $e^i \in \{0,1\}^V$ the vector with i in its i^{th} component and 0 elsewhere. Thus, $N\Delta^N_{\pm,i} f \to \pm\partial_i f$ as $N \to \infty$ for f differentiable. If $f : \mathbb{R} \to \mathbb{R}^n$, we design its derivative by f'.

For convenience, we will use the letter C to denote positive and finite constants, whose precise value is irrelevant and that may change from line to line. It is allowed to depend on λ, a, the interference graph, localization constants and parameters in specific lemmas. It does not depend on quantities that will change when scaling the process like the size of queue lengths or time horizons.

Let $\mathcal{X}$ be a countable set, π a probability measure on $\mathcal{X}$ and f a function on $\mathcal{X}$. If f is integrable with respect to π we write $f \in \mathbb{L}^1(\pi)$. For any $f \in \mathbb{L}^1(\pi)$ we use the bracket notation to denote integration:

$$\pi[f] \coloneqq \sum_{x \in \mathcal{X}} \pi(x) f(x).$$

One of the key elements from standard Markov theory that we will use repeatedly is the martingale problem (or martingale decomposition). If L^0 is the generator of a Markov jump process, let $\mathcal{D}(L^0)$ be the domain of L^0. For any $f \in \mathcal{D}(L^0)$, we can write

$$L^0[f](x) = \sum_{x' \neq x} L^0(x, x') \left(f(x') - f(x)\right).$$

Let Γ^0 be the *carré du champ* operator associated to the generator L^0: for any f such that $f, f^2 \in \mathcal{D}(L^0)$, $\Gamma^0[f]$ is given by

$$\Gamma^0[f](x) = L^0[f^2](x) - 2f(x)L^0[f](x).$$

Elementary computations show that

$$\Gamma^0[f](x) = \sum_{x' \neq x} L^0(x, x') \left(f(x') - f(x)\right)^2.$$

Proposition 2.1
If L^0 is the generator of $(l_t)_{t \geqslant 0}$ a non explosive Markov process, for any function f such that f and $f^2 \in \mathcal{D}(L^0)$ the process

$$M_f(t) = f(l_t) - f(l_0) - \int_0^t L^0[f](l_s)\mathrm{d}s$$

is a local martingale with increasing process

$$\langle M_f \rangle (t) = \int_0^t \Gamma^0[f](l_s)\mathrm{d}s.$$

Proof sketch. We give here an idea of the proof, see Chapter VIII, Lemma 3.68 of [JS03] for a proof of this result. The proof uses notions of semi-martingales and stochastic integration that we did not define here so we will just explain heuristically the identification of the quadratic variation. Let us define $\mathcal{F}_t \coloneqq \sigma(l_u, u \leqslant t)$. To prove that M_f is a martingale, it suffices to check that

$$\mathbb{E}\left[M_f(t+s) - M_f(t) \mid \mathcal{F}_t\right] = 0.$$

To prove the previous equality, one can use the Kolmogorov equation

$$\frac{d}{dt}\mathbb{E}\left[f(l_t)\right] = \mathbb{E}\left[L^0[f](l_t)\right],$$

and obtain the result using Markov property.

More precisely,

$$\begin{aligned}
\mathbb{E}\left[M_f(t+s) \mid \mathcal{F}_t\right] &= \mathbb{E}\left[f(l_{t+s}) \mid \mathcal{F}_t\right] - \mathbb{E}\left[\int_0^{t+s} L^0[f](l_u)\mathrm{d}u \mid \mathcal{F}_t\right] \\
&= \mathbb{E}_{l_t}\left[f(l_s)\right] - \int_0^t L^0[f](l_u)\mathrm{d}u - \mathbb{E}\left[\int_t^{t+s} L^0[f](l_u)\mathrm{d}u \mid \mathcal{F}_t\right] \\
&= f(l_t) + \mathbb{E}_{l_t}\left[\int_0^s L^0[f](l_u)\mathrm{d}u\right] - \int_0^t L^0[f](l_u)\mathrm{d}u - \mathbb{E}\left[\int_t^{t+s} L^0[f](l_u)\mathrm{d}u \mid \mathcal{F}_t\right] \\
&= M_f(t),
\end{aligned}$$

because $\mathbb{E}_{l_t}\left[\int_0^s L^0[f](l_u)\mathrm{d}u\right] = \mathbb{E}\left[\int_t^{t+s} L^0[f](l_u)\mathrm{d}u \mid \mathcal{F}_t\right]$ by the Markov property.

To identify the quadratic variation, let $F_t \coloneqq \int_0^t L^0[f](l_u)\mathrm{d}u$ and $G \coloneqq M_f + f(l_0) - F$, i.e. $G_t = f(l_t)$ for any $t \geqslant 0$. In the same fashion, we define $\widetilde{F}$ and $\widetilde{G}$ from M_{f^2} and $f(l_u)^2$. By definition, $G^2 = \widetilde{G}$. We have

$$\widetilde{G} = \widetilde{G}_0 + M_{f^2} + \widetilde{F}.$$

Using stochastic integrals for Poisson processes and formula I.4.45 of [JS03] we get

$$\widetilde{G} = G^2 = G_0^2 + 2G_- \cdot M_f + 2G_- \cdot F + \langle M_f \rangle.$$

Both M_{f^2} and $2G_- \cdot M_f$ are martingales starting at 0. Similarly, $\widetilde{F}$ and $2G_- \cdot F + \langle M_f \rangle$ have finite variation by definition of $\langle M_f \rangle$ and because $L^0[f](l_u)$ and $L^0[f^2](l_u)$ are bounded. Since $\langle M_f \rangle$ is predictable, by Proposition 4.23 in Chapter I of [JS03], such a decomposition is unique and we get $M_{f^2} = 2G_- \cdot M_f$ and $\widetilde{F} = 2G_- \cdot F + \langle M_f \rangle$.

Finally,

$$\begin{aligned}
\langle M_f \rangle (t) &= \widetilde{F} - 2G_- \cdot F \\
&= \int_0^t L^0[f^2](l_u)\mathrm{d}u - 2\int_0^t f(l_{u-})L^0[f](l_u)\mathrm{d}u \\
&= \int_0^t L^0[f^2](l_u)\mathrm{d}u - 2\int_0^t f(l_u)L^0[f](l_u)\mathrm{d}u \\
&= \int_0^t \Gamma[f](l_u)\mathrm{d}u.
\end{aligned}$$

This concludes the argument. □

We now define the two convergences we will use for random variables: first the convergence in probability,

Definition 2.2
Let X^N and Y be random variables with value in $\mathcal{X}$. If for any $\epsilon > 0$,

$$\mathbb{P}\left(|X^N - Y| \geqslant \epsilon\right) \to 0 \text{ as } N \to +\infty,$$

we say that X^N converges to Y in probability and denote it $X^N \xrightarrow{\mathbb{P}} Y$.

Then convergence in distribution

Definition 2.3
We will say that a sequence of random variables X^N on $\mathcal{X}$ converges in distribution to μ a probability measure on $\mathcal{X}$ and write it $X^N \Rightarrow \mu$ if for any $f : \mathcal{X} \to \mathbb{R}$ measurable, continuous and bounded,

$$\mathbb{E}\left[f(X^N)\right] \to \int_{\mathcal{X}} f(x)\mu(\mathrm{d}x).$$

We will also write $X^N \Rightarrow Y$ if X^N converges in distribution to the distribution of Y.

2.1.2 Model description

We consider a finite set V of n nodes. Each node $v \in V$ represents an $M/M/1$ queue with the FIFO service discipline and vacation, 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 $(\sigma_v(t))_{v\in V} \in \{0,1\}^V$ the activity process: the server at v is active and processing pending requests at unit rate whenever $\sigma_v(t) = 1$, and $\sigma_v(t) = 0$ otherwise. Put differently, $\sigma_v(t)$ is the instantaneous service rate of node v at time t. Next, we define $\lambda \coloneqq (\lambda_v, v \in V)$, $Q(t) \coloneqq (Q_v(t), v \in V)$ and $\sigma(t) \coloneqq (\sigma_v(t), v \in V)$.

Those nodes are placed on a simple undirected graph $G = (V, E)$. An edge between two nodes indicates that they cannot be active at the same time. This is used to model interference constraints in a wireless network. The $\sim$ sign will be used to signify the existence of an edge ($v \sim w \Leftrightarrow \{v, w\} \in E$). We will have two equivalent representations for the schedule: we will sometimes see it as a subset of nodes $\sigma \subset V$ when speaking of "adding" or "removing" a node in the schedule and otherwise as a vector of $\{0,1\}^V$ by identifying nodes currently active in the schedule with non-zero entries of the service rate vector. The admissible service rates can be seen as stable sets of the interference graph: a stable set of G is given by $\sigma \in \{0,1\}^V$ such that $v \sim w \Rightarrow \sigma_v + \sigma_w \leqslant 1$. The admissible service decisions are elements of

$$S(G) \coloneqq \{\sigma \in \{0,1\}^V \mid v \sim w \Rightarrow \sigma_v + \sigma_w \leqslant 1\}.$$

Given the current schedule σ, the queue-length process Q evolves as n independent

$M/M/1$ queues with service rates σ, input rates λ and FIFO service discipline. On the other hand, σ also evolves: given the queue-length process Q, 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 neighboring 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* refers to the time scales in the stochastic averaging principle from Section 1.3.1. Thus, L acts on functions $f : \mathbb{N}^V \times S(G) \to \mathbb{R}$ as

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

with

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

and

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

with $g : \mathbb{N}^V \to \mathbb{R}$ and $h : S(G) \to \mathbb{R}$ arbitrary functions and $e^v \in \{0,1\}^V$ with 0's everywhere except at the vth coordinate equal to 1. One can check that for any $q \in \mathbb{N}^V$, L_{f}^{q} admits a unique reversible distribution denoted π^q. For reasons explained in the introduction, 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 $a > 0$ the parameter of this algorithm. In this case, π^q is given by

$$\pi^q(\sigma) = \frac{\prod_{v \in V} (1 + q_v)^{a \sigma_v}}{\sum_{\rho \in S(G)} \prod_{v \in V} (1 + q_v)^{a \rho_v}}, \sigma \in S(G).$$

With the generator L, the *carré du champ* is given by

$$\begin{aligned}\Gamma[f](q,\sigma) = &\sum_{v\in V}\lambda_v\left(f\left(q+e^v,\sigma\right)-f(q,\sigma)\right)^2 \\ &+\sum_{v\in V}\sigma_v\mathbb{1}_{q_v>0}\left(f\left(q-e^v,\sigma\right)-f(q,\sigma)\right)^2 \\ &+\sum_{v\in V}\left(f(q,\sigma-e^v)-f(q,\sigma)\right)^2\frac{\sigma_v}{1+(q_v+1)^a} \\ &+\sum_{v\in V}\left(f(q,\sigma+e^v)-f(q,\sigma)\right)^2\frac{\prod_{w\sim v}(1-\sigma_w)(1-\sigma_v)}{1+(q_v+1)^{-a}}.\end{aligned} \tag{2.3}$$

An important tool to understand the behavior of the network is the so-called "homogenized process" where the service rates are replaced by their steady state distribution. Let $g:\mathbb{N}^V\to\mathbb{R}$. The generator is given by

$$L_{\mathrm{h}}[g](q)=\sum_{v\in V}\lambda_v\left(g(q+e^v)-g(q)\right)+\sum_{v\in V}\pi^q(\sigma_v=1)\mathbb{1}_{q_v>0}\left(g(q-e^v)-g(q)\right). \tag{2.4}$$

When all queue lengths are large, only the stable sets of largest size will matter in the invariant measure. For any $\sigma\in S(G)$, $|\sigma|$ is its size (the number of active nodes). We can define the constant $\Upsilon\coloneqq\max_{\sigma\in S(G)}|\sigma|$ (we omit the dependency in G) and define

$$S^*\coloneqq\{\sigma\in S(G)\mid|\sigma|=\Upsilon\}.$$

When q_v is larger than 0 for every v, for a large parameter N, $\pi^{Nq}\approx\pi^q_\infty$ given by

$$\pi^q_\infty(\sigma)=0 \text{ if } \sigma\notin S^* \text{ and } \pi^q_\infty(\sigma)=\frac{\prod_{v\in V}q_v^{a\sigma_v}}{\sum_{\rho\in S^*}\prod_{v\in V}q_v^{a\rho_v}} \text{ if } \sigma\in S^*.$$

The instantaneous service rate will be given by $\overline{\pi}^q(v)=\pi^q_\infty(\sigma_v=1)$ for all $v\in V$. There is a partially uniform result:

Lemma 2.4
Let $C_->0$ and $C_+<+\infty$. We have

$$\forall v\in V,\quad \sup_{C_-\leqslant\min_w q_w\leqslant\max_w q_w\leqslant C_+,q\in\frac{1}{N}\mathbb{N}^V}|\pi^{Nq}(\sigma_v=1)-\overline{\pi}^q(v)|\to 0 \text{ as } N\to+\infty. \tag{2.5}$$

Proof sketch. The general idea is to decompose $\pi^{Nq}(\sigma_v=1)$ along the size of independent sets and provide a Taylor expansion in order to obtain explicit bounds depending only on C_-, C_+ and N. We have to check that the influence of stable sets of size smaller than the maximum vanishes as $N\to+\infty$. Notice that

$$\pi^{Nq}(\sigma_v=1)=\sum_{\sigma\in S^*}\sigma_v\pi^{Nq}(\sigma)+\sum_{\rho\in S(G)\setminus S^*}\rho_v\pi^{Nq}(\rho).$$

If $\rho \in S(G) \setminus S^*$, $|\rho| \leqslant \Upsilon - 1$. Then for any $\rho \in S(G) \setminus S^*$,

$$\pi^{Nq}(\rho) = \frac{\prod_{v \in V} (1 + Nq_v)^{a\rho_v}}{\sum_{\eta \in S(G)} \prod_{v \in V} (1 + Nq_v)^{a\eta_v}} \leqslant C \frac{N^{a|\rho|}}{N^{a\Upsilon}} \leqslant CN^{-a} \to 0 \text{ as } N \to +\infty.$$

Similarly, if $\sigma \in S^*$, and $\min_v q_v > 0$

$$\pi^{Nq}(\sigma) = \frac{\prod_{v \in V} (\frac{1}{N} + q_v)^{a\rho_v}}{\sum_{\eta \in S(G)} N^{|\eta| - \Upsilon} \prod_{v \in V} (\frac{1}{N} + q_v)^{a\eta_v}} \approx \frac{\prod_{v \in V} (\frac{1}{N} + q_v)^{a\rho_v}}{\sum_{\eta \in S^*} \prod_{v \in V} (\frac{1}{N} + q_v)^{a\eta_v}} \to \pi^q_\infty(\sigma) \text{ as } N \to +\infty.$$

This concludes the proof for point-wise convergence. See Appendix A for the proof of uniform convergence. □

2.1.3 Localization

On multiple occasions, we will use stopping times constructed from a trajectory to stop the process once it leaves a set convenient for analysis. In Chapters 4 and 5 we will be able to remove this localization using properties of the limiting process. We present the general definition here and will specify as necessary. For any $\epsilon \geqslant 0$, integer p and trajectory $f \in (\mathbb{R}^p_+)^{\mathbb{R}_+}$, and $U \subset \mathbb{R}^p$, let us define

$$\tau^\epsilon(f) \coloneqq \inf \{t > 0,\ \exists k \in \{1, \dots, p\} \mid\ f_k(t) \leqslant \epsilon\}, \tag{2.6}$$

and

$$\bar{\tau}^U(f) \coloneqq \inf\{t \geqslant 0,\ f(t) \notin U\}.$$

To lighten notation, we will omit dependency in the trajectory when it is Q (or a scaled version when unambiguous).

Similarly in both fluid and heavy traffic limits, we will prove convergence up to the time the process escapes a "tube" around the candidate limit and then remove localization. This will help us deal with reflection and regularity issues by bounding queues away from 0. Let $\epsilon > 0$, and $f, g : \mathbb{R}_+ \to \mathbb{R}^V_+$, let us define the stopping time

$$T^\epsilon(f, g) \coloneqq \inf\{t \geqslant 0,\ \|f(t) - g(t)\|_\infty \geqslant \epsilon\}.$$

This will be applied to the scaled queue lengths and their candidate limiting processes.

2.2 Functional analysis

For the rest of the chapter, let L^0 be the infinitesimal generator of a reversible positive recurrent Markov process with invariant probability π on $\mathcal{X}$ a finite set. We will use m^x_t to denote the distribution of X_t the Markov process of generator L^0 at time t conditioned on $X_0 = x \in \mathcal{X}$.

This section will be divided in two parts: in the first one, we will begin by

introducing general notions from functional analysis, mainly spectral gaps and log-Sobolev constants. Then we define two distances for probability measures and link them together. Finally, we give standard results for the convergence speed of Markov chains to their steady state average.

In the second part, we define the notions of Poisson equations and their solutions. This notion will be at the center of the next chapter and is of paramount importance for homogenization. One of the ideas behind the Poisson equation and its solutions is to find a functional application such that $g \mapsto \phi_g$ acts like an inverse for a given generator L^0, i.e., for a given $g \in \mathbb{L}^1(\pi)$, find ϕ such that

$$L^0(\phi)(x) = g(x) - \pi[g],\ \pi[\phi] = 0.$$

Without the $\pi[\phi_g] = 0$ condition, the previous equation could have an infinite number of solutions so this condition serves as a normalization. Poisson equations can be used for homogenization by giving the ability the rewrite the difference between a function and its steady state average. We then bound the norm of the solutions using the log-Sobolev constant.

2.2.1 General notions

We begin this section with the definition of three crucial quantities related to the generator L^0.

Definition 2.5
Let us define these notions:

- *The Dirichlet form of L^0 defined for any $f, g \in \mathbb{L}^1(\pi)$,*

$$\mathcal{E}(f,g) = -\left\langle f, L^0[g]\right\rangle_\pi.$$

- *The spectral gap of L^0, defined by*

$$\ell \coloneqq \inf_{f|\mathrm{Var}_\pi(f)\neq 0} \frac{\mathcal{E}(f,f)}{\mathrm{Var}_\pi(f)},$$

  with

$$\mathrm{Var}_\pi(f) = \sum_{x,y\in\mathcal{X}} |f(x)-f(y)|^2 \pi(x)\pi(y).$$

- *The log-Sobolev constant of L^0, defined by*

$$\alpha = \inf_{\mathcal{L}(f)\neq 0} \frac{\mathcal{E}(f,f)}{\mathcal{L}(f)},$$

  with

$$\mathcal{L}(f) = \sum_{x\in\mathcal{X}} f(x)^2 \log\left(\frac{f(x)^2}{\|f\|_2^2}\right)\pi(x).$$

Next, let us define the total variation distance between two measures:

Definition 2.6
Let μ and ν be two probability measures on $\mathcal{X}$. The total variation distance between μ and ν is given by

$$d_{\mathrm{TV}}(\mu,\nu) \coloneqq \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x)-\nu(x)|\,.$$

On multiple occasions, we will use the relative entropy between two probability measures, also called Kullback Leibler divergence. With the convention $\log(\frac{0}{0})\times 0 = 0$, we get the definition:

Definition 2.7
For any two positive measures μ and ν on $\mathcal{X}$ with μ absolutely continuous with respect to ν,

$$d_{\mathrm{KL}}(\mu,\nu) \coloneqq \sum_{x\in\mathcal{X}} \log\left(\frac{\mu(x)}{\nu(x)}\right)\mu(x).$$

Even though it is not a distance it has some interesting properties exposed in the next proposition:

Proposition 2.8
For any two positive measures μ and ν on $\mathcal{X}$ with μ absolutely continuous with respect to ν,

- $d_{\mathrm{KL}}(\mu,\nu) \geqslant 0$,
- $d_{\mathrm{KL}}(\mu,\nu) = 0 \iff \mu = \nu$,
- *Pinsker inequality:* $2d_{\mathrm{TV}}(\mu,\nu)^2 \leqslant d_{\mathrm{KL}}(\mu,\nu)$.
- $d_{\mathrm{KL}}(\mu,\nu) \leqslant \dfrac{1}{\nu_{\min}}\|\mu-\nu\|_1^2$, *with* $\nu_{\min} = \min_{x\in\mathcal{X}}\nu(x)$.

Proof. See [DSC96] for a proof of these results. The last inequality is standard in the litterature and a proof can be seen in [GSS19]. □

The speed at which a reversible ergodic Markov process converges to its invariant measure is exponential in the log-Sobolev constant.

Proposition 2.9
If L^0 is reversible, $x\in\mathcal{X}$, $t>0$, $d_{\mathrm{KL}}(m_t^x,\pi)) \leqslant \log\left(\frac{1}{\pi(x)}\right)e^{-4\alpha t}$.

Proof. We refer the interested reader to [DSC96] for proofs of this result. □

In addition to the relation between the log-Sobolev constant and the spectral gap, there are some significant links between different deterministic times related to the distribution of a reversible Markov process.

Definition 2.10
We have

- *Let T_{mix} be the mixing time of L^0:*

$$T_{\text{mix}} \coloneqq \inf\left\{t \geq 0 : \max_{x\in\mathcal{X}} d_{\text{TV}}(m_t^x, \pi) < \frac{1}{2e}\right\}.$$

- *Let T_{hit} be its random target hitting time:*

$$T_{\text{hit}} \coloneqq \max_{x\in\mathcal{X}, A\subset\mathcal{X}} \pi(A)\mathbb{E}_x(T_A),$$

with T_A the hitting time of A for X_t:

$$T_A \coloneqq \inf\{t \geq 0,\ X_t \in A\},\ A \subset \mathcal{X}.$$

We now explicit the relation between those times, the spectral gap and the log-Sobolev constant. To lighten notations, let $\pi_{\min} \coloneqq \min_{x\in\mathcal{X}} \pi(x)$.

Proposition 2.11 • *We have*

$$\frac{\ell}{2 - \log(\pi_{\min})} \leqslant \alpha \leqslant \frac{\ell}{2}.$$

- *Also,*

$$\frac{1}{\ell} - 1 \leqslant T_{\text{mix}} \leqslant \frac{1}{\ell}\log(\frac{2e}{\pi_{\min}}).$$

- *There is a constant $c_0 > 0$ independent from the Markov process such that,*

$$T_{\text{mix}} \leq c_0 T_{\text{hit}}.$$

In particular

$$\ell \geqslant \frac{1}{c_0 T_{\text{hit}} + 1} \text{ and } \alpha \geqslant \frac{1}{(2 - \log(\pi_{\min}))(c_0 T_{\text{hit}} + 1)}.$$

Proof. For the first point see [SC97], for the second and third ones see [LPW17] and [Ald81] for a proof of these results in the discrete time case (though the proof readily applies to a continuous case). Combining the two previous results we get the last point directly. □

2.2.2 Poisson equation

The reason we are interested in the log-Sobolev constant is to use it to bound the norm of solutions to the Poisson equation which we define here:

Definition 2.12
For $g \in \mathbb{L}^1(\pi)$, a Poisson equation associated with g, L^0 is the equation in $\phi : \mathcal{X} \to \mathbb{R}$

$$L^0[\phi](x) = g(x) - \pi[g],\ \pi[\phi] = 0 \tag{2.7}$$

These equations have explicit solutions:

Proposition 2.13
For any $g \in \mathbb{L}^1(\pi)$ bounded, the equation (2.7) has a unique solutionϕ_g. It is given by

$$\phi_g(x) = -\int_0^\infty (m_t^x[g] - \pi[g])\,\mathrm{d}t.$$

Proof. We can check that ϕ_g given by

$$\phi_g(x) = -\int_0^{+\infty} (m_t^x[g] - \pi[g])\,\mathrm{d}t$$

is a solution to the Poisson equation. This quantity makes sense because the distance between m_t^x and π decreases exponentially fast with t by Proposition 2.9. Since $\mathcal{X}$ is a finite state space, L^0 commutes with the integral. Thus with this expression for ϕ_g, we get

$$L^0[\phi_g](x) = -\int_0^{+\infty} L^0(m_t^{\cdot}[g])(x)\mathrm{d}t.$$

Using Kolmogorov forward equation, we get that $L^0(m_t^{\cdot}[g])(x) = \partial_t m_t^x[g]$. Replacing in the integral,

$$L^0[\phi_g](x) = -\int_0^{+\infty} \partial_t m_t^x[g]\mathrm{d}t = -[m_t^x[g]]_{t=0}^{+\infty}.$$

At time 0, m_t^x is a Dirac measure on $\{x\}$. When $t \to +\infty$, m_t^x converges in distribution to the invariant measure π, see for instance [LPW17], Theorem 4.9 and thus $m_t^x[g] \to \pi[g]$ by the dominated convergence theorem because g is bounded. We get

$$L^0[\phi_g](x) = g(x) - \pi[g].$$

This also proves the existence of a solution to the Poisson equation. We have $\pi[\phi_g] = 0$ because for any $t \geqslant 0$, $\pi[m_t^{\cdot}[g]] = \pi[g]$ by the invariance property of π.

For uniqueness, consider ϕ and ϕ' solutions to the same Poisson equation. Then

$$\forall x \in \mathcal{X},\ L^0[\phi - \phi'](x) = 0\ .$$

This implies that $(\phi - \phi')(l_t)$ is a martingale by Proposition 2.1. Let $x \in \mathcal{X}$, and recall the definition of T_x:

$$T_x = \inf\{t \geqslant 0,\ X_t = x\}.$$

Since $\mathcal{X}$ is finite and L^0 irreducible, we know that T_x is almost surely finite, which implies that $(\phi - \phi')(l_{t\wedge T_x})$ is also a martingale by the optional stopping theorem. For any $y \in \mathcal{X}$,

$$(\phi - \phi')(y) = \mathbb{E}_y\left[(\phi - \phi')(l_0)\right] = \mathbb{E}_y\left[(\phi - \phi')(l_{t\wedge T_x})\right] \overset{(a)}{=} \mathbb{E}_y\left[(\phi - \phi')(l_{T_x})\right] = (\phi - \phi')(x).$$

Equality (a) is due to the optional stopping theorem. Thus $\phi - \phi'$ is a constant function. □

We now explain how the spectral gap and Poisson equations are linked: recall the notation $\pi_{\min} \coloneqq \min_{x\in\mathcal{X}} \pi(x)$.

Proposition 2.14
For any $g : \mathcal{X} \to \mathbb{R}$, we have

$$\|\phi_g\|_\infty \leqslant \frac{\sqrt{2}}{\ell} \|g\|_\infty (-\log(\pi_{\min}))^{1/2}(2 - \log(\pi_{\min})).$$

If $\pi_{\min} \leqslant e^{-2}$,

$$\|\phi_g\|_\infty \leqslant \frac{2\sqrt{2}(-\log(\pi_{\min}))^{3/2}}{\ell} \|g\|_\infty .$$

Proof. The explicit expression from Proposition 2.13 gives

$$\|\phi_g\|_\infty \leqslant 2 \|g\|_\infty \int_0^{+\infty} d_{\mathrm{TV}}(m_t^x, \pi)\mathrm{d}t$$

and then

$$\|\phi_g\|_\infty \leqslant 2 \|g\|_\infty \int_0^{+\infty} \left(\frac{1}{2} m_t^x \left[\varphi_t^x\right]\right)^{1/2} \mathrm{d}t,$$

where $\varphi_t^x = \log(m_t^x/\pi)$, by Pinsker's inequality from Proposition 2.8, we get by Proposition 2.9 that

$$d_{\mathrm{KL}}(m_t^x, \pi) = m_t^x \left[\varphi_t^x\right] \leqslant -\log(\pi_{\min}) e^{-4\alpha t},$$

while Proposition 2.11 gives

$$\alpha \geq \frac{\ell}{2 - \log(\pi_{\min})}.$$

This gives

$$\|\phi_g\|_\infty \leqslant \|g\|_\infty \sqrt{2}(-\log(\pi_{\min}))^{1/2} \int_0^{+\infty} \exp\left(-\frac{t\ell}{2 - \log(\pi_{\min})}\right) \mathrm{d}t,$$

and thus,

$$\|\phi_g\|_\infty \leqslant \sqrt{2}\frac{(-\log(\pi_{\min}))^{1/2}(2 - \log(\pi_{\min}))}{\ell} \|g\|_\infty \leqslant \frac{\sqrt{2}\|g\|_\infty}{\ell}(-\log(\pi_{\min}))^{1/2}(2-\log(\pi_{\min})).$$

This concludes the proof. □

2.3 Glauber dynamics for QB-CSMA

2.3.1 Historical results

We are mainly interested in the spectral gap of the Glauber dynamics. It gives an idea of the time it takes for the throughput at each queue to reach their steady state equilibrium. The literature on the spectral gap of the Glauber dynamics is rich. See [LY93] for a general bound on the spectral gap of the Glauber dynamics on a cubic

graph. The Glauber dynamic can for instance be used to sample independent sets according to a Gibbs distribution. See [Vig01] where the authors use a coupling argument to bound the mixing time of the Glauber dynamics to estimate the generation speed of independent sets. Their focus was to see how the mixing time scales with the number of nodes. They consider activation rates equal to $\frac{\nu}{1+\nu}$ across the graph and deactivation rates $\frac{1}{1+\nu}$. With Υ the maximum size of an independent set in G and n the number of nodes, if $\nu < \frac{2}{\Upsilon-2}$, the mixing time scales like $O(n\log(n))$ for large n. This result was first proved in triangle free graph in [LV99]. We refer the reader to [RT98] to see how the mixing time, the spectral gap, the log-Sobolev constants and other quantities are used in order to compare the performance of different dynamics related to the Glauber dynamic. This type of bounds has been used for the study of the classical CSMA to prove that the dynamic of the schedule reaches an equilibrium fast enough. If the fugacities are small enough, the mixing time of the dynamic scales polynomially in the number of nodes.

2.3.2 Spectral gap of Glauber dynamics for QB-CSMA

When the goal is to study QB-CSMA, one possible approach is to see how the spectral gap of L_{f}^q (the generator defined in (2.2)) scales with $\|q\|_\infty$. The value of ℓ^q the spectral gap of $\|q\|_\infty$ may not have this form and for instance use information about queues with smaller size but having this bound allows us to bound the spectral gap as soon as the maximum of queue lengths is smaller than a constant. This allows for a practical way of slicing the state space: the set $\left\{q \in \mathbb{N}^V,\ \max_v q_v \leqslant K\right\}$ is finite for any $K < +\infty$. When the activation/deactivation ratio is of the form $e^{f(q_v)}$, most of the bounds on the spectral gap have the form

$$\ell^q \geqslant \exp(-\beta f(\|q\|_\infty)). \tag{2.8}$$

As an example, in [SS12], using Cheeger's inequality the authors prove a bound which indicates that in any interference graph,

$$\ell^q \geqslant \exp\left(-2(\Upsilon+1)f(\|q\|_\infty)\right). \tag{2.9}$$

In an unpublished note, Laurent Miclo proved that we can sometimes improve this bound. Fix a numberings of the nodes $V = \{v_1, \ldots, v_n\}$, we define the sequential boundary :

$$\forall k \in \mathbb{N}, 1 \leqslant k \leqslant n,\ V_k \coloneqq \{v \in V \mid \exists i \in \{1,\ldots,k\} \text{ s.t. } (v,v_i) \in E\} \setminus \{v_1,\ldots,v_k\}\}. \tag{2.10}$$

Let $v(G)$ be the maximum degree of a node in G and $\iota(G)$ the maximum over k of the size of V_k. The bound obtained becomes

$$\ell^q \geqslant C\exp(-\left(2+\min(n-1,(1+v(G))\iota(G))\right)f(\|q\|_\infty)).$$

We can optimize the bound by minimizing $\iota(G)$ over numbering of V.

This method relies on a geometric path argument. See for instance proposition 1 of [SC97]. The general idea is to first assign a path to all oriented pairs $(\eta,\eta') \in S(G)^2$. For any two schedules $(\eta_1,\eta_2) \in S(G)^2$ that are subsequent on at least a path, define a weight that acts like a congestion measure: this weight is computed by summing the weight of each path over all paths containing the η_1 to η_2 transition,

normalized by $\pi^q(\eta_1)L^q_{\mathrm{f}}(\eta_1,\eta_2)$ the steady state rate of transitions from η_1 to η_2. The weight of a path is given by its length multiplied by the steady state average of both extremities. In order to have a bound on the spectral gap, take the maximum over all couples of schedules and minimize over the chosen set of paths to optimize the bound.

From now on, we consider the case of polynomial activation functions, i.e.

$$f(q) = a\log(q+1) \text{ and } \Psi_-(q) = \frac{1}{1+(q+1)^a}.$$

Informally, let

$$\beta_0 = \inf\{b > 0 \mid \ell^q \geqslant C\,\|q+1\|_\infty^{-ab}\ \forall q \in \mathbb{N}^V\},$$

with C a numerical constant depending only on the interference graph. Let

$$\beta_0 \coloneqq \sup_{f \text{ increasing}} \limsup_{q\to+\infty} -\frac{\log(\ell^q)}{f(\|q\|_\infty)}.$$

Because of the previous bounds, for any interference graph, $\beta_0 < +\infty$. Because of (2.9), the supremum over increasing function is finite. In fact, by (2.9), $\beta_0 \leqslant 2(\Upsilon+1)$. By definition of β_0, for any $\epsilon > 0$, and activation function there is $K < +\infty$ such that for any $q \in \mathbb{N}^V$ with $\|q\|_\infty > K$,

$$-\frac{\log(\ell^q)}{f(\|q\|_\infty)} \leqslant \beta + \epsilon,$$

which can be rewritten as

$$\ell^q \geqslant \exp(-(\beta_0+\epsilon)f(\|q\|_\infty)).$$

The constant β_0 is the smallest number with this property. When considering bounds on the spectral gap, we may refer to bounds on β_0 for inequalities such as (2.8), i.e. for instance $\beta_0 \leqslant b$ means that there exists C independent from b such that $\ell^q \geqslant C\,\|q+1\|_\infty^{-ab}$ when $\|q\|_\infty$. Heuristically, a lower β_0 implies a wider range of a for which we can prove homogenization.

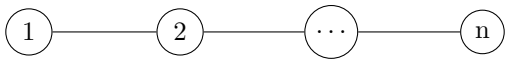

Figure 2.1: line graph with n nodes

Neither of the general bounds discussed in the previous discussion is tight and depending on the graph topology one or the other might yield a better bound for large n. In the case of a complete interference graph, the maximum size on an independent set is 1, the maximum degree is $n-1$ and for any numbering of the nodes, there are $n-1$ nodes in V_1 from (2.10). The bound of [SS12] yields $\beta_0 \leqslant 4$ for any size when the unpublished one gives $\beta_0 \leqslant n+1$ with n the number of nodes. The bound of [SS12] is better as soon as $n \geqslant 3$. If the interference graph is a line of size n (see Fig: 2.2), the largest independent set will have approximately $\frac{n}{2}$ nodes, the maximum degree is always two and if we number the nodes successively on the

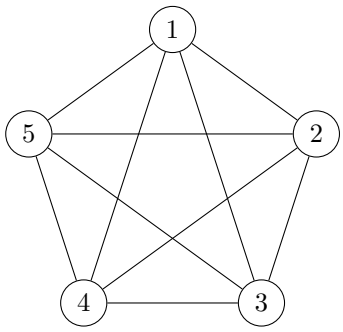

Figure 2.2: Complete interference graph with 5 nodes

line, the size of the sequential boundary from (2.10) will always be one. Thus the first bound gives essentially $n+2$ while the second one gives 5 for any n. The bound from [SS12] is worst as soon as $n \geqslant 3$. The important thing is the existence of this $\beta_0 < +\infty$ depending only on the graph G.

We prove at the end of the section the following lemma.

Lemma 2.15
If G is a complete interference graph, then for any $q \in \mathbb{N}^V$ we have

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

We can also provide a lower bound for the mixing time independently from spectral gap considerations:

Lemma 2.16
For any interference graph, we have

$$T_{\text{mix}}^q \geqslant C \|q+1\|_\infty^a .$$

This implies $\beta \geqslant 1$.

Proof. The mixing time cannot scale slower than $\|q+1\|_\infty^a$. Indeed, if the schedule starts with a queue with a large q_v active, the exit time of the first state for the schedule will be an exponential variable with mean $1+(q_v+1)^a$. Using concentration inequalities for exponential variables, we can prove that the probability of the exit time for the initial state being smaller than $C(q_v+1)^{a\beta}$ converges to 0 as $q_v \to +\infty$ if $\beta < 1$. Thus the mixing time for this dynamic cannot be smaller than $C \|q+1\|_\infty^{a\beta}$ for any $\beta < 1$.

Thus

$$\log(\frac{2e}{\pi^q(\mathbf{0})})\frac{1}{\ell^q} \geqslant T_{\text{mix}}^q \geqslant C \|q+1\|_\infty^a .$$

This is enough for the statement about β because

$$\pi^q(\mathbf{0}) = \frac{1}{1+\sum_{v\in V}(q_v+1)^a} \geqslant \frac{1}{1+n\|q+1\|_\infty^a}.$$

If

$$\ell^q \leqslant C \left\|q+1\right\|_\infty^{-a} \log(2e(1+n\left\|q+1\right\|_\infty^a)).$$

It is not possible to have $\ell^q \geqslant C \left\|q+1\right\|_\infty^{-a\beta}$ and $\beta < 1$ because

$$C \left\|q+1\right\|_\infty^{-a\beta} > C \left\|q+1\right\|_\infty^{-a} \log(2e(1+n\left\|q+1\right\|_\infty^a))$$

for large $\left\|q+1\right\|_\infty$. □

In K-partite complete interference graph, we expect this conjecture to hold:

Conjecture 2.17
For G a complete K-partite graph, the spectral gap of L_{f}^q can be bounded by:

$$\ell^q \geqslant C \left\|q+1\right\|_\infty^{-a\Upsilon}.$$

Proof stub: Our conjecture is true in a complete interference graph but it should also hold for complete partite graphs: to switch between maximal schedules, all the queues in a configuration must deactivate before the next activation. The rate at which all the queues in a given schedule deactivate before an activation is at worst $C\exp\left(-\Upsilon f(\max_v q_v)\right)$. Using again concentration inequalities for exponential variables, we get that from any state, the hitting time of $\mathbf{0}$ should be at most $C\exp\left(\Upsilon f(\max_v q_v)\right)$. The hitting time of $\mathbf{0}$ from any starting schedule σ should also be of the same order of magnitude. Using the strong Markov property, it can be decomposed in failure cycles where it does not visit σ before going back to $\mathbf{0}$ and a success to visit σ. We already established that the time it takes to reach back $\mathbf{0}$ is at most $C\exp\left(\Upsilon f(\max_v q_v)\right)$. Moreover, the number of failures before a success is by construction a geometric variable. Because the activation/deactivation ratio is always greater than 1, the time it takes for a trajectory to go from $\mathbf{0}$ to σ should be smaller than the time it takes to go the other way around. For this reason, we expect the random target hitting time from last section to be $C\exp(\Upsilon f(\max_v q_v))$. □

We will use the random target hitting time, Proposition 2.11 and the probabilistic construction of Markov chains for our bound in the complete interference graph case.

Proof of Lemma 2.15. Because of Proposition 2.11 in order to prove the desired bound, we only need to prove that $T_{\mathrm{hit}}^q \leqslant C\|q+1\|_\infty^a$. In this proof we use the notation $\mathbb{E}_\sigma^q$ to denote the mean of the process generated by L_{f}^q given that the starting state is $\sigma \in \mathcal{S}(G)$. Recall the definition of T_{hit}:

$$T_{\mathrm{hit}} = \max_{x\in\mathcal{X}, A\subset\mathcal{X}} \pi(A)\mathbb{E}_x(T_A).$$

Necessarily for any $A \subset S(G)$, $T_A \leqslant \sum_{\sigma\in A} T_\sigma$, because by definition of T_A, there exists $\sigma^0 \in A$ such that $T_A = T_{\sigma^0}$, and thus

$$T_{\mathrm{hit}}^q \leqslant \max_{\sigma^0\in S(G)} \sum_{\sigma\in S(G)} \mathbb{E}_{\sigma^0}^q(T_\sigma).$$

Since a schedule in the case of a complete interference graph is either an active node

or the empty schedule, this actually reduces to proving that

$$\mathbb{E}_v^q(T_{\mathbf{0}}) \leq C\|q+1\|_\infty^a \quad \text{and} \quad \mathbb{E}_{\mathbf{0}}^q(T_v) \leq C\|q+1\|_\infty^a, \tag{2.11}$$

for any $v \in V$ by identifying the schedule $\{v\}$ with the node v. Indeed, for $v^0 \neq v \in V$ the process η needs to pass through $\mathbf{0}$ to go from v^0 to v and so the strong Markov property gives

$$\mathbb{E}_{v^0}^q(T_v) = \mathbb{E}_{v^0}^q[T_{\mathbf{0}}] + \mathbb{E}_{\mathbf{0}}^q[T_v].$$

So let us prove (2.11). The bound on $\mathbb{E}_v^q(T_{\mathbf{0}})$ is obvious since by definition $T_{\mathbf{0}}$ under $\mathbb{E}_v^q$ is an exponential random variable with parameter $\Psi_-(q_v)$ so that

$$\mathbb{E}_v^q[T_{\mathbf{0}}] = \frac{1}{\Psi_-(q_v)} = 1 + (q_v+1)^a \leq C\|q+1\|_\infty^a.$$

Let us now prove that $\mathbb{E}_{\mathbf{0}}^q[T_v] \leq C\|q+1\|_\infty^a$. Under $\mathbb{P}_{\mathbf{0}}^q$, decompose the trajectory into cycles away from $\mathbf{0}$: in the k-th cycle, the schedule stays in $\mathbf{0}$ for a duration X_k, then moves to some $i \in V$ where it stays for a duration Y_k and then comes back to $\mathbf{0}$. Y_k depends on the chosen node but it can be bounded above by $\widetilde{Y}_k$ an exponential variable with parameter $\frac{1}{1+(\max_v(q_v)+1)^a}$ independently from everything. Similarly, X_k can be bounded above by $\widetilde{X}_k$ the minimum of n exponential with parameter $\frac{1}{2}$. If $K \in \mathbb{N}$ denotes the number of cycle before the schedule visits v, we can thus write

$$T_v = \sum_{k=1}^{K}(\widetilde{X}_k + \widetilde{Y}_k) + \widetilde{X}_{K+1}.$$

It is possible to bound $K+1$ with a geometric variable with a parameter depending only on the number of nodes: to do that use the probabilistic construction of Markov process with independent exponential variables for each jump. Indeed, for any $q \in E$ the intensity of activation of node v is at least $\frac{1}{2}$. Similarly, the intensity of activation of any node is at most 1. Hence the probability that a given queue activates after any given idle period is greater than $\frac{1/2}{1/2+(n-1)} = \frac{1}{2n-1}$. Since the trials are independent, $K+1$ can be coupled with a geometric variable with parameter $\frac{1}{2n-1}$ which is always larger. In particular,

$$T_v \leq \sum_{k=1}^{G}(\widetilde{X}_k + \widetilde{Y}_k)$$

with G a geometric random variable with parameter $\frac{1}{2n-1}$ independent from the $\widetilde{X}_k$ and $\widetilde{Y}_k$'s. it follows that

$$\mathbb{E}_{\mathbf{0}}^q[T_v] \leqslant \mathbb{E}[G]\left(\mathbb{E}_{\mathbf{0}}^q\left[\widetilde{X}_1\right] + \mathbb{E}_{\mathbf{0}}^q\left[\widetilde{Y}_1\right]\right),$$

with

$$\mathbb{E}_{\mathbf{0}}^q\left[\widetilde{X}_1\right] \leqslant \frac{2}{n},$$

and

$$\mathbb{E}_{\mathbf{0}}^q\left[\widetilde{Y}_1\right] \leqslant C\|q+1\|_\infty^a.$$

Gathering the previous bounds yields the desired result. □

2.3.3 Influence of threshold function

Recall the activation rates from [RSS09], [GS10] and [SS12]: for any $q \in \mathbb{N}^V$,

$$\Psi_+^v(q) = \frac{e^{W_v(q)}}{1+e^{W_v(q)}}, \quad \text{and } \Psi_-^v(q) = 1 - \Psi_+^v(q),$$

with

$$W_v(q) = \max(f(q_v), h(f(\max_w q_w))). \tag{2.12}$$

We will explain in this section the reason why such a function is used and the drawbacks for this function h. When some queues are too small, a small variation in q can make big changes in the invariant measure of the dynamic with fixed q. Even though the schedule reaches its equilibrium faster than the time it takes for $\|q\|_\infty$ to change significantly, if the invariant measure of the dynamic with fixed q can evolve even faster, the equilibrium reached by the schedule will not be the one corresponding to the current state of the queue lengths. The proof of [GS10] can be tweaked to prove a time scale separation for polynomial rates but this requires h to be linear. If $f(q) = a\log(q)$, Υ the maximum size of an independent set in G and $h(x) = \delta x$, there is a time scale separation if $a\beta < \delta$, with β satisfying (2.8). This ensures that the speed at which the invariant measure evolves is slower than the mixing time of the dynamic. The logarithmic f corresponds to polynomial activation rates. If $\|q\|_\infty$ is large, the mixing time of the dynamic with fixed q is essentially

$$\exp(\beta f(\|q\|_\infty)).$$

The "speed" at which the invariant measure of the fixed q dynamic evolves with q is given by

$$\pi^q(\sigma) - \pi^{q\pm e^i} \approx f'(f^{-1}(h(f(\|q\|_\infty)))).$$

Since the mixing time is exponential, f cannot grow faster than log. Otherwise, but cummulating increments, the time it takes for distribution of the Markov process of generator L_{f}^q to reach π^q is too long compared to the time it takes for $\pi^{Q(t)}$ to evolve significantly. With f increasing slower than a logarithm, we get the "speed" of the invariant measure smaller than

$$\exp(-\frac{\delta}{a} f(\|q\|_\infty)),$$

using the linear threshold. The $a\beta < \delta$ condition ensures that the invariant measure evolves slower than the mixing time of the fixed q dynamic.

This information about $\max_w q_w$ induces some communication overhead between nodes to be able to estimate this quantity. The goal of this threshold function is to make it so that even when queues are small, their activation/deactivation rates do not change fast. The evolution speed of the weights of queues that have a small queue lengths is slowed because of the inclusion of $h \neq 0$ in the weights. The authors of [RSS09] proposed a distributed scheme to estimate $\max_w q_w$ using only gossiping between neighboring nodes. Even when reducing the communication overhead required, the problem with this threshold function is that it degrades the approximation of Max-Weight by the invariant measure. With the weights described in (1.2) the invariant measure of L_{f}^q concentrates exponentially on

$$\arg\max_{\sigma \in S(G)} \langle W(q), \sigma \rangle$$

when q becomes large. Once of the advantage of QB-CSMA is that its invariant measure approximates decisions of Max-Weights. The next discussion shows that with weights of the form (2.12), Max-Weights can give bad service decisions.

When $h(x) = \delta x$ and δ is chosen too big, it can even prevent Max-Weight with weight W from being stable: picture a network with $n+1$ nodes arranged in a star. There is one central node numbered 0 and Υ outer nodes. If we used weights such as in (2.12) and take $\Upsilon\delta > 1$, when the queue lengths are large, queue 2 can never have the maximum weight because even if queue 2 has the longest queue length the combined weight of the Υ outer queues is $\Upsilon\delta f(\max_v q_v)$.

Lemma 2.18
The Max-Weight algorithm on a star interference graph (Figure 2.3) is transient with weight of the form

$$W_v(q) = \max(f(q_v), \delta f(\max_w q_w)),$$

for any function f and $\lambda \in \mathbb{R}_+^V$, as soon as $\delta\Upsilon > 1$.

Proof. For any $q \in \mathbb{N}^V$, $\max_{\sigma\in S(G)} \langle W(q), \sigma\rangle = \delta\Upsilon f(\|q\|_\infty)$ because $\delta\Upsilon > 1$. In order to always schedule the independent set of maximum weight, the central node can never be scheduled and thus grows to infinity even when the arrival rates are in the capacity region. □

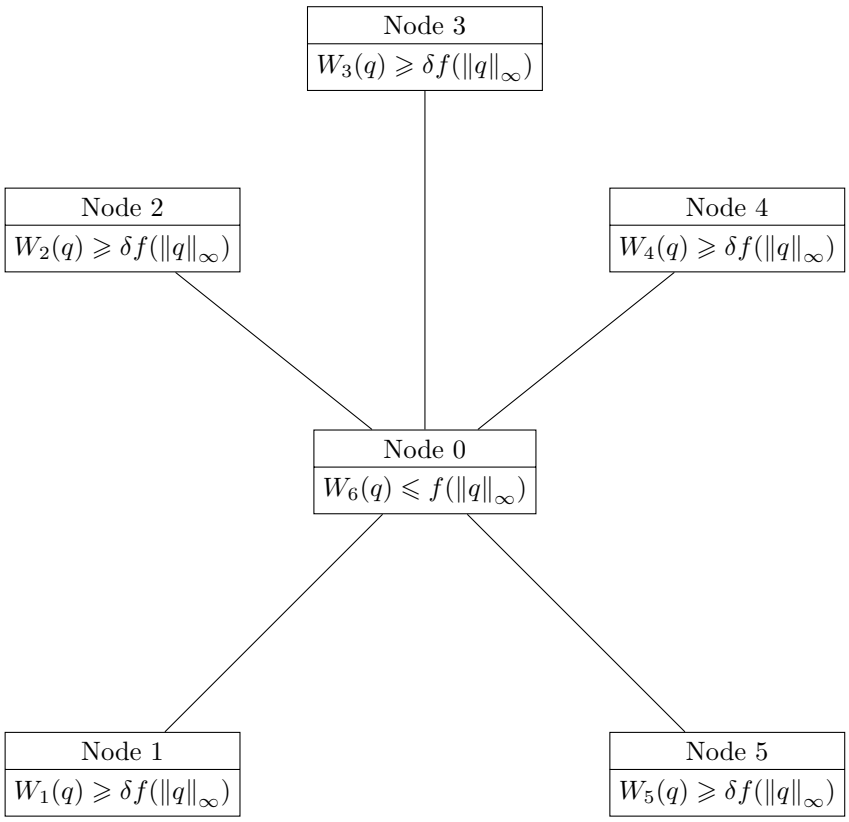

Figure 2.3: Star interference graph with 6 nodes

If $c\delta > 1$ and there is a time scale separation, the fraction of the time the central schedule is chosen is negligible as $\|q\|_\infty$ regardless of the configuration of queue lengths. Even if max-Weight is stable with weights W, the next lemma explains how the function h "degrades" the approximation of Max-Weight with weights $f(q)$ for QB-CSMA. Even when throughput optimality is preserved, approximating Max-Weight with weights W means it is not maximizing with weights $f(q)$ anymore. It is expected that delay is worse with weights W than with weights $g(q)$.

Lemma 2.19
There is a distinction between two cases:

- $h(x) = \delta x$*: For any* $\widetilde{\epsilon} > \Upsilon\delta$*, there exists* q^1 *such that for any configuration* $q \in \mathbb{N}^V$ *with* $\|q\|_\infty > q^1$*,*

$$\pi^q\left[\langle f(q), \cdot\rangle\right] \geqslant (1-\widetilde{\epsilon}) \max_{\rho\in S(G)} \langle f(q), \rho\rangle .$$

- $h(x) = o(x)$*: For any* $\widetilde{\epsilon} > 0$*, , there exists* q^1 *such that for any configuration* $q \in \mathbb{N}^V$ *with* $\|q\|_\infty > q^1$

$$\pi^q\left[\langle f(q), \cdot\rangle\right] \geqslant (1-\widetilde{\epsilon}) \max_{\rho\in S(G)} \langle f(q), \rho\rangle .$$

  We get the result by letting $\epsilon' \to 0$*.*

Proof. Fix $\widetilde{\epsilon} > 0$ and $\mathcal{X}^q := \{\eta \in S(G) : \langle f(q), \eta\rangle < (1-\widetilde{\epsilon}) \max_{\rho\in S(G)} \langle f(q), \rho\rangle\}$. We aim at proving that if $\|q\|_\infty$ is large enough, $\pi^q(\mathcal{X}^q)$ can be as small as desired.

$$\begin{aligned}
\pi^q(\mathcal{X}^q) &= \sum_{\rho\in\mathcal{X}^q} \pi^q(\rho) \\
&= \sum_{\rho\in\mathcal{X}^q} \frac{e^{\langle W(q),\rho\rangle}}{\sum\limits_{\rho\in S(G)} e^{\langle W(q),\rho\rangle}} \\
&\leqslant \sum_{\rho\in\mathcal{X}^q} \frac{\exp(\langle f(q) + h(f(\|q\|_\infty)), \rho\rangle)}{\sum\limits_{\rho\in S(G)} e^{\langle W(q),\rho\rangle}} \\
&\leqslant \sum_{\rho\in\mathcal{X}^q} \frac{\exp((1-\widetilde{\epsilon}) \max\limits_{\rho\in S(G)} \langle f(q), \rho\rangle) \exp(\Upsilon h(f(\|q\|_\infty)))}{\sum\limits_{\rho\in S(G)} e^{\langle W(q),\rho\rangle}}
\end{aligned}$$

Furthermore

$$\sum_{\rho\in S(G)} e^{\langle W(q),\rho\rangle} \geqslant \sum_{\rho\in S(G)} e^{\langle f(q),\rho\rangle} \geqslant e^{\max\limits_{\rho\in S(G)} \langle f(q),\rho\rangle}$$

and so $\pi^q(\mathcal{X}^q) \leqslant 2^n \exp(\Upsilon h(f(\|q\|_\infty)) - \widetilde{\epsilon} W_{max})$. Whenever $h(f(\|q\|_\infty)) = o(f(\|q\|_\infty))$, this bound goes to 0 when W_{max} grows to $+\infty$ for any $\widetilde{\epsilon} > 0$. If $h(f(\|q\|_\infty)) = \delta f(\|q\|_\infty)$, there is q^1 for which $\pi(\mathcal{X}) \leqslant \epsilon'$ as soon as $\|q\|_\infty > q^1$ if $\Upsilon\delta < \widetilde{\epsilon}$.

Under the conditions stated above, in both cases, for any $\epsilon' > 0$,

$$\pi^q \left(\sigma \in S(G),\, \langle f(q), \sigma \rangle < (1 - \widetilde{\epsilon}) \max_{\rho \in S(G)} \langle f(q), \rho \rangle \right) < \epsilon',$$

and so for any $\epsilon' > 0$,

$$\pi^q [\langle f(q), \cdot \rangle] \geqslant (1 - \widetilde{\epsilon})(1 - \epsilon') \max_{\rho \in S(G)} \langle f(q), \rho \rangle \quad ,$$

for $\max_v q_v$ large enough, which concludes the proof by letting $\epsilon' \to 0$. □

Chapter 3

Homogenization through Poisson equation

Contents

3.1 Main result 49
3.2 Proof of the main result 51
3.2.1 Control in terms of solutions to the Poisson equation 51
3.2.2 Control of solutions to the Poisson equation 55
3.3 Scaled process 57
3.3.1 General considerations 57
3.3.2 Fluid limits in a complete interference graph 59

3.1 Main result

In this Chapter, we present explicit bounds for the "homogenization error" for the model described in Section 2.1.2. Recall the generator of this process given in (2.1) and (2.2), and the homogenized version from (2.4). Recall that for any $q \in \mathbb{N}^V$ and $\sigma \in S(G)$,

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

and

$$L_{\mathrm{h}}[g](q) = \sum_{v\in V}\lambda_v\left(g(q+e^v)-g(q)\right) + \sum_{v\in V}\pi^q(\sigma_v=1)\mathbb{1}_{q_v>0}\left(g(q-e^v)-g(q)\right).$$

Notice that the $0-1$ service rate in L_{f}^q is replaced by a steady state average in L_{h}. The main goal in this section is to provide some bounds on the difference between

those two generators when integrated over finite time intervals:

$$\left| \int_0^T \left(L_{\mathrm{s}}^{\sigma(s)} - L_{\mathrm{h}} \right) [g](Q(s)) \mathrm{d}s \right|, \tag{3.1}$$

with $g : \mathbb{R}_+^V \to \mathbb{R}_+$ smooth enough and $T < +\infty$. This quantity allows us to compare the dynamic of the slow process with its homogenized version, the queueing process of generator L_{h}. Let $U \subset (0, +\infty)^V$ and recall that $\mathbf{0}$ is the empty schedule and

$$\bar{\tau}^U = \inf\{t > 0,\ Q(t) \notin U\}.$$

We will prove the following theorem.

Theorem 3.1
Assume U is such that $0 < \min_{q \in U} \pi^q(\boldsymbol{0})$. Assume $g : U \to \mathbb{R}$ is bounded and twice differentiable. Then, for any $T > 1$

$$\mathbb{E}\left[\sup_{t \leqslant T \wedge \bar{\tau}^U} \left| \int_0^t (L_s^{\sigma(s)} - L_h)[g](Q(s)) \mathrm{d}s \right| \right] \leqslant C B_0 \max_{v \in V} \|\partial_v g\|_{\infty, U}\, T + C\sqrt{T}\Omega_0 \left(\max_{v \in V} \|\partial_v g\|_{\infty, U} + \sqrt{T} \max_{w,v \in V} \left\|\partial^2_{w,v} g\right\|_{\infty, U} \right),$$

with

$$\Omega_0 = (-\log(\min_{q \in U} \pi^q(\boldsymbol{0})))^{3/2} \frac{1}{\min_{q \in U} \ell^q},$$

and

$$B_0 = \frac{\Omega_0^2}{\min_{v \in V, q \in U} q_v^{1+a}} + \frac{\Omega_0}{\min_{v \in V, q \in U} q_v}.$$

We now define the main tool of this chapter:

Definition 3.2
For any function $g : S(G) \to \mathbb{R}$ and any $q \in \mathbb{N}^V$ we denote by $\phi_g(q, \cdot)$ the unique solution to the Poisson equation associated to the fast generator L_{f}^q and the function g, i.e., $\phi_g(q, \cdot)$ is the unique solution to the equation with unknown ϕ

$$L_{\mathrm{f}}^q[\phi] = g - \pi^q[g],\ \pi^q[\phi] = 0. \tag{3.2}$$

In particular, $\phi_v(q, \cdot)$ is the solution to (3.2) with $g(\sigma) = \sigma_v$ for $v \in V$, which therefore satisfies for any $q \in \mathbb{N}^V$ and any $\eta \in S(G)$

$$L_{\mathrm{f}}^q \left[\phi_v(q, \cdot)\right](\eta) = \eta_v - \pi^q(\sigma_v = 1) \text{ and } \pi^q[\phi_v(q, \cdot)] = 0. \tag{3.3}$$

The proof of Theorem 3.1 has two steps: first, provide a bound in terms of solutions to the Poisson equation (3.2) and then controlling these solutions. These two steps are performed in Sections 3.2.1 and 3.2.2 respectively. To see how the Poisson equation comes into play, let us proceed with the following preliminary computation. Using (2.1) and (2.4), we can rewrite the difference in generator as

$$(L_{\mathrm{s}}^{\sigma(s)} - L_{\mathrm{h}})[g](q) = \sum_{v \in V} (\sigma_v - \pi^q(\sigma_v = 1)) \mathbb{1}_{q_v > 0} (g(q - e^v) - g(q)).$$

We get from (3.3)

$$\sigma_v - \pi^q(\sigma_v = 1) = L_{\mathrm{f}}^q[\phi_v(q, \cdot)](\sigma).$$

Fix a function $g \in D(L_{\mathrm{h}})$ and define $f_v(q) \coloneqq g(q - e^v) - g(q)$. Since f_v does not depend on σ, this makes it possible to rewrite

$$(\sigma_v - \pi^q(\sigma_v = 1))\, f_v\,(q) = L_{\mathrm{f}}^q\,[F_v(q, \cdot)]\,(\sigma)$$

with $F_v(q, \sigma) = \phi_v(q, \sigma) f_v(q)$. Integrating over a trajectory of (Q, σ) and stopping before $\bar{\tau}^U$ means that $\mathbb{1}_{Q_v(s)>0} = 1$ for all $v \in V$ and $s \leqslant \bar{\tau}^U$ because $U \subset (0, +\infty)^V$. We omit the indicator in the next computation. In addition, making use of Proposition 2.1, by integrating over a trajectory, we finally rewrite this as

$$\begin{aligned}\int_0^t \left(L_{\mathrm{s}}^{\sigma(s)} - L_{\mathrm{h}}\right)[g](Q(s))\mathrm{d}s &= \sum_{v\in V} \int_0^t \left(\sigma_v(s) - \pi^{Q(s)}(\sigma_v = 1)\right) f_v\,(Q(s))\,\mathrm{d}s \quad (3.4)\\ &= \sum_{v\in V} \int_0^t L_{\mathrm{f}}^{Q(s)}\,[F_v(Q(s), \cdot)]\,(\sigma(s))\mathrm{d}s.\end{aligned}$$

In order to prove Theorem 3.1, we will individually bound all

$$\int_0^t \left(\sigma_v(s) - \pi^{Q(s)}(\sigma_v = 1)\right) f\,(Q(s))\,\mathrm{d}s = \int_0^t L^{Q(s)}[\phi_v(Q(s), \cdot) f(Q(s))](\sigma(s))\mathrm{d}s$$

for a generic function f in Lemma 3.3 using the same decomposition and then sum over v. The Poisson equation gives us an alternative representation of $\sigma_v - \pi^q(\sigma_v = 1)$. Using the fact that for any $h \in \mathcal{D}(L)$,

$$L[h](q, \sigma) = L_{\mathrm{f}}^q[h(q, \cdot)](\sigma) + L_{\mathrm{s}}^\sigma[h(\cdot, \sigma)](q),$$

and Proposition 2.1, we get

$$\begin{aligned}\int_0^t L_{\mathrm{f}}^{Q(s)}\,[F_v(Q(s), \cdot)]\,(\sigma(s))\mathrm{d}s &= [F_v(Q(t), \sigma(t)) - F_v(Q(0), \sigma(0))]\\ &\quad - M_{F_v}(t) - \int_0^t L_{\mathrm{s}}^{\sigma(s)}\,[F_v(\cdot, \sigma(s))]\,(Q(s))\mathrm{d}s. \quad (3.5)\end{aligned}$$

It is with this expression that we will prove Lemma 3.3.

3.2 Proof of the main result

3.2.1 Control in terms of solutions to the Poisson equation

This section provides a first step toward the proof of Theorem 3.1. We will first derive a quantity useful to bound (3.4) in terms of the following constants:

$$\Omega \coloneqq \sup_{q\in U,\, \|g\|_\infty \leqslant 1} \|\phi_g(q, \cdot)\|_\infty\,, \text{ and } B \coloneqq \sup_{q\in U,\, \|g\|_\infty \leqslant 1} \max_{v\in V, \sigma\in S(G)} |\phi_g(q \pm e^v, \sigma) - \phi_g(q, \sigma)|\,.$$

During this section, we consider $v \in V$ fixed. Recall that $\bar{\tau}^U$ is the exit time of U for the process of queue lengths and we use C as a numerical constant that may

depend on the interference graph, the arrival rates and whose value is irrelevant to the proof. The goal will be to prove this lemma:

Lemma 3.3
For any $v \in V$, finite time horizon $T > 0$ and any $f : U \to \mathbb{R}_+$ differentiable, we have

$$\mathbb{E}\left[\sup_{0\leqslant t\leqslant T\wedge\bar{\tau}^U}\left|\int_0^t\left(\sigma_v(s)-\pi^{Q(s)}(\sigma_v=1)\right)f\left(Q(s)\right)\mathrm{d}s\right|\right]$$
$$\leqslant C\Omega\left\|f\right\|_{\infty,U}+C\sqrt{T}\left[\left\|f\right\|_{\infty,U}\left(\Omega+B(1+\sqrt{T})\right)+\max_{v\in V}\left\|\partial_v g\right\|_{\infty,U}\Omega(1+\sqrt{T})\right].$$

From (3.5), we get

$$\int_0^t\left(\sigma_v(s)-\pi^{Q(s)}(\sigma_v=1)\right)f\left(Q(s)\right)\mathrm{d}s=\left[F(Q(t),\sigma^N(t))-F(Q(0),\sigma(0))\right]-M_F(t)$$
$$-\int_0^t L_{\mathrm{s}}^{\sigma(s)}\left[F(Q(s),\,\cdot\,)\right](\sigma(s))\mathrm{d}s, \quad (3.6)$$

with $F(q,\sigma)=\phi_v(q,\sigma)f(q)$. We will deal with the martingale term in Lemma 3.4. For the F terms, we will use

$$|F(q,\sigma)|\leqslant\left\|f\right\|_{\infty,U}\Omega. \quad (3.7)$$

Recall that

$$L_{\mathrm{s}}^{\sigma}[f](q)=\sum_{w\in V}\left[\lambda_w\left(f(q+e^w)-f(q)\right)+\sigma_w\mathbb{1}_{q_w>0}\left(f(q-e^w)-f(q)\right)\right],$$

so for the $L_{\mathrm{s}}^{\sigma}[F]$ term,

$$|F(q\pm e^w,\sigma)-F(q,\sigma)|\leqslant\max_{w\in V}\left\|\partial_w f\right\|_{\infty,U}\Omega+\left\|f\right\|_{\infty,U}B, \quad (3.8)$$

because $f(q\pm e^w)-f(q)=\int_0^1\partial_w f(q\pm ue^w)\mathrm{d}u$.

We now explain how we control the martingale term:

Lemma 3.4
Recall $F(q,\sigma)=\phi_v(q,\sigma)f(q)$. We have

$$\mathbb{E}\left[\sup_{0\leqslant t\leqslant T\wedge\bar{\tau}^U}|M_F(t)|\right]\leqslant C\Omega\sqrt{T}\left(\left\|f\right\|_{\infty,U}+\max_{j\in V}\left\|\partial_j f\right\|_{\infty,U}\right)+C\sqrt{T}\left\|f\right\|_{\infty,U}B.$$

Proof. Using Doob's inequality for the first inequality, Itô's isometry for the first equality and Proposition 2.1 for the second equality, we have:

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

According to (2.3), we have

$$\begin{aligned}\Gamma[F](q,\sigma) = &\sum_{w\in V}\lambda_w\left(F(q+e^w,\sigma)-F(q,\sigma)\right)^2\\ &+\sum_{w\in V}\sigma_w\mathbb{1}q_w>0\left(F(q-e^w,\sigma)-F(q,\sigma)\right)^2\\ &+\sum_{w\in V}\left(F(q,\sigma-e^w)-F(q,\sigma)\right)^2\frac{\sigma_w}{1+(q_w+1)^a} \qquad (3.9)\\ &+\sum_{w\in V}\left(F(q,\sigma+e^w)-F(q,\sigma)\right)^2\frac{\prod_{w\sim v}(1-\sigma_w)(1-\sigma_v)}{1+(q_w+1)^{-a}}. \qquad (3.10)\end{aligned}$$

We integrate this quantity over the trajectory (Q,σ) for $t\leqslant T\wedge\bar\tau^U$: along this trajectory we bound the terms $\sigma_v(s)$, $1/(1+(Q_w(s)+1)^a)$ and $1/(1+(Q_w(s)+1)^{-a})$ by one. Recall that $F(q,\sigma)=f(q)\phi_v(q,\sigma)$: thus for $q\in U$, using (3.7) and (3.8), we obtain

$$\mathbb{E}\left[\int_0^{T\wedge\bar\tau^U}\Gamma[F](Q(s),\sigma(s))\mathrm{d}s\right]\leqslant 2nT\left(\max_{v\in V}\|\partial_v f\|_{\infty,U}\,\Omega+\|f\|_{\infty,U}B\right)^2+2nT\|f\|_{\infty,U}^2\Omega^2.$$

Using $\sqrt{x^2+y^2}\leqslant|x|+|y|$ and Cauchy-Schwartz inequality, we obtain

$$\mathbb{E}\left[\sup_{0\leqslant t\leqslant T\wedge\bar\tau^U}M_F(t)\right]\leqslant\Omega\sqrt{2nT}\left(\|f\|_{\infty,U}+\max_{v\in V}\|\partial_v f\|_{\infty,U}\right)+\sqrt{2nT}\,\|f\|_{\infty,U}\,B.$$

This gives the announced result. □

Remark 3.5
This lemma can be improved in the case of a complete interference graph: in this case we use the bounds

$$\frac{\sigma_v}{1+(q_v+1)^a}\leqslant\frac{1}{\min_{q\in U,w\in V}q_w^a}$$

and

$$\prod_{w\sim v}(1-\sigma_w)(1-\sigma_v)=\mathbb{1}_{\sigma=0}$$

in (3.9) *and* (3.10). *Recall that* $\sigma_{\boldsymbol{0}} := \mathbb{1}_{\sigma=\boldsymbol{0}}$. *We will then control*

$$\int_0^T \sigma_{\boldsymbol{0}}(s)\,\mathrm{d}s.$$

The bound then becomes

$$\mathbb{E}\left[\sup_{0\leqslant t\leqslant T\wedge\bar\tau^U}|M_F(t)|\right] \leqslant C\Omega\sqrt{T}\,\|f\|_{\infty,U}\sqrt{\frac{1}{\min_{q\in U,v\in V}q_v^a}}+C\Omega\sqrt{T}\max_{v\in V}\|\partial_v f\|_{\infty,U}$$
$$+C\sqrt{\mathbb{E}\left[\int_0^{T\wedge\bar\tau^U}\sigma_{\boldsymbol{0}}(s)\mathrm{d}s\right]}\Omega\,\|f\|_{\infty,U}+C\sqrt{T}\,\|f\|_{\infty,U}\,B.$$

With this improvement, and $T\geqslant 1$ *the bound in Theorem* 3.1 *becomes*

$$\mathbb{E}\left[\sup_{t\leqslant T\wedge\bar\tau^U}\left|\int_0^t\left(L_s^{\sigma(s)}-L_h\right)[g](Q(s))\mathrm{d}s\right|\right]\leqslant CB_0\max_{v\in V}\|\partial_v g\|_{\infty,U}\,T$$
$$+CT\Omega_0\max_{v,w\in V}\left\|\partial^2_{v,w}g\right\|_{\infty,U}+C\Omega_0\max_{j\in V}\|\partial_j g\|_{\infty,U}\left(\sqrt{\frac{T}{\min_{q\in U,v\in V}q_v^a}}+\sqrt{\mathbb{E}\left[\int_0^{T\wedge\bar\tau^U}\sigma_{\boldsymbol{0}}(s)\mathrm{d}s\right]}\right)$$
$$+C\Omega_0\max_{j\in V}\|\partial_j g\|_{\infty,U}.$$

Next let's prove Lemma 3.3 and explain how to use this bound in order to control the homogenization term.

Proof of Lemma 3.3. Using decomposition as in (3.6), we obtain

$$\sup_{0\leqslant t\leqslant T\wedge\bar\tau^U}\left|\int_0^t\left(\sigma_v(s)-\pi^{Q(s)}(\sigma_v=1)\right)f\left(Q(s)\right)\mathrm{d}s\right|\leqslant|F(Q(0),\sigma(0))|$$
$$+\sup_{0\leqslant t\leqslant T\wedge\bar\tau^U}|F(Q(t),\sigma(t))|$$
$$+\sup_{0\leqslant t\leqslant T\wedge\bar\tau^U}\left|\int_0^t L_{\mathrm{s}}^{\sigma(s)}\left[F\left(\cdot,\sigma(s)\right)\right](Q(s))\mathrm{d}s\right|$$
$$+\sup_{0\leqslant t\leqslant T\wedge\bar\tau^U}|M_F(t)|.$$

As $Q(t)\in U$ for $t\leqslant\bar\tau^U$, using (3.7) and (3.8), we have a control on the three first terms in the right-hand side of the previous display. First by definition of F and Ω,

$$|F(Q(0),\sigma(0))|+\sup_{0\leqslant t\leqslant T\wedge\bar\tau^U}|F(Q(t),\sigma(t))|\leqslant C\|f\|_{\infty,U}\Omega.$$

Second, recall that

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

so it is essentially a sum of discrete derivatives, and thus by (3.8),

$$\sup_{0\leqslant t\leqslant T\wedge\bar{\tau}^U}\left|\int_0^t L_{\mathrm{s}}^{\sigma(s)}\left[F\left(\,\cdot\,,\sigma(s)\right)\right]\left(Q(s)\right)\mathrm{d}s\right|\leqslant CT\Omega\max_{j\in V}\left\|\partial_j f\right\|_{\infty,U}+CT\|f\|_{\infty,U}B.$$

Third, by Lemma 3.4,

$$\mathbb{E}\left[\sup_{0\leqslant t\leqslant T\wedge\bar{\tau}^U}\left|M_F(t)\right|\right]\leqslant C\Omega\sqrt{T}\left(\left\|f\right\|_{\infty,U}+\max_{j\in V}\left\|\partial_j f\right\|_{\infty,U}\right)+C\sqrt{T}\left\|f\right\|_{\infty,U}B.$$

Combining these bounds gives the result:

$$\begin{aligned}\sup_{0\leqslant t\leqslant T\wedge\bar{\tau}^U}\left|\int_0^t\left(\sigma_v(s)-\pi^{Q(s)}(\sigma_v=1)\right)f\left(Q(s)\right)\mathrm{d}s\right|\leqslant{}&C\Omega\sqrt{T}\left\|f\right\|_{\infty,U}\\&+C\Omega\sqrt{T}\max_{j\in V}\left\|\partial_j f\right\|_{\infty,U}\\&+C\sqrt{T}\left\|f\right\|_{\infty,U}B\\&+CT\Omega\max_{j\in V}\left\|\partial_j f\right\|_{\infty,U}\\&+CT\|f\|_{\infty,U}B\\&+C\|f\|_{\infty,U}\Omega.\end{aligned}$$

Factorizing this inequality gives the result. □

3.2.2 Control of solutions to the Poisson equation

In the previous section we have established a bound on some averaging quantity in terms of the constants Ω and B. The goal of this section is to provide a bound on those quantities.

For $q\in\mathbb{N}^V$ introduce α^q and ℓ^q are the log-Sobolev constant and spectral gap associated to L_{f}^q, respectively, and recall $\phi_g(q,\,\cdot\,)$ the solution to the Poisson equation

$$L_{\mathrm{f}}^q[\varphi]=g-\pi^q[\varphi],\ \pi^q[\varphi]=0.$$

Lemma 3.6
For $q\in\mathbb{N}^V$ and $v\in V$ let

$$\Omega(q)=\left(-\log(\pi^q(\boldsymbol{0}))\right)^{3/2}\frac{1}{\ell^q}$$

and

$$B_v(q)=\frac{\Omega(q)}{q_v+1}+\frac{\Omega(q)^2}{q_v^{1+a}}.$$

Then

$$\left\|\phi_g(q,\cdot)\right\|_\infty\leqslant C\left\|g\right\|_\infty\Omega(q)\ \textit{and}\ \max_{\sigma\in S(G)}\left|\phi_g(q\pm e^v,\sigma)-\phi_g(q,\sigma)\right|\leqslant C\left\|g\right\|_\infty B_v(q).\tag{3.11}$$

Moreover,

$$\sup_{q\in U}\Omega(q)\leqslant\Omega_0\ \textit{and}\ \sup_{q\in U,v\in V}B_v(q)\leqslant B_0.\tag{3.12}$$

Proof. The bounds from (3.12) are direct consequences of the definitions of U, $\Omega(q)$, Ω_0, $B_v(q)$ and B .
The first bound of (3.11) on $\|\phi_g(q,\,\cdot\,)\|_\infty$ is actually the result of Proposition 2.14.
We now prove the second bound of (3.11). Fix temporarily $v \in V$, and let $q \in \mathbb{N}^V$ with $q_v > 0$. Let $\Phi \coloneqq \phi_g(q-e^v,\,\cdot\,) - \phi_g(q,\,\cdot\,)$ and

$$H(q,\sigma) \coloneqq L_{\mathrm{f}}^q[\Phi](\sigma).$$

Since $\pi^q[L_{\mathrm{f}}^q[f]] = 0$ for any f, $\pi^q[H(q,\cdot)] = 0$ by the previous identity. In addition, again by the previous identity, Φ is the solution to the Poisson equation associated to $H(q,\cdot)$: $\Phi(q,\cdot) = \phi_{H(q,\cdot)}(q,\cdot)$ and the first bound in (3.11) implies

$$\|\phi_g(q-e^v,\,\cdot\,) - \phi_g(q,\,\cdot\,)\|_\infty \leqslant \Omega(q)\,\|H\|_{\infty,U}\,.$$

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

$$\begin{aligned} H(q,\sigma) &= L_{\mathrm{f}}^q[\phi_g(q-e^v,\,\cdot\,) - \phi_g(q,\,\cdot\,)](\sigma) \\ &= \left(L_{\mathrm{f}}^q - L_{\mathrm{f}}^{q-e^v}\right)[\phi_g(q-e^v,\,\cdot\,)](\sigma) + L_{\mathrm{f}}^{q-e^v}[\phi_g(q-e^v,\,\cdot\,)](\sigma) - L_{\mathrm{f}}^q[\phi_g(q,\,\cdot\,)](\sigma) \\ &= \left(L_{\mathrm{f}}^q - L_{\mathrm{f}}^{q-e^v}\right)[\phi_g(q-e^v,\,\cdot\,)]\,(\sigma) - \sum_{\rho\in S(G)} (\pi^{q-e^v}(\rho) - \pi^q(\rho))g(\rho) \end{aligned}$$

and so

$$\|H(q,\cdot)\|_\infty \leqslant \left\|\left(L_{\mathrm{f}}^{q-e^v} - L_{\mathrm{f}}^q\right)[\phi_g(q-e^v,\,\cdot\,)]\right\|_\infty + \|g\|_\infty \sum_\sigma \left|\pi^{q-e^v}(\sigma) - \pi^q(\sigma)\right|.$$

For any function h we have according to (2.2)

$$\begin{aligned} \left(L_{\mathrm{f}}^{q-e^v} - L_{\mathrm{f}}^q\right)[h](\sigma) = \sigma_v\,(\Psi_-(q_v-1) - \Psi_-(q_v))\,(h(\sigma-e^v) - h(\sigma)) \\ + \sigma_{\mathbf{0}}\,(\Psi_+(q_v-1) - \Psi_+(q_v))\,(h(\sigma+e^v) - h(\sigma)) \end{aligned}$$

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

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

Therefore, using again the bound (3.11) gives

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

Direct calculation yields

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

and so $\left|\Psi_-'(q_v-u)\right| \leqslant \frac{q_v^{a-1}}{(1+q_v^a)^2} \leqslant q_v^{-1-a}$ as long as $u \leqslant 1$. Similarly, we now need to compute the partial derivative for $\pi^q(\sigma)$. One can check that

$$\partial_v \pi^q(\mathbf{0}) = -\frac{a\pi^q(\mathbf{0})\pi^q(\sigma_v = 1)}{q_v+1},$$

and if σ is not the empty schedule,

$$\partial_v \pi^q(\sigma) = -\frac{a\pi^q(\sigma)\pi^q(\sigma_v = 1)}{q_v + 1} + \frac{a\pi^q(\sigma)}{q_v + 1}\mathbb{1}_{\sigma_v=1}.$$

In any case, $|\partial_v \pi^q(\sigma)| \leqslant \frac{C}{q_v+1}$. The same reasoning still applies for $q + e^v$. Gathering the previous bounds gives the result. □

Proof of Theorem 3.1. In Lemma 3.3, take $f_v(q) \coloneqq g(q-e^v)-g(q)$. Notice that if g is twice differentiable, for every v, f_v is also twice differentiable. Notice as well that $\|f_v\|_{\infty,U} \leqslant \max_v \|\partial_v g\|_{\infty,U}$ and $\max_w \|\partial_w f_v\|_{\infty,U} \leqslant \max_{v,w} \left\|\partial^2_{v,w} g\right\|_\infty$. Lemma 3.6 states that $\Omega \leqslant \Omega_0$ and $B \leqslant B_0$ so we get the result by summing over V. □

3.3 Scaled process

3.3.1 General considerations

The explicit bounds that we obtained in the previous sections allow us to go further and obtain some asymptotic results for renormalizations of the process. We give a proof for a bound here and refer to the next chapters for more applications and detailed discussions. We consider queue size starting from a large order of magnitude N. We want to study the evolution of $\dfrac{Q}{N}$. Since we rescale the process in space, it takes a long time to evolve, that is why we also speed up time by N^θ to have a non-trivial dynamic. We consider $\theta > 0$ fixed in this section. Let

$$Q^N(t) \coloneqq \frac{Q(N^\theta t)}{N}$$

and

$$\sigma^N(t) \coloneqq \sigma(N^\theta t),$$

with $\theta \in \{1, 1+a\}$. The case $\theta = 1$ will give us a first order approximation *à la* law of large number, we will also consider $\theta = 1 + a$ in the critical case to obtain a heavy traffic result. We compare the evolution of the network with the evolution in the homogenized case. The generators for the scaled queueing process thus become for any $(q, \sigma) \in \frac{1}{N}\mathbb{N}^V \times S(G)$,

$$L^\sigma_{\mathrm{s},N}[f](q) \coloneqq N^\theta \sum_{v\in V} \lambda_v \left(f\left(q + \frac{e^v}{N}\right) - f(q)\right) + \sigma_v \mathbb{1}_{q_v>0} \left(f\left(q - \frac{e^v}{N}\right) - f(q)\right).$$

The generator for the fast process becomes for any $(q, \sigma) \in \frac{1}{N}\mathbb{N}^V \times S(G)$,

$$\begin{aligned} L^q_{\mathrm{f},N}[h](\sigma) \coloneqq{}& N^\theta \sum_{v\in V} \sigma_v \Psi_-(Nq_v)\left(h(\sigma - e^v) - h(\sigma)\right) \\ &+ N^\theta \sum_{v\in V} \prod_{w\sim v} (1-\sigma_w)(1-\sigma_v)\Psi_+(Nq_v)\left(h(\sigma + e^v) - h(\sigma)\right), \end{aligned}$$

and the homogenized generator:

$$L_{\mathrm{h},N}[f](q) \coloneqq N^\theta \sum_{v\in V}\left[\lambda_v\left(f\left(q+\frac{e^v}{N}\right)-f(q)\right)+\pi^{Nq}(\sigma_v=1)\mathbb{1}_{q_v>0}\left(f\left(q-\frac{e^v}{N}\right)-f(q)\right)\right].$$

This leads us to consider

$$\begin{aligned}
&\int_0^t \left(L_{\mathrm{s},N}^{\sigma^N(s)}-L_{\mathrm{h},N}\right)[g](Q^N(s))\mathrm{d}s\\
=&N^\theta\int_0^t\sum_{v\in V}\left(\sigma_v^N(s)-\pi^{NQ^N(s)}(\sigma_v=1)\right)\left(g\left(Q^N(s)-\frac{e^v}{N}\right)-g(Q^N(s))\right)\mathrm{d}s\\
=&\int_0^{N^\theta t}\sum_{v\in V}\left(\sigma_v(s)-\pi^{Q(s)}(\sigma_v=1)\right)\left(g\left(Q^N(s)-\frac{e^v}{N}\right)-g(Q^N(s))\right)\mathrm{d}s\\
=&\int_0^{N^\theta t}\left(L_{\mathrm{s}}^{\sigma(s)}-L_{\mathrm{h}}\right)[g^N](Q(s))\mathrm{d}s, \qquad (3.13)
\end{aligned}$$

with $g^N(q)=g(\frac{q}{N})$. For localization, we will need a candidate limiting process for explicit bounds. The localization set U for Q^N will take the form:

$$U \coloneqq \{q\in\mathbb{R}^V \mid q\in(C_-,C_+)^V\}, \qquad (3.14)$$

with $C_-<0$ and $C_+>0$ independent from N. We then use Theorem 3.1 with g^N and $T'=N^\theta T$. In Chapter 4 we will choose $\theta=1$ to obtain a fluid limits first order approximation. In Chapter 5 we will choose $\theta=1+a$. We will see that $\theta=1+a$ is the correct time scale to see the evolution of the sum of coordinates in the critical case in a complete interference graph. We will derive a second order heavy traffic approximation. See the discussions in Chapter 5 for more details on this time scale

Corollary 3.7
Assume $Q^N(0)\to q^0\in(C_-,C_+)^V$, let $g:\mathbb{R}_+^V\to\mathbb{R}$ be twice differentiable. Assume that there exist β and $C\in(0,+\infty)$ such that for any $q\in U$ and $N\geqslant 1$,

$$\ell^{Nq}\geqslant C\left\|Nq+1\right\|_\infty^{-a\beta}.$$

Then for any $\theta>0$ and N large enough,

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\leqslant T\wedge\bar\tau^U(Q^N)}\left|\int_0^t\left(L_{\mathrm{s},N}^{\sigma^N(s)}-L_{\mathrm{h},N}\right)(g)(Q^N(s))\mathrm{d}s\right|\right]&\leqslant CT\max_v\left\|\partial_v g\right\|_{\infty,U}N^{\theta+a(2\beta-1)-2}\log(N)^3\\
&+CT\max_{v,w\in V}\left\|\partial^2_{v,w}g\right\|_{\infty,U}N^{\theta+a\beta-2}\log(N)^{3/2}\\
&+C\sqrt{T}\max_v\left\|\partial_v g\right\|_{\infty,U}N^{\frac{\theta}{2}+a\beta-1}\log(N)^{3/2}.
\end{aligned}$$

In the case of a complete interference graph and $\theta=1+a$, we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\leqslant T\wedge\bar\tau^U(Q^N)}\left|\int_0^t\left(L_{\mathrm{s},N}^{\sigma^N(s)}-L_{\mathrm{h},N}\right)(g)(Q^N(s))\mathrm{d}s\right|\right]&\leqslant CT\max_v\left\|\partial_v g\right\|_{\infty,U}N^{a-\frac12}\log(N)^3\\
&+CT\max_{v,w\in V}\left\|\partial^2_{v,w}g\right\|_{\infty,U}N^{2a-1}\log(N)^{3/2}.
\end{aligned}$$

Proof. Picking up from (3.13), we need to provide a bound on

$$\mathbb{E}\left[\sup_{0\leqslant t\leqslant N^\theta T\wedge\bar\tau^U}\int_0^{N^\theta t}\left(L_{\mathrm{s}}^{\sigma(s)}-L_{\mathrm{h}}\right)[g^N](Q(s))\mathrm{d}s\right]$$

Recall the result from Theorem 3.1

$$\mathbb{E}\left[\sup_{t\leqslant T\wedge\bar\tau^U}\left|\int_0^t(L_s^{\sigma(s)}-L_h)[g](Q(s))\mathrm{d}s\right|\right]\leqslant CB_0\max_{v\in V}\|\partial_v g\|_{\infty,U}\,T$$
$$+C\sqrt{T}\Omega_0\left(\max_{v\in V}\|\partial_v g\|_{\infty,U}+\sqrt{T}\max_{w,v\in V}\left\|\partial^2_{w,v}g\right\|_{\infty,U}\right),$$

For the first part let's apply Theorem 3.1 with a time horizon $N^\theta T$ and g^N. By stopping the process before $\tau^U(Q^N)$, using Lemma 3.6 and the localization, we are able to prove that $\Omega_0\leqslant CN^{a\beta}\log(N)^{3/2}$. From Lemma 2.16, we get that β cannot be smaller than 1. For this reason $a\beta-1\leqslant 2a\beta-1-a$ and

$$B_0\leqslant CN^{2a\beta-1-a}\log(N)^3+CN^{a\beta-1}\log(N)^{3/2}\leqslant CN^{2a\beta-1-a}\log(N)^3.$$

Next, for any $v\in V$, $\left\|\partial_v g^N\right\|_{\infty,NU}=\frac{1}{N}\left\|\partial_v g\right\|_{\infty,U}$.

$$\mathbb{E}\left[\sup_{0\leqslant t\leqslant N^\theta T\wedge\bar\tau^U}\int_0^{N^\theta t}\left(L_{\mathrm{s}}^{\sigma(s)}-L_{\mathrm{h}}\right)[g^N](Q(s))\mathrm{d}s\right]$$
$$\leqslant CN^{2a\beta-1-a}\log(N)^3N^{-1}\max_{v\in V}\|\partial_v g\|_{\infty,U}\,N^\theta T$$
$$+C\sqrt{N^\theta T}N^{a\beta}\log(N)^{3/2}\left(\max_{v\in V}\|\partial_v g\|_{\infty,U}\,N^{-1}+\sqrt{N^\theta T}N^{-2}\max_{w,v\in V}\left\|\partial^2_{w,v}g\right\|_{\infty,U}\right).$$

Gathering the terms, we get the first result.

The second part is a direct consequence of Lemma 2.15, Remark 3.5 and Lemma 5.8. Lemma 2.15 states that in the case of a complete interference graph $\beta=1$, we will prove in Lemma 5.8 in Chapter 5 that,

$$\mathbb{E}\left[\int_0^{N^{1+a}(T\wedge\bar\tau^U(Q^N))}\sigma_{\mathbf{0}}(s)\mathrm{d}s\right]\leqslant CT(N+N^{2a}).$$

The localization argument will allow us to bound scaled queue lengths away from 0 and thus $\frac{1}{\min_{q\in U^N,v\in V}q_v}\leqslant CN^{-a}$. For more details regarding either scalings, see the respective chapter. □

3.3.2 Fluid limits in a complete interference graph

Let's consider the case $\theta=1$. Integrate the scaled T and $\|\partial_v g\|_{\infty,U}$ and $\max_{w,v\in V}\left\|\partial^2_{w,v}g\right\|_{\infty,U}$ in the constant C. In the full interference graph, $\beta=1$ and the term of leading

order of magnitude becomes $N^{a-\frac{1}{2}}$ and we get

$$\mathbb{E}\left[\sup_{t\leqslant T\wedge\bar{\tau}^U(Q^N)}\left|\int_0^t\left(L_{\mathrm{s},N}^{\sigma^N(s)}-L_{\mathrm{h},N}\right)(g)(Q^N(s))\mathrm{d}s\right|\right]\leqslant CN^{a-\frac{1}{2}}\log(N)^{3/2}$$

Our condition for the homogenization error to vanish as $N\to+\infty$ becomes $a<\frac{1}{2}$ and that may be somewhat surprising. On the one hand when queue lengths are of an order of magnitude N, they take a time of order N to evolve. On the other hand, the mixing time of L_{f}^{Nq} is of order N^a by Lemma 2.15 so we could expect homogenization to occur as soon as $a<1$ because in this case the schedule homogenize faster than the time it takes for queue lengths to evolve. Instead, we are able to prove that homogenization holds as soon as $a<1/2$. We do not know whether our condition for homogenization is tight. We may have lost possible values for a because of the rough absolute value bound we use

$$|\phi_v(q,\sigma)-\phi_v(q,\sigma')|\leqslant 2\Omega$$

in (3.9) and (3.10) for instance. We would either need another proof method or better knowledge of the solution to some Poisson equations to check if the condition is also necessary. Whether the error terms from Corollary 3.7 continue to vanish for $1/2<a<1$ and $\theta=1$ constitutes in our view an interesting open problem, which also testifies from the difficulty to prove fully coupled stochastic averaging principles even in seemingly simple cases.

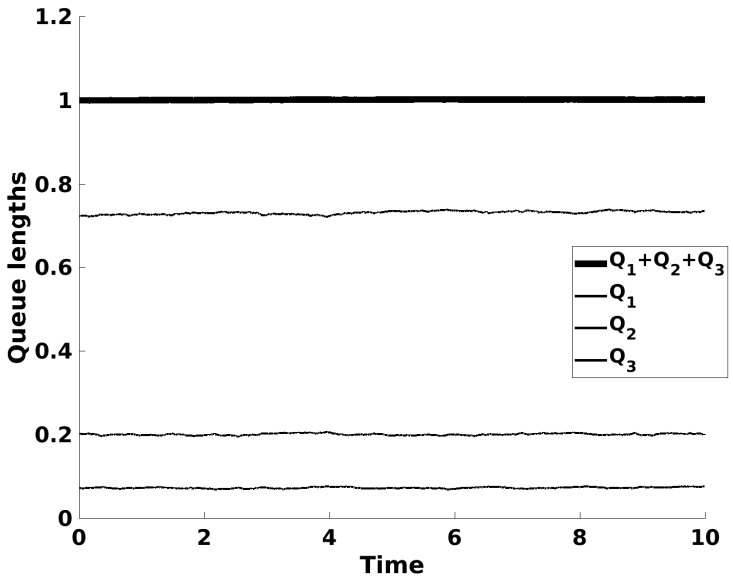

Figure 3.1: Evolution of 3 nodes with QB-CSMA, a=0.4, N=1e7

Both figures represent the evolution of queue lengths for the same $\lambda=(0.2,0.3,0.5)$ but different values for a. In both cases, the sum of coordinates converges to a constant process but when $a>0.5$, it is unclear wether the convergence still holds. If it does, the convergence seems a lot slower as soon as $a>0.5$.

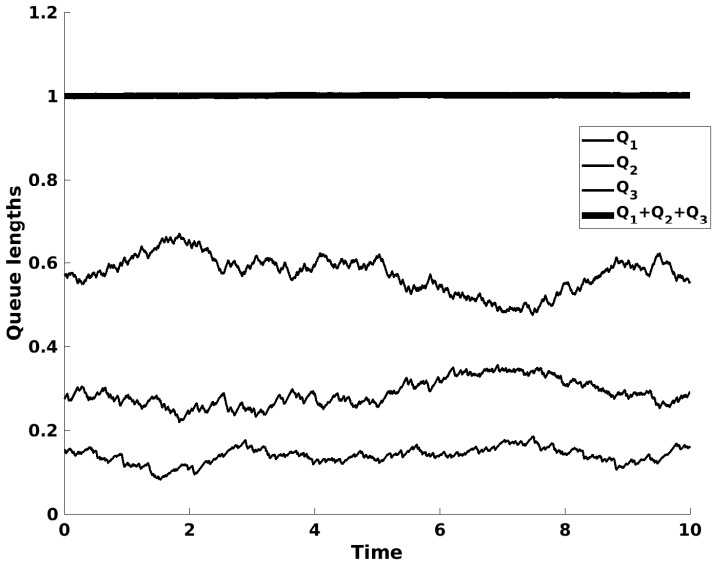

Figure 3.2: Evolution of 3 nodes with QB-CSMA, a=0.7, N=1e7

Chapter 4

Fluid Limits of QB-CSMA

Contents

4.1 Context 63
4.2 Main results and heuristic 64
4.2.1 Limiting process: general interference graph 64
4.2.2 Identification of the limit: Complete interference graph . . . 66
4.2.3 Main result 68
4.2.4 Heuristic, line network 69
4.3 Preliminary 73
4.3.1 Localization and homogenization 75
4.3.2 Proof of preliminary results 76
4.4 General interference graph up to $\tau^0(q^*)$, Theorem 4.5 . . 80
4.5 Complete interference graph 83
4.5.1 Proof of Theorem 4.6 83

4.1 Context

In this chapter, we will prove fluid limit results using the homogenization result from last chapter. As before we ignore the dependency in G for all notations. We will distinguish two cases: general and full interference graphs and specify only when necessary. Without any information on the interference graph, starting from a positive initial condition at each node, we are able to prove convergence of the fluid limits to a deterministic process governed by an ODE up to the time this ODE hits 0 in one of its coordinates. In the case of a complete interference graph, we are able to go further and distinguish (sub/super)critical cases giving three possible behaviors and we give a proof for convergence of the fluid limits for any initial condition, any arrival rates and any finite time horizon. In this chapter, the process

of interest will be denoted (Q^N, σ^N) and is given by

$$Q^N(t) := \frac{Q(Nt)}{N} \text{ and } \sigma^N(t) := \sigma(Nt).$$

To be rigorous, we would have to add a superscript to the unscaled process: we actually consider a sequence of networks where the initial condition may depend on N. For the sake of clarity of exposition, we will omit this dependency in N for the initial state. We will consider situations where $\frac{Q(0)}{N} \to q^0 > 0$ as $N \to +\infty$. We will use the scaled generators from Section 3.3 $L^{\sigma}_{\mathrm{s},N}$ and $L^{q}_{\mathrm{f},N}$ with $\theta = 1$.

In this case the identification step from [Kur92] mentioned in the introduction Section 1.3.1 is not clear because deactivation rates converge to 0 in the limit. In particular, the asymptotic generator

$$L_{\mathrm{f}}^{\infty,q} := \lim_{N\to+\infty} L_{\mathrm{f}}^{Nq}$$

describes a reducible dynamic: it starts at $\mathbf{0}$ and then jumps to one of the possible states $e^v \in S(G)$ from which schedules where nodes interfering with v are active become inaccessible as long as the scaled queue lengths remain positive. A situation with the absence of uniqueness has been studied in [HK94] for a general loss network. They show that any accumulation point must be a linear combination of the different stationary measures but no general method seems to exist to characterize this combination. To be more precise, it is possible to identify any limiting point of the occupation measure of the schedule as an invariant measure of $L_{\mathrm{f}}^{\infty,q}$, see Theorem 3 of [HK94]. In our case, for any bounded $f : S(G) \to \mathbb{R}_+$ and any q bounded away from 0, $L_{\mathrm{f}}^{\infty,q}(f)$ is given by

$$L_{\mathrm{f}}^{\infty,q}(f)(\sigma) = \sum_{v\in V} \prod_{w\sim v} (1-\sigma_w)(1-\sigma_v)\left(h(\sigma+e^v) - h(\sigma)\right).$$

Any Dirac measure on maximal stable sets (or mixture of such measures) is an invariant measure for this generator and thus this will not be enough to identify the limit. We postpone the discussion about the method from [LN13] to the next chapter as the interest of our method becomes more obvious.

4.2 Main results and heuristic

4.2.1 Limiting process: general interference graph

Recall the definition of S^*:

$$S^* = \{\sigma \in S(G) \mid |\sigma| = \Upsilon\},$$

and the definition of the stoping times τ^ϵ:

$$\tau^\epsilon(f) = \inf\{t > 0, \min_v f_v(t) \leqslant \epsilon\}.$$

We begin by identifying a potential limit using the homogenized process. On the homogenized process, because of Lemma 2.4, we will localize the process such that

the asymptotic service rate will be given for any $v \in V$ by

$$\bar{\pi}^q(v) = \pi^q_\infty(\sigma_v = 1) = \sum_{\sigma \in S^*} \frac{\sigma_v \prod\limits_{w \in V} q_w^{a\sigma_w}}{\sum\limits_{\rho \in S^*} \prod\limits_{w \in V} q_w^{a\rho_w}}.$$

Let

$$\begin{array}{ccc} g : (0,+\infty)^V & \to [-1,1]^V \\ q & \longmapsto \lambda - \bar{\pi}^q \end{array} \tag{4.1}$$

be the difference between arrival rates and homogenized departure rates. The function g is locally Lipschitz on $(0,+\infty)^V$ by differentiability. Let $q^0 \in (0,+\infty)^V$. We will consider the equation with unknown f with values in $(0,+\infty)^V$:

$$\begin{cases} f' = g(f) \\ f(0) = q^0 \end{cases}. \tag{4.2}$$

Lemma 4.1
For any $q^0 \in (0,+\infty)^V$, there exists at least one solution of (4.2) defined in an open interval $D \subset \mathbb{R}$ containing 0.
If D_1 and D_2 are two open intervals containing 0 such that $q^1 : D_1 \to (0,+\infty)^V$ and $q_2 : D_2 \to (0,+\infty)^V$ are two solutions of (4.2), then $q^1(t) = q^2(t)$ for any $t \in D_1 \cap D_2$.

Proof. Since g is locally Lipschitz on $(0,+\infty)^V$, this is a direct application of the Cauchy Lipschitz Theorem, see Chapter 5 Section 3.1 and 3.3 of [Dem16]. □

Technically, each definition set I gives a different solution but since for any given $t \in \mathbb{R}$ there is at most one possible value for the solution evaluated at time t, any solution can be extended to a unique maximal solution that has the largest (for the inclusion) definition interval. There exists an open interval $D^* \subset \mathbb{R}$ such that any solution to (4.2) is a restriction of the maximal solution defined on D^*. By the maximality criterion of [Dem16] Chapter 5 Section 2.6, the solution $q^{\max} : D^* \to (0,+\infty)^V$ is maximal if it escapes any compact of $(0,+\infty)^V$ when $t \to \sup D^*$ or $\sup D^* = +\infty$ and similarly, $\inf D^* = -\infty$ or the solution escapes compacts of $(0,+\infty)^V$ before that time. By the characterization of compact of $(0,+\infty)^V$ this means that $\sup D^* = +\infty$ or $\lim_{t \to \sup D^*} q^*_v(t) = 0$ or $+\infty$ for some $v \in V$. The exit time only depends on the initial condition and we denote it

$$T_{\text{ext}}(q^0) \coloneqq \sup D^*.$$

In fact we are interested in solutions with a domain contained in $[0,+\infty)$.

Definition 4.2
For any $q^0 \in (0,+\infty)^V$, we call

$$q^*(\cdot, q^0) : [0,+\infty) \to \mathbb{R}_+^V$$

the restriction on $[0,+\infty)$ of the maximal solution of (4.2) with initial condition q^0. The solution is extended after $T_{ext}(q^0)$ by stating that

$$\forall s \geqslant 0,\ q^*(T_{\text{ext}}(q^0) + s, q^0) = \lim_{t \to T_{\text{ext}}(q^0)} q^*(t, q^0).$$

4.2.2 Identification of the limit: Complete interference graph

In this section, we go further in the case where G is a complete interference graph. Let's define a slightly different equation. We introduce the next equation in order to specify what happens when approaching $\mathbf{0}$ or starting in this state. When G is a complete interference graph, let's define q_c^* as the only (existence and uniqueness are proved in Lemma 4.3) function $f : \mathbb{R}_+ \to \mathbb{R}_+^V$ that is continuous, differentiable almost everywhere (everywhere except at $t = 0$ if $q^0 = \mathbf{0}$ or $\inf\{t > 0, f(t) = \mathbf{0}\}$), such that

$$\begin{cases} f' = g(f) & \text{if } f \neq \mathbf{0} \\ s_1(f)(t) = \max(s_1(q^0) + (s_1(\lambda) - 1)t, 0) & \\ f(0) = q^0 & \end{cases} \tag{4.3}$$

We now give an informal description of two important lemmas that will justify this definition and link solutions to this initial value problem with solutions to (4.2).

We will show in Lemma 4.4 that the only possibility for queue lengths to reach 0 is if they do it at the same time. Imposing this behavior for the sum of coordinates of the solution to (4.3) ensures that either the sum of coordinates is increasing and is never null for $t > 0$ or the sum of coordinates decreases to 0 and stays absorbed. Hence this behavior is consistent with the extension of solutions to (4.2) from Definition 4.2. We will see in Lemma 4.3 that this is enough to uniquely characterize the solution.

As long as $q_c^*(t) \neq \mathbf{0}$ the condition on the sum is redundant by summing the derivatives over V. Without the condition on the sum, the initial value problem is not well defined when $q^0 = \mathbf{0}$. We explain in the next discussion why the condition on the sum allows for a unique characterization of the solution. By imposing the condition

$$s_1(q_c^*)(t) = \max(s_1(q^0) + (s_1(\lambda) - 1)t, 0),$$

it ensures that if $s_1(\lambda) > 1$ and $q^0 = \mathbf{0}$ the solution escapes $\mathbf{0}$ instantaneously and if $s_1(\lambda) \leqslant 1$ and $s_1(q_c^*(t)) = \mathbf{0}$ then $s_1(q_c^*(t+s)) = \mathbf{0}$ for any $s \geqslant 0$.

Lemma 4.3 states that $q^*(\cdot, q^N)$ with $q^N \in (0, +\infty)^V$ has a unique limit as $q^N \to q^0 \in \mathbb{R}_+^V$. This limit is noted $q_c^*(\cdot, q^0)$.

- If $q^0 \in (0, +\infty)^V$ this limit is the solution to (4.2) with initial condition q^0.
- If $q^0 \in \mathbb{R}_+^V \setminus \left((0, +\infty)^V \cup \{\mathbf{0}\}\right)$, the initial value problem (4.2) has a unique solution on $[0, +\infty)$ for a complete graph.
- If $q^0 = \mathbf{0}$ and $s_1(\lambda) \leqslant 1$ the limit is the process constant equal to $\mathbf{0}$.
- If $q^0 = \mathbf{0}$ and $s_1(\lambda) > 1$, there exists a unique continuous function differentiable everywhere except at $t = 0$ such that $f(0) = \mathbf{0}$ and $s_1(f)(t) = (s_1(\lambda) - 1)t$ for any $t \geqslant 0$ and $f'(t) = g(f(t))$ for any $t > 0$. We also call it $q_c^*(\cdot, \mathbf{0})$ and we have $q^*(\cdot, q^N) \to q_c^*(\cdot, \mathbf{0})$.

If $q^0 \in \mathbb{R}_+^V \setminus \left((0, +\infty)^V \cup \{\mathbf{0}\}\right)$, let's extend the definition of T_{ext}:

$$T_{\text{ext}}(q^0) \coloneqq \lim_{N \to +\infty} T_{\text{ext}}(q^N).$$

If $s_1(\lambda) > 1$, we will see that $T_{\text{ext}}(q^0) = +\infty$ so there is a unique solution to (4.2) whose domain is $[0, +\infty)$. If $s_1(\lambda) = 1$ and $s_1(q^0) > 0$ the same behavior occurs.

If $s_1(\lambda) < 1$, all the coordinates reach 0 at the same time. After that time, we will prove that the process of queue lengths remains absorbed in $\mathbf{0}$ on the fluid scale. Because of that, it is natural to further extend the domain of the solution q^* up to $(0, +\infty)$ by stating that in this case once $q^*(t) = \mathbf{0}$, $q^*(t + s) = \mathbf{0}$ for any $s \geqslant 0$. We now give a more formal description of the behavior mentioned above.

Lemma 4.3
Let $q^N \in (0, +\infty)^V$ such that $q^N \to q^0 \in \mathbb{R}_+^V$.

- *If $q_v^0 > 0$ for all $v \in V$, $q^*(\cdot, q^N) \to q^*(\cdot, q^0)$ uniformly on compact time sets .*
- *If there exists $v \in V$ such that $q_v^0 = 0$, the limit is still unique and*

$$\lim_{N \to +\infty} q^*(\cdot, q^N) \coloneqq q^*(\cdot, q^0)$$

  is a solution to the initial value problem (4.3).
- *For any $q^0 \in \mathbb{R}_+^V$, the solution to the initial value problem (4.3) is unique.*

Proof. See Section 4.3.2 for a proof of this result. □

Uniqueness in this lemma is in the same sense as in the previous one. To prove this lemma, we explain in the next lemma why the complete interference graph constitutes a particular case: the dynamic (4.3) ensures that all coordinates touch 0 at the same time if they do at all. Recall the definition of the stopping times

$$\tau^0(f) = \inf\{t > 0,\ \min_v f_v(t) \leqslant 0\}.$$

The next lemma is required to prove uniqueness of solutions of (4.3). We will justify the existence of solutions in the proof of this result and then prove that the properties desribed hold for any solution to (4.3).

Lemma 4.4
Let $q^0 \in \mathbb{R}_+^V$ and q^* any solution to (4.3). Then $\tau^0(q_c^*) = \tau^0(s_1 \circ q_c^*)$. In addition,

- *If $s_1(\lambda) > 1$, $\tau^0(s_1 \circ q_c^*) = +\infty$.*
- *If $s_1(\lambda) < 1$, $\tau^0(s_1 \circ q_c^*) = \frac{s_1(q^0)}{1 - s_1(\lambda)}$*
- *If $s_1(\lambda) = 1$ and $q^0 \neq \mathbf{0}$, $\tau^0(s_1 \circ q_c^*) = +\infty$.*
- *If $s_1(\lambda) = 1$ and $q^0 = \mathbf{0}$, $\tau^0(s_1 \circ q_c^*) = 0$.*

Proof stub. The general idea is that no coordinate cannot be decreasing when too small compared to the sum. The expressions for $\tau^0(s_1 \circ q_c^*)$ are direct consequences of the description of the sum. See Section 4.3.2 for a proof of this result. □

Because of this Lemma, we keep the $q^*(\cdot, q^0)$ notation for solutions to (4.3) when unambiguous: they coincide with the extension given in Definition 4.2 when $q^0 \in (0, +\infty)^V$ and q_c^* is the only solution of (4.2) defined on $[0, +\infty)$ when $q_v^0 = 0$ for some v.

4.2.3 Main result

The main theorem of this chapter concerns the fluid limit of the system. It states that over small enough time horizon, the scaled process of queue lengths converges uniformly in probability to q^* given in the previous section. Recall the definition of the process

$$Q^N = \frac{Q(N\cdot)}{N},$$

and β is such that

$$\ell^q \geqslant C \left\| q + 1 \right\|_\infty^{-a\beta}.$$

For instance $\beta = 1$ in the case of a complete interference graph. Recall that $\xrightarrow{\mathbb{P}}$ denotes the convergence in probability from Definition 2.2.

Theorem 4.5
Assume that the two following assumptions hold:

- $2a\beta < 1$;
- $Q^N(0) \to q^0$ *for some* $q^0 \in (0, +\infty)^V$.

Then $Q^N(\cdot) \xrightarrow{\mathbb{P}} q^*(\cdot, q^0)$ *as* $N \to +\infty$ *uniformly over compact sets of* $[0, T_{ext}(q^0))$.

Because of Lemma 4.4, this theorem is enough to state that in the case of a complete interference graph, if $s_1(\lambda) \geqslant 1$ and $Q^N(0) \to q^0$ with $q_v^0 > 0$ for every $v \in V$,

$$Q^N \xrightarrow{\mathbb{P}} q^*(\cdot, q^0)$$

uniformly over compact time intervals: in this case, $T_{\text{ext}}(q^0) = +\infty$. The comparison to an ODE is troublesome when some queues start empty or after they reach 0 because homogenization may fail when some queues lengths are too small but this does not happen in the super critical case. In general, for any interference constraints and initial condition q^0 such that $T_{\text{ext}}(q^0) = +\infty$ and $q_v^0 > 0$ for every $v \in V$, this theorem is enough to prove convergence of Q^N uniformly over compact of $(0, +\infty)$. We now justify why if $\lambda \in \Lambda^*(S(G))$, $T_{\text{ext}}(q^0) < +\infty$ for any $q^0 \in \mathbb{N}^V$. If $q \neq \mathbf{0}$, $\pi_\infty^q(S^*) = 1$ and thus,

$$\sum_{v \in V} \bar{\pi}^q(v) = \sum_{v \in V} \pi_\infty^q[\sigma_v] = \Upsilon.$$

For any t such that $q^*(t) > 0$ for all $v \in V$,

$$\begin{aligned} s_1(q^*)'(t) &= \sum_{v \in V} \lambda_v - \bar{\pi}^{q^*(t)}(v) \\ &= s_1(\lambda) - \Upsilon. \end{aligned} \tag{4.4}$$

If $\lambda \in (1-\epsilon)\Lambda^*(S(G))$, $s_1(\lambda) \leqslant (1-\epsilon)\Upsilon$: any element of $\Lambda^*(S(G))$ can be expressed as a convex combination of stable sets of size Υ. If $\lambda \in (1-\epsilon)\Lambda^*(S(G))$,

$$s_1(\lambda) \leqslant (1-\epsilon) \max_{\sigma \in S(G)} s_1(\sigma) = \Upsilon.$$

Thus, $s_1(q^*)'(t) \leqslant -\epsilon\Upsilon$ as long as $q^*(t) \neq \mathbf{0}$. For any $\epsilon > 0$, if $\lambda \in (1-\epsilon)\Lambda^*(S(G))$, $s_1(\lambda) \leqslant (1-\epsilon)\Upsilon$ and $s_1(q^*)$ is decreasing linearly. It continues to decrease until at least one of the coordinates reaches 0. When G is a complete graph, it is possible to control potential reflections at the boundary separately to obtain the following result:

Theorem 4.6
Let $q^0 \in [0,+\infty)^V$ and $\lambda \in \mathbb{R}_+^V$ be fixed. Assume the following:

- *$Q^N(0) \to q^0$ as $N \to +\infty$,*
- *G is a complete interference graph,*
- *and $a < \frac{1}{2}$.*

Then $Q^N \xrightarrow{\mathbb{P}} q^*$ uniformly over compact time interval for q^* characterized as the unique solution to (4.3).

The proof of the first theorem relies heavily on two essential results that we prove in Section 4.3. First Q^N is tight and any limiting point is a continuous function. As usual with fluid limits, there does not lie the difficulty and the proof is straightforward. Second the homogenization result Corollary 4.13, a straightforward application of Corollary 3.7.

Theorem 4.5 is a direct consequence of these results. Using homogenization, we are able to bound the distance between Q^N and its homogenized version until q^* reaches the boundary of its state space. We will see in the next section how the boundary behavior can make the problem more complex and how the complete interference graph allows us to deal with reflections at the boundary: we prove in Lemma 4.14 that when starting with some null queues, all queues become positive for positive time and then use property of the limiting process to prove convergence. In the next section we will discuss an example where this is not the case: the line graph where some boundary issues arise in the analysis, and the time horizons for which we can prove convergence are bounded above.

4.2.4 Heuristic, line network

4.2.4.1 Homogenization objective

The rest of the section is dedicated to outlining the difficulties in proving a theorem such as 4.6. We will have to deal with boundary behaviors and potential reflections: the bounds we obtained for homogenization do not scale well when some but not all coordinates are large. These boundary behaviors are the most challenging part of this chapter: the homogenization bounds we obtain in the previous chapter may

fail but we are able to prove convergence nonetheless for a complete interference graph. In this section, we outline the difficulties arising in a line interference graph, as in Figure 2.2. Doing so, we will give insight as to the method we use to prove Theorem 4.6.

Consider a line interference graph with three nodes (Figure 2.2 with $n = 3$) and number the nodes successively on the line: 1 and 3 are called "outer nodes" and 2 the "central node". Because of Lemma 2.4, if all the queues start with an order of magnitude N, the service rate of the central node will remain at 0, at least up until the time one of the queue reaches 0 on the fluid scale. More formally, on the line graph, any solution of (4.2) with positive initial condition gives

$$\begin{cases} q^*(0) = q^0 \\ q_1^*(t) = q_1^0 + t(\lambda_1 - 1) \\ q_3^*(t) = q_3^0 + t(\lambda_3 - 1) \\ q_2^*(t) = q_2^0 + t\lambda_2 \end{cases},$$

until one of the coordinates reaches 0. This happens in finite time if $\min(\lambda_1, \lambda_3) < 1$. Indeed node 2 is not in an independent set of maximum size and $\bar{\pi}^q(|\sigma| < \Upsilon) = 0$. In this case, $\pi^q(\sigma)$ is given by

$$\pi^q(\sigma_v = 1) = \begin{cases} \dfrac{(q_1+1)^a(q_3+1)^a + (q_v+1)^a}{(q_1+1)^a(q_3+1)^a + (q_1+1)^a + (q_2+1)^a + (q_3+1)^a} & \text{if } v \in \{1,3\} \\ \dfrac{(q_2+1)^a}{(q_1+1)^a(q_3+1)^a + (q_1+1)^a + (q_2+1)^a + (q_3+1)^a} & \text{if } v = 2 \end{cases} \tag{4.5}$$

When all three nodes are "large" it is reasonable to expect $(Q_1(Nt), Q_2(Nt), Q_3(Nt))$ to reach a state with order of magnitude $(N^\gamma, N, N^{1-\gamma})$: if $Q_1(Nt)Q_3(Nt) \gg Q_2(Nt)$ node 2 does not receive any service, and if $Q_1(Nt)Q_3(Nt) \ll Q_2(Nt)$, nodes 1 and 3 do not receive any service. This type of order of magnitude ensures that both outer and central queues have a non-trivial homogenized service rate. It may be reasonable to expect that one queue reaches 0 before the other and we have a situation where $\gamma = 0$ with only one of the outer queue being of order N while the other is 1. We do not know how to handle transitions between order of magnitudes for the scaled process so we focus on the $\gamma = 1/2$ as a "most balanced" case. Assume that the localization set is such that each queue length remains of the same order of magnitude throughout the studied period. To be more precise, the localization set U^N for Q is of the form

$$U^N := \{q \in \mathbb{R}^V,\ \forall v \in V,\ \frac{q_v}{N^{\gamma_v}} \in (C_-, C_+)\}, \tag{4.6}$$

with $\gamma_1 = \gamma_3 = 1/2$ and $\gamma_2 = 1$. Say we choose a sequence of starting states $Q^{(N)}(0)$ such that $\frac{Q_v^{(N)}(0)}{\sqrt{N}} \to q_v^0$ for $v = 1, 3$ and $\frac{Q_2^{(N)}(0)}{N} \to q_2^0$, and define $\widetilde{Q}^N(t) = (\frac{Q_v^{(N)}(Nt)}{N^{\gamma_v} t})_{v=1,2,3}$. This gives us a scaled generator for $\widetilde{Q}^N$:

$$\widetilde{L}^\sigma_{\mathrm{s},N}(f)(q) := N \sum_{v \in V} \left[\lambda_v \left(f\left(q + \frac{e^v}{N^{\gamma_v}}\right) - f(q) \right) + \sigma_v \mathbb{1}_{q_v > 0} \left(f\left(q - \frac{e^v}{N^{\gamma_v}}\right) - f(q) \right) \right].$$

Corollary 3.7 is not enough to prove homogenization in this case:

Lemma 4.7

For any interference graph, let g be a function on $\mathbb{R}_+^V$ twice differentiable with bounded derivatives. Assume that $T > 1$ and

$$\ell^q \geqslant C \left\| q + 1 \right\|_\infty^{-a\beta}.$$

With the localization set defined in (4.6), for any N large enough,

$$\mathbb{E}\left[\sup_{t \leqslant T \wedge \bar{\tau}^{U^N}(Q)/N} \left| \int_0^t \left(\widetilde{L}_{\mathrm{s},N}^{\sigma^N(s)} - \widetilde{L}_{\mathrm{h},N} \right) [g](\widetilde{Q}^N(s)) \mathrm{d}s \right| \right] \leqslant CTN^{2a\beta - a/2 - 1} \log(N)^3 + CTN^{a\beta} \log(N)^{3/2}.$$

Proof. Similarly to Corollary 3.7, we apply Theorem 3.1 with g^N scaled in space defined by $g^N(q) = g((\frac{q_v}{N^{\gamma_v}})_{v \in V})$. Because of the scaling, we have $\max_{v \in V} \left\| \partial_v g^N \right\|_{\infty, U} \leqslant N^{-1/2}$, and $\frac{1}{\min_{q \in U^N, v \in V} q_v} \geqslant CN^{-1/2}$. Similarly, $\max_{v,w \in V} \left\| \partial^2_{v,w} g^N \right\|_{\infty, U} \leqslant N^{-1}$. □

Remark 4.8
It is possible to generalize the previous result: let us define a localization set

$$\widetilde{U}^N \coloneqq \{ q \in \mathbb{R}^V,\ \forall v \in V,\ \frac{q_v}{N^{\gamma_v}} \in (C_-, C_+) \}, \tag{4.7}$$

with $\gamma_v \in [0,1]$ and $\gamma = \min_v \gamma_v$. The closer to 0 queue lengths are, the less efficient our homogenization result becomes: for any interference graph, let g be a function on $\mathbb{R}_+^V$ twice differentiable with bounded derivatives, under the same spectral gap assumption as the previous lemma. With the localization set defined in (4.7), for any N large enough,

$$\mathbb{E}\left[\sup_{t \leqslant T \wedge \bar{\tau}^{\widetilde{U}^N}(Q)/N} \left| \int_0^t \left(\widetilde{L}_{\mathrm{s},N}^{\sigma^N(s)} - \widetilde{L}_{\mathrm{h},N} \right) (g)(Q^N(s)) \mathrm{d}s \right| \right] \leqslant CTN^{2a\beta - \gamma(2+a)} \log(N)^3 + CTN^{a\beta + \frac{1}{2} - \gamma} \log(N)^{3/2}.$$

Unfortunately, to have the convergence to 0 of (4.11) below, it is not possible to take $\gamma \leqslant 1/2$ because as soon as $a > 0$, $1/2 < 1/2 + a\beta$. In order for the right-hand side to vanish, we need $\gamma > \max(a\beta + 1/2, \frac{2a\beta}{2+a})$. For the complete interference graph, $\gamma = 1$ is sufficient so we just need $a\beta + \frac{1}{2} < 1$. In the case of the line $\gamma = 1/2$ is indeed the best bound we can obtain but it is not enough. In fact, we can prove convergence in Theorem 4.5 when queue start with an order of manitude N until they reach an order of magnitude $N^{\max(a\beta+1/2, \frac{2a\beta}{2+a})}$. A minimal condition for homogenization is $a\beta < 1$ in order to let the schedule homogenize before the queue lengths change significantly. Under this assumption, the leading term is $N^{a\beta+1/2}$ so the bound on the homogenization error we get with this result does not vanish as $N \to +\infty$.

4.2.4.2 Homogenized process

Assuming homogenization, we have some ideas about the behavior of the limit: whenever $Q^{(N)}(0) \in U^N$, looking at the limit for $\pi^{Q^{(N)}(0)}(\sigma_v = 1)$ we can expect that in some specific arrival configurations, the scaled process may keep this order of magnitude over finite time intervals. Let $q_v^N = q_v^0 \times N^{\gamma_v}$ with $\gamma_1 = \gamma_3 = 1/2$ and $\gamma_2 = 1$. From (4.5), by factorizing N^a, we get

$$\pi^{q^N}(\sigma_v = 1) = \frac{(q_1^0 q_3^0)^a}{(q_2^0)^a + (q_1^0 q_3^0)^a} + \frac{(q_v^0)^a}{N^{a/2}((q_2^0)^a + (q_1^0 q_3^0)^a)} + O(N^{-a}) \text{ if } v = 1, 3$$

or

$$\pi^{q^N}(\sigma_2 = 1) \approx \frac{(q_2^0)^a}{(q_2^0)^a + (q_1^0 q_3^0)^a} + O(N^{-a}).$$

The generator of the homogenized process is given by

$$\widetilde{L}_{\mathrm{h},N}[f](q) \approx N \sum_{v \in V} \left[\lambda_v \left(f\left(q + \frac{e^v}{N_v}\right) - f(q) \right) + \pi^{(N^{\gamma_w} q_w)_{w \in V}}(\sigma_v = 1) \left(f\left(q - \frac{e^v}{N_v}\right) - f(q) \right) \right].$$

We have a handy approximation for the homogenized generator for $q \in U^N$ for large N:

$$\widetilde{L}_{\mathrm{h},N}[f](q) \approx N \sum_{v \in V} \left[\lambda_v \left(f\left(q + \frac{e^v}{N_v}\right) - f(q) \right) + \frac{\mathbb{1}_{v=2} q_2^a + \mathbb{1}_{v \in (1,3)} (q_1 q_3)^a}{q_2^a + (q_1 q_3)^a} \left(f\left(q - \frac{e^v}{N_v}\right) - f(q) \right) \right].$$

For any $q \in \mathbb{R}_+^3$, let's define

$$Z^q \coloneqq q_2^a + (q_1 q_3)^a.$$

Using the martingale problem on the homogenized process $\widetilde{Q}^N$ of generator $\widetilde{L}_{\mathrm{h},N}$ and the approximation previously described for the invariant measure, we get

$$\begin{aligned}
\widetilde{Q}_1^N(t) &\approx q_1^0 + \sqrt{N} \int_0^t \left(\lambda_1 - \frac{(\widetilde{Q}_1^N(s) \widetilde{Q}_3^N(s))^a}{Z^{\widetilde{Q}^N(s)}} \right) \mathrm{d}s + o(\sqrt{N}) + M_1^N(t), \\
\widetilde{Q}_3^N(t) &\approx q_3^0 + \sqrt{N} \int_0^t \lambda_3 - \frac{(\widetilde{Q}_1^N(s) \widetilde{Q}_3^N(s))^a}{Z^{\widetilde{Q}^N(s)}} \mathrm{d}s + o(\sqrt{N}) + M_3^N(t), \text{ and} \\
\widetilde{Q}_2^N(t) &\approx q_2^0 + \int_0^t \left(\lambda_2 - \frac{(\widetilde{Q})_2^N(s))^a}{Z^{\widetilde{Q}^N(s)}} \right) \mathrm{d}s + M_2^N(t).
\end{aligned}$$

The martingale term in the dynamic of node 2 vanishes as $N \to +\infty$. Because of the $\sqrt{N}$ scaling in space for the outer queues, M_1^N and M_3^N do not vanish for large N but they remain of order of magnitude 1. For instance the *carré du champ* of M_1^N is given by

$$\Gamma_1^N(t) = \int_0^t \frac{\lambda_1^2 + \sigma_1(Ns)^2}{(\sqrt{N})^2} \mathrm{d}s = \lambda_1^2 t + \int_0^t \sigma_1(s) \mathrm{d}s.$$

Because of the $\sqrt{N}$ in front of the integral for the dynamic of node 1 and 3 tightness

is not obvious for $\widetilde{Q}^N$ as the process is not necessarily bounded. If nodes 1 and 3 have the same arrival rate

$$\lambda_0 \coloneqq \lambda_1 = \lambda_3,$$

nodes 1 and 3 have the same first order asymptotic dynamic. If $\widetilde{Q}^N$ is tight, we can even expect to reach an equilibrium state where

$$\lambda_0 - \frac{1}{1+(\frac{\widetilde{Q}_2^N(t)}{\widetilde{Q}_1^N(t)\widetilde{Q}_3^N(t)})^a} \approx \frac{1}{\sqrt{N}}$$

to ensure that the right hand side is bounded. Such a state space collapse would allow us to compare the dynamic of $\widetilde{Q}_2^N$ to the solution to an ODE and give us a candidate limiting process: if $\lambda_0 - \frac{1}{1+(\frac{\widetilde{Q}_2^N(t)}{\widetilde{Q}_1^N(t)\widetilde{Q}_3^N(t)})^a} \approx 0$, we have

$$\widetilde{Q}_1^N(t)\widetilde{Q}_3^N(t) \approx \widetilde{Q}_2^N(t)(\frac{\lambda_0}{1-\lambda_0})^{1/a}.$$

We can rewrite the dynamic of $\widetilde{Q}_2^N$:

$$\begin{aligned}\widetilde{Q}_2^N(t) &\approx q_2^0 + \int_0^t \lambda_2 - \frac{(\widetilde{Q}_2^N(s))^a}{(\widetilde{Q}_2^N(s))^a(1+\frac{\lambda_0}{1-\lambda_0})}\mathrm{d}s + M_2^N(t)\\ &\approx q_2^0 + t(\lambda_2+\lambda_0-1).\end{aligned}$$

If $\lambda_0+\lambda_2 \geqslant 1$ and $q_v^0 > 0$ for all v, this would ensure that it is impossible for any queue to reach 0 during any finite time interval if we assume that queues remain bounded. We could define a bounded and bounded away from 0 localization set U such that $\widetilde{Q}^N(t)$ remains in U over any finite time intervals in the super-critical regime. Since both of the outer queues have the same asymptotic dynamic, in the subcritical case, it seems possible that they do not reach 0 on the fluid scale at the same time if they have different starting points. Looking further at a second order approximation of the service rate, we see that in fact the distance between Q_1^N and Q_3^N has a negative drift component. Using the approximation of the invariant measure given at the begining of the section, we get that

$$\widetilde{Q}_1^N(t) - \widetilde{Q}_3^N(t) \approx q_1^0 - q_3^0 + N^{1/2(1-a)}\int_0^t \frac{\widetilde{Q}_3^N(s)^a - \widetilde{Q}_1^N(s)^a}{Z^{\widetilde{Q}^N(s)}}\mathrm{d}s + M_1^N(t) - M_3^N(t).$$

Since the martingales have an order of magnitude 1, this ensures that $\left|\widetilde{Q}_1^N - \widetilde{Q}_3^N\right|$ decreases extremely fast. If we managed to prove the homogenization result from Lemma 4.7, we would be able to have a result similar to Theorem 4.6 for a line interference graph. One of the problem is that the process $Q_1^N Q_3^N$ evolves much faster than Q_2^N because of their size. Queues that are of order $\sqrt{N}$ take a time of order $\sqrt{N}$ to evolve significantly.

4.3 Preliminary

We will begin this section by proving a tightness result for the process of queue lengths. Then, we will spend the next subsection defining the localization set and

proving the homogenization result as a direct consequence of Corollary 4.13 derivative of Corollary 3.7. Finally, we will prove the results stated in the preliminary section.

Recall that C denotes a numerical constant allowed to depend on a, n, λ, G and ϵ. We begin with a definition:

Definition 4.9
The martingale from 2.1 associated with (Q^N, σ^N) and $f(q,\sigma) = q_v$ is M_v^N.

$$\forall t \geqslant 0,\ M_v^N(t) = Q_v^N(t) - \int_0^t \left(\lambda_v - \sigma_v^N(s)\right) \mathrm{d}s$$

Having bounds on this martingale will prove to be useful so we provide one here.

Lemma 4.10
For any initial condition,

$$\mathbb{E}\left[\sup_{t\leqslant T}|M_v^N(t)|\right] \leqslant 2\sqrt{\frac{T(\max_v \lambda_v + 1)}{N}} \to 0 \text{ as } N \to +\infty. \tag{4.8}$$

Proof. Indeed,

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\leqslant T}|M_v^N(t)|\right] &\overset{(a)}{\leqslant} \sqrt{\mathbb{E}\left[\sup_{t\leqslant T} M_v^N(t)^2\right]} \\
&\overset{(b)}{\leqslant} 2\sqrt{\mathbb{E}\left[M_v^N(T)^2\right]} \\
&= 2\sqrt{\mathbb{E}\left[\langle M_v^N\rangle(T)\right]} \\
&\overset{(c)}{=} 2\sqrt{\mathbb{E}\left[N\int_0^T \frac{\lambda_v + \sigma_v^N(s)}{N^2}\mathtt{d}s\right]} \\
&\leqslant 2\sqrt{\frac{T(\max_v \lambda_v + 1)}{N}}.
\end{aligned}$$

Inequality (a) comes from Jensen's inequality, (b) comes from Doob's inequality and (c) comes from the definition of the *carré du champ*. □

Proposition 4.11
For any tight sequence initial condition $Q^N(0)$, Q^N is tight for the topology of uniform convergence over compact time sets. Any limiting point is a distribution on continuous functions.

Proof. To obtain this property, recall the definition of the modulus of continuity of

Q^N from Chapter 2 of [Bil99]: for any $\delta > 0$,

$$\begin{aligned}\omega_{Q^N}(\delta) &= \sup_{s,t\leqslant T,|t-s|\leqslant\delta,v\in V} |Q_v^N(t) - Q_v^N(t)| \\ &= \sup_{s,t\leqslant T,|t-s|\leqslant\delta,v\in V} |\int_s^t (\lambda_v - \sigma_v^N(u))\mathrm{d}u + M_v^N(t) - M_v^N(s)| \\ &\leqslant \delta \max_v(\lambda_v \vee |\lambda_v - 1|) + 2 \sup_{t\leqslant T,v\in V} |M_v^N(t)|\end{aligned}$$

To conclude the proof of this proposition, use Markov inequality on (4.8) to obtain:

$$\mathbb{P}\left(\omega_{Q^N}(\delta) \geqslant \delta(\max_v(\lambda_v \vee |\lambda_v - 1|) + \nu)\right) \to 0 \text{ as } N \to +\infty \text{ for any } \nu > 0, \quad (4.9)$$

which implies the result with a slight modification of Theorem 7.3 in [Bil99]. □

4.3.1 Localization and homogenization

Let us define a localization set of the form $U \subset \mathbb{R}_+^V$ such that

$$U \coloneqq \left\{q \in \mathbb{R}_+^V : \forall v \in V,\, q_v \in (C_-, C_+)\right\}, \quad (4.10)$$

Recall q^* the solution of (4.2) or (4.3) depending on the context, and

$$\tau^\epsilon(f) = \inf\left\{t > 0,\, \min_{v\in V} f_v(t) \leqslant \epsilon\right\}.$$

By continuity of q^*, for any $T < +\infty$, there exists a C_T such that

$$\sup_{t\leqslant T} \max_v q_v^*(t) < C_T.$$

For any fixed $q^0 \in (0, +\infty)^V$ and $0 < \epsilon < \min_v q_v^0$, let us fix T a finite horizon such that $T \leqslant \tau^\epsilon(q^*)$. To obtain convergence uniformly over compact sets contained in $[0, \tau^\epsilon(q^*))$ define U^ϵ in (4.10) with $C_- = \frac{\epsilon}{3}$ and $C_+ = 3C_{\tau^\epsilon(q^*)}$ (we emphasize the dependency in ϵ only here but omit it until needed in the rest of the chapter). Recall the exit time of Q^N from U^N:

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

Define also

$$T_{\text{tube}}^N \coloneqq T^{(\frac{\epsilon}{3}\wedge C_{\tau^\epsilon(q^*)})}(Q^N, q^*) = \inf\left\{t > 0,\, \left\|Q^N(t) - q^*(t)\right\|_\infty > \min(C_-, C_+/3)\right\}.$$

We present here a lemma used to remove the localization by replacing it with a "tube condition" around the limiting process. The behavior of the limiting process allows us to remove this tube condition as well in some cases.

Lemma 4.12

Almost surely $T^N_{\text{tube}} < \bar{\tau}^U(Q^N)$. *In particular,* $Q^N(t \wedge T^N_{\text{tube}}) \in U^N$ *for all* $t \leqslant T$.

Proof. Recall that by definition of the finite time horizon

$$\epsilon < \inf_{0 \leqslant t \leqslant T, v \in V} q^*_v(t) \leqslant \sup_{0 \leqslant t \leqslant T, v \in V} q^*_v(t) < C_{\tau^\epsilon(q^*)},$$

or equivalently

$$3C_- < \inf_{0 \leqslant t \leqslant T, v \in V} q^*_v(t) < \sup_{0 \leqslant t \leqslant T, v \in V} q^*_v(t) \leqslant C_+/3.$$

Since the jump size of Q^N is $\frac{1}{N}$ we have $\sup_{t \leqslant T^N_{\text{tube}}} \left\| Q^N(t) - q^*(t) \right\|_\infty \leqslant \min(C_-, C_+/3) + \frac{1}{N}$. In addition, for large N, $\frac{1}{N}$ is smaller than $\min(C_-, C_+/3)$. A direct consequence is that

$$3C_- - 2C_- < \inf_{0 \leqslant t \leqslant T^N_{\text{tube}} \wedge T} Q^N_v(t) \leqslant \sup_{0 \leqslant t \leqslant T^N_{\text{tube}} \wedge T} Q^N_v(t) \leqslant 3C_+/3.$$

Said otherwise, $T^N_{\text{tube}} < \bar{\tau}^U(Q^N)$ almost surely. □

Using the localization set from (4.10), we are able to state the homogenization result that is used in this chapter.

Corollary 4.13
Assume that

$$\ell^q \geqslant C \left\| q + 1 \right\|_\infty^{-a\beta}.$$

For any $v \in V$, $t \leqslant T$, *if* $a\beta < 1/2$ *we have*

$$\mathbb{E}\left[\sup_{0 \leqslant t \leqslant T \wedge \bar{\tau}^U(Q^N)} \left| \int_0^t \left(\sigma^N_v(s) - \pi^{NQ^N(s)}(v) \right) \mathrm{d}s \right| \right] \leqslant C N^{a\beta - 1/2} \log(N)^{3/2}. \quad (4.11)$$

Proof. Using Corollary 3.7, we just have to notice that for any $a > 0, \beta > 0$, $a\beta - \frac{1}{2} > a\beta - 1$. Notice also that $a\beta - \frac{1}{2} > a(2\beta - 1) - 1$ as long as $a\beta < \frac{1}{2}$. □

4.3.2 Proof of preliminary results

We begin by proving the preliminary results stated earlier in the section. First Lemma 4.4, then 4.3.

Proof of Lemma 4.4. Let $q^0 \in \mathbb{R}^V_+$. If $q^0 \in (0, +\infty)^V$, let $q^* = q^*(\cdot, q^0)$ solution to (4.2). If $q^0 \in \mathbb{R}^V_+ \setminus (0, +\infty)^V$, assume the existence of a solution (they necessarily exist for $q^0 \in \mathbb{R}^V_+ \setminus \{\mathbf{0}\}$, see the proof of Lemma 4.3 for more details). We use q^* to denote an arbitrary solution to (4.3). If $q^0 \neq \mathbf{0}$, the expression of $\tau^0(s_1 \circ q^*)$ is a direct consequence of the expression of the sum. For any $q \in \mathbb{N}^V$, there always exists $q_v \geqslant \frac{s_1(q)}{n}$ because otherwise $\frac{n s_1(q)}{n} < s_1(q)$ by summing over V. Hence, $s_a(q) \geqslant (\frac{s_1(q)}{n})^a$.

We now prove that $\tau^0(s_1 \circ q^*) = \tau^0(q^*)$ if $s_1(\lambda) > 1$. For any $q^0 \in \mathbb{R}_+^V$, $t > 0$ we get

$$\begin{aligned}(q_v^*)'(t) &= \lambda_v - \frac{(q_v^*(t))^a}{s_a(q^*(t))} \\ &\geqslant \lambda_v - \left(\frac{nq_v^*(t)}{s_1(q^*(t))}\right)^a \\ &= \lambda_v - \left(\frac{nq_v^*(t)}{(s_1(\lambda)-1)t}\right)^a.\end{aligned}$$

If $q_v^*(t) < (\frac{\lambda_v}{2})^{1/a}\frac{(s_1(\lambda)-1)}{n}t$ we get $(q_v^*)'(t) > \frac{\lambda_v}{2}$. If $\frac{\lambda_v}{2} < (\frac{\lambda_v}{2})^{1/a}\frac{(s_1(\lambda)-1)}{n}$, $q_v^*(t)$ cannot become smaller than $\frac{\lambda_v}{2}t$: if at some point $q_v^*(t) \leqslant (\frac{\lambda_v}{2})^{1/a}\frac{(s_1(\lambda)-1)}{n}t$, by continuity of q^*, for any $s > 0$,

$$q_v^*(t+s) \geqslant q_v^*(t) + \frac{\lambda_v}{2}s.$$

Similarly, If $\frac{\lambda_v}{2} \geqslant (\frac{\lambda_v}{2})^{1/a}\frac{(s_1(\lambda)-1)}{n}$, by continuity, $q_v^*(t)$ cannot become smaller than $(\frac{\lambda_v}{2})^{1/a}\frac{(s_1(\lambda)-1)}{n}t$ because as soon as $q_v^*(t) \leqslant (\frac{\lambda_v}{2})^{1/a}\frac{(s_1(\lambda)-1)}{n}t$, we get

$$(q_v^*)'(t) \geqslant \frac{\lambda_v}{2} \geqslant (\frac{\lambda_v}{2})^{1/a}\frac{(s_1(\lambda)-1)}{n}.$$

Thus for every $v \in V$, and $t \geqslant 0$,

$$q_v^*(t) \geqslant \min\left(\frac{\lambda_v}{2}, (\frac{\lambda_v}{2})^{1/a}\frac{(s_1(\lambda)-1)}{n}\right)t,$$

which implies $\tau^0(q^*) = \tau^0(s_1 \circ q^*) = +\infty$.

We next prove that $\tau^0(q^*) = \tau^0(s_1 \circ q^*)$ if $q^0 \neq \mathbf{0}$ and $s_1(\lambda) \leqslant 1$. The first step is done now: we prove that coordinates which started empty become positive for any small enough positive time. If $s_1(\lambda) = 1$, $s_1(q^*)$ is constant and thus always greater than $\frac{s_1(q^0)}{2}$. If $s_1(\lambda) < 1$, for $t \leqslant \frac{s_1(q^0)}{2|s_1(\lambda)-1|}$, we have $s_1(q^*(t)) \geqslant \frac{s_1(q^0)}{2}$. In both cases, there is $t^0 > 0$ and $\epsilon > 0$ such that $\inf s_1(q^*)(t) \geqslant \epsilon$ for any $t \leqslant t^0$. Moreover, for any $\epsilon > 0$, if $s_1(q^*(t)) \geqslant \epsilon$, we can give a lower bound to the derivative in time for solutions to (4.3). Using the definition of the ODE, and $s_a(q) \geqslant \left(\frac{s_1(q)}{n}\right)^a$, we get

$$\begin{aligned}(q_v^*)'(t) &= \lambda_v - \frac{(q_v^*(t))^a}{s_a(q^*(t))} \\ &\geqslant \lambda_v - \left(\frac{nq_v^*(t)}{s_1(q^*(t))}\right)^a \\ &\geqslant \lambda_v - \left(\frac{nq_v^*(t)}{\epsilon}\right)^a\end{aligned} \tag{4.12}$$

For any $t \geqslant 0$ if $s_1(\lambda) = 1$ or $t \leqslant \frac{s_1(q^0)}{2|s_1(\lambda)-1|}$ if $s_1(\lambda) < 1$, whenever $q_v^*(t) \leqslant \left(\frac{\lambda_v}{2}\right)^{1/a}\frac{s_1(q^0)}{2n}$, it follows that $(q_v^*)'(t) > \frac{\lambda_v}{2}$. So there is $t^0, \epsilon' > 0$ such that any v

with $q_v^0 \leqslant \epsilon'$ gets q_v^* increasing for $t \in (0, t^0)$. Even if some queues start null, they become positive for $t > 0$ small enough.

We now prove that if they do touch 0, all coordinates reach it simultaneously. It is only possible for queue lengths to touch 0 if $s_1(\lambda) < 1$. By (4.12), if $q_v^*(t) \leqslant \left(\frac{\lambda_v}{2}\right)^{1/a} \frac{s_1(q^*(t))}{n}$, we have $(q_v^*)'(t) \geqslant \frac{\lambda_v}{2}$. If $\tau^0(q^*) < \tau^0(s_1 \circ q^*)$, this would mean that $s_1(q^*)(\tau^0(q^*)) > 0$. Without loss in generality, assume that $q_v^*(t) \to 0$ as $t \to \tau^0(q^*)$. Since q_v^* is continuous and $q_v^*(t) \to 0$ as $t \to \tau^0(q^*)$, there exists a time interval $(t_-, \tau^0(q^*))$ such that $q_v^*(t) \leqslant \left(\frac{\lambda_v}{2}\right)^{1/a} \frac{s_1(q^*)(\tau^0(q^*))}{n}$ and q_v^* is decreasing. Since $s_1(\lambda) < 1$, $s_1 \circ q^*$ is decreasing so $s_1(q^*)(t) \geqslant s_1(q^*)(\tau^0(q^*))$ for any $t \in (t_-, \tau^0(q^*))$. This leads to a contradiction because $(q_v^*)'(t) \geqslant \frac{\lambda_v}{2}$ when $q_v^*(t) \leqslant \left(\frac{\lambda_v}{2}\right)^{1/a} \frac{s_1(q^*(t))}{n}$. Hence $\tau^0(q^*) = \tau^0(s_1 \circ q^*)$. □

We now prove Lemma 4.3.

Proof of Lemma 4.3. Let $q^N \in (0, +\infty)^V$ be such that $q^N \to q^0 \in \mathbb{R}_+^V$. Let the maximal solution to (4.2) on a complete interference graph with $\bar{q}^N(0) = q^N$ be denoted

$$\bar{q}^N : D^N \to (0, +\infty)^V.$$

For any N, there is $T_{\text{ext}}(q^N) = \sup D^N > 0$. Because of Lemma 4.4,

$$T_{\text{ext}}(q^N) = \tau^0(s_1 \circ \bar{q}^N).$$

Lemma 4.4 states that if $s_1(\lambda) > 1$, $T_{\text{ext}}(q^N) = +\infty$, if $s_1(\lambda) < 1$ $T_{\text{ext}}(q^N) = \frac{s_1(q^N)}{1 - s_1(\lambda)}$ and if $s_1(\lambda) = 1$, $T_{\text{ext}}(q^N) = +\infty$ but we will see that the limiting process may reach $\mathbf{0}$ if $q^N \to \mathbf{0}$.

If $s_1(\lambda) < 1$ and $q^0 = \mathbf{0}$, $\tau^0(s_1 \circ \bar{q}^N) \to 0$ as $N \to +\infty$ so $q^*(\cdot, q^N) \to \mathbf{0}$ uniformly over compact time sets.

If $s_1(\lambda) \geqslant 1$ or $q^0 \neq \mathbf{0}$, let us consider $D = (0, \liminf_N T_{\text{ext}}(q^N))$ and $\widetilde{q}^N : D \to \mathbb{R}_+^V$, the restriction of $\bar{q}^N$ on D. The first step is to prove convergence of $\widetilde{q}^N$ for the uniform convergence and then extend to the positive half real line. Because $T_{\text{ext}}(q^N) = \frac{s_1(q^N)}{1 - s_1(\lambda)}$ and $q^0 \neq \mathbf{0}$, we have $\liminf_N T_{\text{ext}}(q^N) > 0$, and thus there is $t^0 > 0$ such that $(0, t^0) \subset D$. Since $q^N \to q^0 \in \mathbb{R}_+^V$, $\|g\|_\infty \leqslant 1$ and $(\widetilde{q}^N)'(t) = g(\widetilde{q}^N)(t)$, $\widetilde{q}^N$ is relatively compact in $C(I, \mathbb{R}_+^V)$ for the convergence uniform over compact time sets by the Arzelà-Ascoli theorem (see Theorem 7.2 in Chapter 2 of [Bil99]). Indeed, for any $\delta > 0$,

$$\sup_{t,s \leqslant T, |t-s| \leqslant \delta} \|q_v^*(t) - q_v^*(s)\|_\infty \leqslant \|g\|_\infty \delta.$$

Let N_k be a subsequence such that $\widetilde{q}^{N_k} \to \bar{q}$ uniformly over compact sets of D. Note that compact sets of D are bounded away from 0. We will prove that $\bar{q}$ must be a solution to (4.3).

First we prove that $g(\widetilde{q}^N) \to g(\bar{q})$ uniformly over compact time sets of D. Fix a

time horizon $T < \sup I$ and $\nu > 0$. By definition of D,

$$\inf_{\nu \leqslant t \leqslant T, v \in V} \widetilde{q}_v^N(t) \geqslant \eta > 0.$$

First, g is Lipschitz continuous on $(\eta, C_T)^V$ because it has bounded derivative, so it is uniformly continuous. By definition of uniform continuity, for any $\epsilon > 0$, there exists $\delta > 0$ such that for any $q, q' \in (\eta, C_T)^V$ with $\|q - q'\|_\infty \leqslant \delta$,

$$\|g(q) - g(q')\|_\infty \leqslant \epsilon.$$

Let N_0 be such that $\sup_{t \leqslant T} \left\|\widetilde{q}^N(t) - \bar{q}(t)\right\|_\infty \leqslant \delta$ for any $N \geqslant N_0$. Then for any $N > N_0$,

$$\sup_{0 < t \leqslant T} \left\|g(\widetilde{q}^N)(t) - g(\bar{q}(t))\right\|_\infty \leqslant \epsilon,$$

i.e. $g(\widetilde{q}^N) \to g(\bar{q})$ uniformly over compact time sets of D.

Since $q^N \to q^0 \in \mathbb{R}_+^V$, $(\widetilde{q}^N)' = g(\widetilde{q}^N)$, $\widetilde{q}^N \to \bar{q}$ and $g(\widetilde{q}^N) \to g(\bar{q})$ uniformly over compact time sets of $(0, t^0)$, by Theorem 7.17 of [Rud76], $\bar{q}$ is differentiable and $(\widetilde{q}^N)' \to \bar{q}'$ uniformly over compact time sets of $(0, t^0)$. For any $0 < t < t^0$, by uniqueness of the limit $\bar{q}'(t) = g(\bar{q})(t)$, and thus $\bar{q}$ must be a continuous function such that for any $t \in (0, t^0)$ we have $\bar{q}'(t) = g(\bar{q})(t)$. This also proves the existence of solutions to (4.3). We only used Lemma 4.4 for solutions of (4.2) with positive initial condition so there is no logic loop.

We now prove uniqueness of solutions to (4.3). Let $\overline{q} : D_1 \to \mathbb{R}_+^V$ and $\widetilde{q} : D_2 \to \mathbb{R}_+^V$ be two solutions starting from q^0 at time $t = 0$. For $t \in D_1 \cap D_2$, the sum can be expressed as

$$\zeta(t) \coloneqq \max(s_1(q^0) + (s_1(\lambda) - 1)t, 0).$$

If $s_1(\lambda) \geqslant 1$ and $q^0 \neq \mathbf{0}$, ζ will never reach 0. If $s_1(\lambda) < 1$, it will in finite time.

Let $t_d \in D_1 \cap D_2$ be the first time that $\overline{q}(t) \neq \widetilde{q}(t)$. We first make the assumption that t_d is also the infimum of times such that $s_a(\overline{q}(t)) > s_a(\widetilde{q}(t))$ and deal with the complementary later. By continuity there is $t_d > 0$ and $\delta > 0$ such that $\overline{q}(s) \neq \widetilde{q}(s)$ and $s_a(\overline{q}(s)) > s_a(\widetilde{q}(s))$for any $s \in (t_d, t_d + \delta) \subset D_1 \cap D_2$.

For any v such that $\overline{q}_v(s) < \widetilde{q}_v(s)$, we have $\frac{\overline{q}_v(s)}{s_a(\overline{q}(s))} < \frac{\widetilde{q}_v(s)}{s_a(\widetilde{q}(s))}$ and thus,

$$\overline{q}_v(s) - \widetilde{q}_v(s) < 0 \Rightarrow (\overline{q}'_v(s) - \widetilde{q}'_v(s)) > 0.$$

By continuity of the solutions, since $\bar{q}(t_d) = \widetilde{q}(t_d)$, $s_a(\overline{q}(s)) > s_a(\widetilde{q}(s))$ implies $\overline{q}_v(s) \geqslant \widetilde{q}_v(s)$ for all $v \in V$, for all $s \in (t_d, t_d + \delta)$ because as soon as $\bar{q}_v(t)$ would become smaller than $\widetilde{q}_v(t)$, $\bar{q}_v(t)$ would become increasing. Since $s_1(\overline{q}) = s_1(\widetilde{q})$ the only possibility is that $\overline{q}_v(s) = \widetilde{q}_v(s)$ for all $v \in V$ which contradicts $s_a(\overline{q}(s)) > s_a(\widetilde{q}(s))$ for any $s \in (t_d, t_d + \delta)$. The first time $\widetilde{q}$ and $\bar{q}$ are different is different from the first time $s_a \circ \widetilde{q}$ and $s_a \circ \bar{q}$ are different.

Let us introduce for any $t \in [t_d, t_d + \delta)$,

$$\zeta_a(t) \coloneqq s_a(\overline{q}(t)) = s_a(\widetilde{q}(t)).$$

We can rewrite (4.3) as

$$\begin{cases} f'(t) & = \lambda - \dfrac{f(t)^a}{\zeta_a(t)} \text{ if } \zeta_a(t) > 0 \\ s_1(f) & = \zeta \\ f(0) & = q^0 \end{cases} \tag{4.13}$$

For any $v \in V$, and $t > 0$, such that $\overline{q}_v(t) - \widetilde{q}_v(t) \geqslant 0$ we have

$$\overline{q}'_v(t) - \widetilde{q}'_v(t) = \frac{\widetilde{q}_v(t)^a - \overline{q}_v(t)^a}{\zeta_a(t)} \leqslant 0 \text{ because } a > 0.$$

Once again by continuity of the solutions, this implies $\overline{q}_v(t) \leqslant \widetilde{q}_v(t)$ for any $t \in D_1 \cap D_2$ and equality since $s_1(\overline{q}) = s_1(\widetilde{q})$. Thus there is a unique maximal solution defined up to the exit time of $(0, +\infty)^V$. By Lemma 4.4,

$$\lim_{t \to T_{\text{ext}(q^0)}} \bar{q}(t) = \lim_{t \to T_{\text{ext}(q^0)}} \widetilde{q}(t) = \mathbf{0}.$$

We extend the solution as in Definition 4.2 to obtain convergence on $\mathbb{R}_+$. □

4.4 General interference graph up to $\tau^0(q^*)$, Theorem 4.5

Recall the definitions:

$$U \coloneqq \left\{ q \in \mathbb{R}_+^V : \forall v \in V,\, q_v \in (C_-, C_+) \right\},$$

and

$$T^N_{\text{tube}} = T^{(\frac{\epsilon}{3} \wedge C_{\tau^\epsilon(q^*)})}(Q^N, q^*) = \inf \left\{ t > 0,\ \left\| Q^N(t) - q^*(t) \right\|_\infty > \min(C_-, C_+/3) \right\},$$

with $C_- > 0$ and $C_+ < +\infty$. To prove Theorem 4.5, we will establish its equivalent for the stopped process $Q^N(\,\cdot \wedge T^N_{\text{tube}} \wedge T)$ using Gronwall's lemma. We then transfer the result on the stopped process to $Q^N(\cdot \wedge T)$ using Lemma 4.12. This gives us uniform convergence over compact sets of $[0, \tau^\epsilon(q^*))$ for any $\epsilon > 0$. We also give the arguments to extend up to $\tau^0(q^*)$ at the end of the proof.

Proof of Theorem 4.5. Recall that for any $t \leqslant T \leqslant \tau^\epsilon(q^*)$,

$$q^*(t) = q^0 + \int_0^t \left(\lambda - \overline{\pi}^{q^*(s)} \right) \mathrm{d}s.$$

Using the martingale problem and adding/subtracting

$$\int_0^t \left(\pi^{NQ^N(s)}(\sigma_v = 1) + \bar{\pi}^{Q^N(s)}(v)\right) \mathrm{d}s,$$

we get

$$Q_v^N(t) - q_v^*(t) = Q_v^N(0) - q_v^0 + \int_0^t \left(\overline{\pi}^{q^*(s)}(v) - \overline{\pi}^{Q^N(s)}(v)\right) \mathrm{d}s + M_v^N(t)$$
$$+ \int_0^t \left(\overline{\pi}^{Q^N(s)}(v) - \pi^{NQ^N(s)}(\sigma_v = 1) + \pi^{NQ^N(s)}(\sigma_v = 1) - \mathbb{1}_{Q_v^N(s)>0}\sigma_v^N(s)\right) \mathrm{d}s.$$

Define

$$\Theta_v^N = \left|Q_v^N(0) - q_v^0\right| + \eta_v^N(T) + h_v^N(T) + \mu_v^N(T),$$

were

$$\eta_v^N(T) = \sup_{t \leqslant T} \left| \int_0^{t \wedge T_{\text{tube}}^N} \left(\overline{\pi}^{Q^N(s)}(v) - \pi^{NQ^N(s)}(\sigma_v = 1)\right) \mathrm{d}s \right|,$$

$$h_v^N(T) = \sup_{t \leqslant T} \left| \int_0^{t \wedge T_{\text{tube}}^N} \left((\pi^{NQ^N(s)}(\sigma_v = 1) - \mathbb{1}_{Q_v^N(s)>0}\sigma_v^N(s)\right) \mathrm{d}s \right|$$

and

$$\mu_v^N(T) = \sup_{t \leqslant T \wedge T_{\text{tube}}^N} \left|M_v^N(t)\right|.$$

We get for all $v \in V$,

$$\left|Q_v^N(t \wedge T_{\text{tube}}^N) - q_v^*(t \wedge T_{\text{tube}}^N)\right| \leqslant \Theta_v^N + \int_0^{t \wedge T_{\text{tube}}^N} \left|\overline{\pi}^{q^*(s)}(v) - \overline{\pi}^{Q^N(s)}(v)\right| \mathrm{d}s.$$

Remark $f : U \to \mathbb{R}^V$ defined by $f(q) = \overline{\pi}^q$ is Lipschitz. So there exists $C < +\infty$ such that for any $q, q' \in U$

$$\|f(q) - f(q')\| \leqslant C \|q - q'\|,$$

and thus,

$$\int_0^t \left|\overline{\pi}^{q^*(s)}(v) - \overline{\pi}^{Q^N(s)}(v)\right| \mathrm{d}s \leqslant C \int_0^t \left\|Q^N(s) - q^*(s)\right\| \mathrm{d}s.$$

To sum up, for any $t \leqslant T \wedge T_{\text{tube}}^N$, by summing over v, and denoting $\Theta^N = \sum_{v \in V} \Theta_v^N$,

$$\left\|Q^N(t) - q^*(t)\right\|_1 \leqslant \Theta^N + C \int_0^t \left\|Q^N(s) - q^*(s)\right\|_1 \mathrm{d}s.$$

By Gronwall lemma ([EK86], Appendix 5), for any $t \leqslant T \wedge T^N_{\text{tube}}$, we get

$$\left\|Q^N(t) - q^*(t)\right\|_1 \leqslant \Theta^N \exp(Ct) \leqslant \Theta^N \exp(CT).$$

By Lemma 2.4 for all $v \in V$, $\mathbb{E}\left[\eta^N_v(T)\right] \to 0$ as $N \to +\infty$. Further Corollary 4.13 states that $\mathbb{E}\left[h^N_v(T)\right] \to 0$ as $N \to +\infty$ if $a\beta < 1/2$. Every μ^N_v converges to 0 in the mean, by Lemma (4.10). Finally, $Q^N_v(0) - q^0_v \to 0$ as $N \to +\infty$. Thus $\mathbb{E}\left[\Theta^N\right] \to 0$ as $N \to +\infty$. To conclude the first part of this proof, we can state that for any $q^0 \in (0, +\infty)^V$, $\epsilon < \min_v q^0_v$, $T \leqslant \tau^\epsilon(q^*)$ and $\delta > 0$,

$$\mathbb{E}\left[\sup_{0 \leq t \leq T \wedge T^N_{\text{tube}}} \left\|Q^N(t) - q^*(t)\right\|\right] \to 0 \text{ as } N \to +\infty.$$

Let us now remove the localization and prove that $Q^N \xrightarrow{\mathbb{P}} q^*$ uniformly on $[0, T]$. In order to do so, it is enough to show that $\mathbb{P}(T^N_{\text{tube}} \geq T) \to 1$: indeed for any $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\left(\sup_{t \leqslant T} \left\|Q^N(t) - q^*(t)\right\|_\infty \geqslant \delta\right) &\leqslant \mathbb{P}\left(\sup_{t \leqslant T \wedge T^N_{\text{tube}}} \left\|Q^N(t) - q^*(t)\right\|_\infty \geqslant \delta, T^N_{\text{tube}} \geqslant T\right) \\
&\quad + \mathbb{P}\left(T^N_{\text{tube}} < T\right) \\
&\leqslant \mathbb{P}\left(\sup_{t \leqslant T \wedge T^N_{\text{tube}}} \left\|Q^N(t) - q^*(t)\right\|_\infty \geqslant \delta\right) + \mathbb{P}\left(T^N_{\text{tube}} < T\right).
\end{aligned}$$

We already proved that the first term on the right hand side converges to 0, $\mathbb{P}(T^N_{\text{tube}} \geq T) \to 1$ would mean that the second term also vanishes as $N \to +\infty$. By definition of T^N_{tube}, we have

$$\left\|Q^N(T^N_{\text{tube}}) - q^*(T^N_{\text{tube}})\right\|_1 \mathbb{1}_{T^N_{\text{tube}} < +\infty} \geq \min(C_-, C_+/3) \mathbb{1}_{T^N_{\text{tube}} < +\infty}.$$

Since $T^N_{\text{tube}} \wedge T = T^N_{\text{tube}}$ on the event $\{T^N_{\text{tube}} \leq T\}$, this entails

$$\mathbb{P}\left(T^N_{\text{tube}} \leq T\right) = \mathbb{P}\left(\left\|Q^N(T^N_{\text{tube}} \wedge T) - q^*(T^N_{\text{tube}} \wedge T)\right\|_1 \geq \min(C_-, C_+/3),\, T^N_{\text{tube}} \leq T\right).$$

Since we have proved that $Q^N(\,\cdot \wedge T^N_{\text{tube}}) \xrightarrow{\mathbb{P}} q^*$ uniformly on $[0, T]$ for any $T \leqslant \tau^\epsilon(q^*)$, the previous probability vanishes.

To sum up, for any $\epsilon, \delta > 0$, for any $T \leqslant \tau^\epsilon(q^*)$,

$$\mathbb{P}\left(\sup_{0 \leqslant t \leqslant T} \left\|Q^N(t) - q^*(t)\right\|_\infty \geqslant \delta\right) \to 0 \text{ as } N \to +\infty.$$

By continuity of q^*, since $\min_v q^0_v > 0$, we have

$$\lim_{\epsilon \to 0} \tau^\epsilon(q^*) = \tau^0(q^*)$$

and

$$q^*(\tau^\epsilon(q^*)) \to q^*(\tau^0(q^*)) \text{ as } \epsilon \to 0.$$

For any $T < \tau^0(q^*)$, there is $\epsilon > 0$ such that $T < \tau^\epsilon(q^*)$, and thus $Q^N \to q^*$ uniformly over compact time sets of $[0, \tau^0(q^*))$. □

4.5 Complete interference graph

This section only deals with a complete interference graph so we omit dependency in the graph in all of the statements. Recall the definitions:

$$\tau^{\epsilon}(s_1 \circ Q^N) = \inf\{t > 0,\ s_1(Q^N(t)) \leqslant \epsilon\},$$

and

$$\tau^{\epsilon}(s_1 \circ q^*) = \inf\{t > 0,\ s_1(q^*)(t) \leqslant \epsilon\}.$$

Introduce as well

$$T_{-}^{\epsilon}(f) \coloneqq \inf\{t > 0,\ \min_v f_v(t) \geqslant \epsilon\}.$$

We will prove that queue lengths become positive for small positive times and then identify the limit as the solution to (4.3) using Theorem 4.5 up to $T_{\text{ext}}(q^0)$. This is where we have to distinguish three cases:

- (I) $s_1(\lambda) < 1$,
- (II) $s_1(\lambda) > 1$,
- (III) $s_1(\lambda) = 1$.

In (I) we will show that the sum of queue lengths stays absorbed at 0 when it reaches it. In case (II) we will show that queue lengths converge to $+\infty$ linearly in all coordinates. For case (III) we will rely on the two previous cases: if $q^0 = \mathbf{0}$ the process stays absorbed at 0, if $q^0 \neq \mathbf{0}$ all coordinates are positive in positive time and they do not reach 0 in finite time. In the next section, we will prove Theorem 4.6.

4.5.1 Proof of Theorem 4.6

Theorem 4.6 strengthens Theorem 4.5 to include any initial condition and stop at any finite time horizon. The proof will rely on Theorem 4.5. We then handle the behavior when some queues are small using a lemma roughly stating that it takes a time very small for the limit of Q^N to cross small thresholds:

Lemma 4.14
If $q^0 \neq \mathbf{0}$ or $\sum_{v\in V} \lambda_v > 1$, for any $\epsilon > 0$ small enough, there is $t_\epsilon > 0$, such that

$$\mathbb{P}_{q^0}\left(T_{-}^{\epsilon}(Q^N) \geqslant t_\epsilon\right) \to 0 \text{ as } N \to +\infty.$$

In addition, $t_\epsilon \to 0$ as $\epsilon \to 0$.

Proof stub: The proof is based on an induction argument: we will recursively add nodes in a non-decreasing subset of nodes with bounded below queue lengths. To do that, for any $c > 0$ we will define a succession of thresholds ϵ_c^v, stopping times $\varphi_v^N(c)$ and time horizons t_c^v such that the set of nodes

$$J_i^N \coloneqq \{v \in V,\ Q_v^N(\varphi_i^N(c)) > \epsilon_c^i\}$$

is increasing with i and $\varphi_v^N(c) \leqslant t_c^v$ almost surely. At step k of the induction argument, we will prove that the probability of $\{k \in J_k^N\}$ goes to 1 as $N \to +\infty$. The thresholds are defined in a way such that if v is not in $J_{v-1}^N(c)$, Q_v^N increases on the mean, at least until the next threshold. On the other hand, once $v \in J_k^N$, we choose the time intervals such that for any $k' \geqslant k$, $v \in J_{k'}^N$ with high probability. Because the proof is quite technical for such an intuitive result and the proof does not give major insights we refer the interested reader to Appendix B for a proof of this result. □

Given this lemma, the proof of Theorem 4.6 is straightforward using tightness and uniqueness of solutions of (4.3):

Proof of Theorem 4.6: By Proposition 4.11, Q^N is tight and any limiting point is a continuous function. Let N_k be a subsequence such that $Q^{N_k} \Rightarrow \bar{q}$ as $k \to +\infty$ and $Q^{N_k}(0) \to q^0 \in [0, +\infty)^V$. We know that $\bar{q}$ must be almost surely continuous by Proposition 4.11. Let's sum up the proof method for this theorem.

- In case (I) the queue process reaches $\mathbf{0}$ in finite time and stays absorbed.
- In case (II), after some time all queues are positive and the dynamic evolves as in Theorem 4.5. Starting from an initial condition positive at each node, Theorem 4.5 is enough to obtain a convergence uniform over compact sets.
- The case (III) can be handled similarly to case (I) or (II) depending on the initial state: the variation of the sum of queue lengths is only due to the fraction of the time no queue is active and thus it takes a long time to evolve. On the fluid scale, the scaled sum remains constant. If $s_1(q^0) = 0$, all components remain at 0 as in case (I). Starting from a non trivial state, the queue lengths will evolve according to the ODE given by (4.3). If $s_1(q^0) \neq 0$, even if some queues start empty, all coordinates will be positive for positive time. Like for case (II), $\tau^0(q^*) = +\infty$. See Section 5.1.1 for more details on the critical case.

We first prove convergence of Q^N up to $T_{\text{ext}}(q^0)$ starting from any initial condition. If $q_v^0 > 0$ for all $v \in V$, this is Theorem 4.5 so we assume $\min_v q_v^0 = 0$ for the first part of this proof. Since we already know that Q^N is tight and any limiting point is in the space of continuous functions, we simply need to uniquely identify any potential limit.

First, $\min_v \bar{q_v}$ is a continuous function almost surely because $\bar{q}$ is as well. Let us call $\bar{q}^{-1}$ the generalized inverse of $\min_v \bar{q_v}$:

$$\bar{q}^{-1}(t) \coloneqq \inf\{s \geqslant 0,\ \min_v \bar{q_v}(s) \geqslant t\}.$$

By Lemma 2.10 in Chapter 6 of [JS03], $\bar{q}^{-1}$ is almost surely càg and the set

$$\mathcal{J} \coloneqq \{t \geqslant 0,\ \mathbb{P}\left(\bar{q}^{-1} \text{ is not continuous in } t\right) > 0\},$$

of fixed time discontinuities of $\bar{q}^{-1}$ is countable. Let $(\epsilon_p)_{p\in\mathbb{N}} \in \mathcal{J}^c$ be a sequence of real number converging to 0. By definition of $\mathcal{J}$, $\bar{q}^{-1}$ is almost surely continuous at ϵ_p for every $p \in \mathbb{N}$ and $\epsilon_p \to 0$ as $p \to +\infty$. Said otherwise,

$$\mathbb{P}\left(\bar{q}^{-1} \text{continuous at } \epsilon_p\right) = 1$$

for all $p \in \mathbb{N}$. Consider T_-^ϵ as an operator on càdlàg functions. By Proposition 2.11 of Chapter 6 in [JS03], $T_-^{\epsilon_p}$ is almost surely continuous at $\bar{q}$ because by definition of ϵ_p, $\bar{q}^{-1}$ is continuous at ϵ_p. By the continuous mapping theorem (Theorem 2.7 from [Bil99]), since the limit of Q^N is $\bar{q}$ and it is continuous and $T_-^{\epsilon_p}$ is almost surely continuous at $\bar{q}$, for any $p \in \mathbb{N}$,

$$(Q^{N_k}, T_-^{\epsilon_p}(Q^N)) \Rightarrow (\bar{q}, T_-^{\epsilon_p}(\bar{q})).$$

Using the joint convergence from the continuous mapping theorem, for any $p \in \mathbb{N}$,

$$Q^{N_k}(T_-^{\epsilon_p}(Q^{N_k}) + \cdot) \Rightarrow \bar{q}(T_-^{\epsilon_p}(\bar{q}) + \cdot) \text{ as } k \to +\infty.$$

Next, we explain how shifting the trajectory with the stopping time $T_-^{\epsilon_p}(Q^{N_k})$ gives another description of the limit using Theorem 4.5. By definition of $T_-^{\epsilon_p}(Q^{N_k})$,

$$Q^{N_k}(T_-^{\epsilon_p}(Q^{N_k})) \geqslant \epsilon_p.$$

We get $Q^{N_k}(T_-^{\epsilon_p}(Q^{N_k})) \Rightarrow \bar{q}(T_-^{\epsilon_p}(\bar{q}))$, with $\bar{q}_v(T_-^{\epsilon_p}(\bar{q})) \geqslant \epsilon_p$. By Theorem 4.5, and the strong Markov property, as $k \to +\infty$,

$$Q^{N_k}(T_-^{\epsilon_p}(Q^{N_k} + \cdot)) \Rightarrow q^*(\cdot, \bar{q}(T_-^{\epsilon_p}(\bar{q}))).$$

By uniqueness of the limit, it is necessary to have

$$\bar{q}(T_-^{\epsilon_p}(\bar{q}) + \cdot) = q^*(\cdot, \bar{q}(T_-^{\epsilon_p}(\bar{q}))).$$

By Lemma 4.14, for any $p > 0$, there is t_p such that

$$\mathbb{P}\left(T_-^{\epsilon_p}(\bar{q}) > t_p\right) = 0,$$

with $t_p \to 0$ as $p \to +\infty$. Since $T_-^{\epsilon_p}(\bar{q}) \leqslant t_p$ almost surely and $t_p \to 0$ as $p \to +\infty$, $T_-^{\epsilon_p}(\bar{q}) \to 0$ almost surely, and by continuity of $\bar{q}$, $\bar{q}(T_-^{\epsilon_p}(\bar{q})) \Rightarrow q^0$ as $p \to +\infty$.

In order to use uniform convergence to exchange the limits in k and p, we use Skorohod representation theorem (see for instance Theorem 6.7 of [Bil99]). It states that since $\mathbb{R}_+^V$ is separable, and $Q^{N_k} \Rightarrow \bar{q}$ uniformly over compact sets, it is possible to construct a probability space such that $Q^{N_k} \to \bar{q}$ almost surely for the topology of uniform convergence on compact sets. Until the end of the proof, consider Q^{N_k} and $\bar{q}$ constructed in such a way.

To conclude, uniformly over compact sets of $[0, T_{\text{ext}}(q^0))$: for any p and N greater than 0,

$$\begin{aligned}\sup_{0\leq t\leq T} \left\|Q^{N_k}(t) - q^*(t, q^0)\right\|_\infty \leqslant & \sup_{0\leq t\leq T} \left\|Q^{N_k}(t) - Q^{N_k}(t + T^{\epsilon_p}(Q^N))\right\|_\infty \\ & + \sup_{0\leq t\leq T} \left\|Q^{N_k}(t + T_-^{\epsilon_p}(Q^{N_k})) - q^*(t, \bar{q}(T_-^{\epsilon_p}(\bar{q})))\right\|_\infty \\ & + \sup_{0\leq t\leq T} \left\|q^*(t, \bar{q}(T_-^{\epsilon_p}(\bar{q}))) - q^*(t, q^0)\right\|_\infty .\end{aligned}$$

By letting $N \to +\infty$ then $p \to +\infty$, we get

$$Q^{N_k} \xrightarrow{\mathbb{P}} q^*(\cdot, q^0).$$

In cases where $T_{\text{ext}}(q^0) = +\infty$, Theorem 4.6 is proved. This regroups the case

where $s_1(\lambda) > 1$ and the case where $s_1(\lambda) = 1$ and $q^0 \neq \mathbf{0}$.

We now focus on the case $s_1(\lambda) < 1$. The first part of the proof stated that starting from any initial condition, the scaled sum of queue lengths will reach 0 in finite time ($\tau^0(q^*) < +\infty$). We will now prove that when $Q^N(0) \to \mathbf{0}$, the limit of the sum of queue lengths remains absorbed at $\mathbf{0}$. Recall the definition of the stopping times $T_-^\epsilon(s_1 \circ Q^N)$ given by

$$T_-^\epsilon(s_1 \circ Q^N) \coloneqq \inf\{t > 0,\ s_1(Q^N(t)) \geqslant \epsilon\}.$$

Here we emphasize the dependency in the trajectory but to ease notations later, we will only emphasize in the computation when it depends on the trajectory shifted after some time. Introduce the set $R^N(\epsilon)$ by

$$R^N(\epsilon) \coloneqq \{q \in \frac{1}{N}\mathbb{N}^V,\ s_1(q) \in [\frac{\epsilon}{2}; \frac{\epsilon}{2} + \frac{1}{N})\}.$$

Note that for any $\epsilon > 0$, $R^N(\epsilon)$ is a finite set. For $q \in \mathbb{R}_+^V$, the notation $\mathbb{P}_q$ denotes $\mathbb{P}$ conditionally on $Q^N(0) = q$. Let's fix $\epsilon > 0$ and call

$$f^N(t) \coloneqq \sup_{q \in R^N(\epsilon)} \mathbb{P}_q\left(T_-^\epsilon \leqslant t\right).$$

Since $R^N(\epsilon)$ is a finite set, there exists $q^N(t) \in R^N(\epsilon)$ such that $f^N(t) = \mathbb{P}_{q^N(t)}\left(T_-^\epsilon \leqslant t\right)$.

Fix $t < +\infty$ and $\epsilon > 0$, assume $Q^N(0) \to \mathbf{0}$ as $N \to +\infty$. We will prove that for $\epsilon > 0$, the probability that the sum exceeds ϵ in finite time goes to 0 as $N \to +\infty$. The argument is essentially the same as with a deterministic system: to reach a level ϵ, the sum should first reach $\epsilon/2$ and then be increasing between $\epsilon/2$ and ϵ which is not possible. For any ν probability measure, we use $\mathbb{P}_\nu$ to denote the probability measure $\sum_{q \in \mathbb{N}^V} \nu(q)\mathbb{P}\left(\cdot \mid Q(0) = q\right)$, if $\nu = \delta_q$ a Dirac measure, $\mathbb{P}_\nu = \mathbb{P}_q$. We get

$$\begin{aligned}
\mathbb{P}_{Q^N(0)}\left(T_-^\epsilon \leqslant t\right) &= \mathbb{P}_{Q^N(0)}\left(T_-^{\epsilon/2} \leqslant t, T_-^\epsilon(s_1 \circ (Q^N(\cdot + T_-^{\epsilon/2})) \leqslant t - T_-^{\epsilon/2}\right) \\
&\leqslant \mathbb{P}_{Q^N(0)}\left(T_-^{\epsilon/2} \leqslant t, T_-^\epsilon(s_1 \circ (Q^N(\cdot + T_-^{\epsilon/2})) \leqslant t\right) \\
&= \mathbb{E}_{Q^N(0)}\left[\mathbb{1}_{T_-^{\epsilon/2} \leqslant t} \mathbb{P}_{Q^N(T_-^\epsilon)}\left(T_-^{\epsilon/2} \leqslant t\right)\right], \qquad (4.14)
\end{aligned}$$

where the last equality is obtained using the strong Markov property. As it turns out, since jump size is $\frac{1}{N}$, almost surely on $\{T_-^{\epsilon/2} \leqslant t\}$, we have $Q^N(T_-^{\epsilon/2}) \in R^N(\epsilon)$. For this reason,

$$\mathbb{P}_{Q^N(0)}\left(T_-^\epsilon \leqslant t\right) \leqslant \mathbb{1}_{T_N^{c,\epsilon/2} \leqslant t} \mathbb{P}_{Q^N(T_N^{c,\epsilon/2})}\left(T_-^\epsilon \leqslant t\right) \leqslant f^N(t). \qquad (4.15)$$

We now prove that $f^N(t) \to 0$ for all t. Since f^N is bounded by 1, for any t $f^N(t)$ is tight as a sequence of real numbers and we only need to prove uniqueness

of the limit in terms of subsequence. For any t, there exists a subsequence such that $q^{N_k}(t) \to \widetilde{q}(t)$ because they too are tights. Any $\widetilde{q}(t)$ obtained in that way has $s_1(\widetilde{q}(t)) = \frac{\epsilon}{2}$ because $q^N(t) \in R^N(\epsilon)$.
By the first part of the proof, starting from any $q^N \in R^N(\epsilon)$ bounded away from $\mathbf{0}$, the sum of queue lengths should decrease to 0 before being able to reach ϵ. Necessarily, $\tau^0(s_1 \circ Q^N) \xrightarrow{\mathbb{P}} \frac{\epsilon}{2(1-\sum_{v\in V}\lambda_v)}$ by the first part of the proof. Define $t^* := \frac{\epsilon}{2(1-\sum_{v\in V}\lambda_v)}$. Assume that $t \leqslant t^*$, we get

$$\begin{aligned} f^N(t) &= \mathbb{P}_{q^N(t)}\left(T^\epsilon_- \leqslant t\right) \\ &= \mathbb{P}_{q^N(t)}\left(T^\epsilon_- \leqslant t, T^\epsilon_- \leqslant t^*\right) + \mathbb{P}_{q^N(t)}\left(T^\epsilon_- \leqslant t, T^\epsilon_- > t^*\right) \\ &\leqslant \mathbb{P}_{q^N(t)}\left(T^\epsilon_- \leqslant t^*\right) + \mathbb{P}_{q^N(t)}\left(T^\epsilon_-(s_1 \circ Q^N(\cdot + t^*)) \leqslant t - t^*\right) \end{aligned}$$

Similarly, by the Markov property,

$$\mathbb{P}_{q^N(t)}\left(T^\epsilon_-(Q^N(.+t^*)) \leqslant t - t^*\right) = \mathbb{E}_{q^N(t)}\left[\mathbb{P}_{Q^N(t^*)}\left(T^\epsilon_- \leqslant t - t^*\right)\right]$$

Finally, since for any sequence $q^N \in R^N(\epsilon)$, under $\mathbb{P}_{q^N}$, $Q^N(t^*) \xrightarrow{\mathbb{P}} \mathbf{0}$, we can apply the same reasoning as in (4.14) to prove that

$$\mathbb{P}_{Q^N(t^*)}\left(T^\epsilon_- \leqslant t - t^*\right) \leqslant f^N(t - t^*)\mathbb{1}_{t \geqslant t^*}$$

To sum this up, $f^N(t) \leqslant f^N(t^*) + f^N(t - t^*)\mathbb{1}_{t\geqslant t^*}$, and iterating, we get

$$f^N(t) \leqslant \lceil \frac{t}{t^*} \rceil f^N(t^*).$$

We can conclude the proof by using convergence up to $\tau^0(q^*)$ to prove that $f^N(t^*) \to 0$ as $N \to +\infty$: $f^N(t^*) = \mathbb{P}_{q^N(t^*)}\left(T^\epsilon_-(s_1 \circ Q^N) \leqslant t^*\right)$ but since $q^N(t^*) \in R^N(\epsilon)$, $T_{\text{ext}}(q^N(t^*)) \to t^*$ and the limiting process will reach $\mathbf{0}$ at t^* because $s_1(q^N(t^*)) \to \frac{\epsilon}{2}$.

□

Chapter 5

Heavy traffic of QB-CSMA in a complete interference graph

Contents

5.1 Intuition and discussion 90
5.1.1 Fluid limits in the critical case, State space collapse 90
5.1.2 Idleness in random-access settings 93
5.1.3 Nonstandard behavior 94
5.1.4 Beyond $a < \frac{1}{2}$ 94
5.1.5 Initial state, limiting ODE 95
5.2 Main result 95
5.3 Notation and main steps of the proof 97
5.3.1 General notations 97
5.3.2 Localization, constants 98
5.3.3 Distance to I 98
5.3.4 Main steps 99
5.4 State space collapse 101
5.4.1 Proof of $\mathbb{E}(\Xi^N) \to 0$ 103
5.5 Proof of main result 104
5.5.1 First step: convergence of the sum to S 104
5.5.2 Second step: proof of Theorem 5.4 107

In this chapter, we will prove a heavy traffic result for Queue-Based CSMA on a complete interference graph. Once again, homogenization plays a central role. In addition, we will prove that the process of queue lengths collapses to a one-dimensional manifold. The process of interest will be

$$(Q^N(t), \sigma^N(t))_{t\geqslant 0} \coloneqq (\frac{Q(N^{1+a}t)}{N}, \sigma(N^{1+a}t))_{t\geqslant 0}.$$

5.1 Intuition and discussion

5.1.1 Fluid limits in the critical case, State space collapse

In this chapter we prove a heavy traffic result with the scheduling algorithm defined in Section 2.1.2. Since at most one queue can be active at any given time in a complete interference graph, if $s_1(\lambda) < 1$ one can find a scheduling algorithm such that the process of queue lengths is ergodic. For instance schedule q_v a fraction of the time $\frac{\lambda_v}{s_1(\lambda)}$ for all $v \in V$. If $s_1(\lambda) > 1$, for any scheduling algorithm the queueing process is transient. The critical case is then $s_1(\lambda) = 1$. To understand the behavior of the queueing process for long time horizons, it is insightful to consider the limiting ODE for the fluid limits. Recall q^* from Definition 4.2. From (4.3) we obviously get $s_1(q^*(t)) = s_1(q^0)$ for all $t \geqslant 0$. We now explain why the process of queue lengths experiences a state space collapse. To understand why this phenomenon occurs, let's look at the long term behavior of the solution to (4.3). Let

$$\begin{aligned} I &\coloneqq \left\{x \in \mathbb{R}_+^V : \forall v \in V,\ \lambda_v = \bar{\pi}^x(v)\right\} \\ &= \left\{x \in \mathbb{R}_+^V : \lambda_v^{-1/a} x_v = \lambda_w^{-1/a} x_w,\ v, w \in V\right\} \\ &= \left\{x \in \mathbb{R}_+^V : x_v = \frac{\lambda_v^{1/a}}{s_{1/a}(\lambda)} s_1(x),\ v \in V\right\}. \end{aligned} \tag{5.1}$$

We omit the dependency in q^0 as it will remain constant during this section. If $q^0 \in I$, by definition of I and (4.3), $q^*(t) = q^0$ for any $t \geqslant 0$. It is also possible to prove that even when the initial state does not lie in I, Proposition 5.1 states that the distance between $q^*(t)$ and I converges to 0 asymptotically. To be more precise, we will actually show in this section that the distance to the manifold is decreasing monotonically.

Proposition 5.1
For any $q^0 \in \mathbb{R}_+^V \setminus \{\mathbf{0}\}$, if $s_1(\lambda) = 1$ and G is a complete interference graph,

$$(d^\infty \circ q^*)(t) \leqslant d^\infty(q^*)(0) \exp(-\frac{C'a}{\epsilon} t)$$

In order to control the distance to the invariant manifold I given by (5.1), i.e. to control the state space collapse property, we will use the Kullback-Leibler divergence between λ and $(\bar{\pi}^q(v), v \in V)$ (note that both are probability measures on V). More precisely, for $q \in (0, +\infty)_+^V$, let

$$d^\infty(q) \coloneqq d_{\mathrm{KL}}(\lambda, \bar{\pi}^q) = \sum_{v \in V} \lambda_v \log\left(\frac{\lambda_v}{\bar{\pi}^q(v)}\right).$$

Note that $d^\infty(q) = 0$ if and only if $q \in I$, and $d^\infty(q) \geqslant 0$ so d^∞ can indeed be seen as a pseudo-distance to I.

Lemma 5.2

For any $v \in V$, $q \in (0, +\infty)^V$,

$$\partial_v d^\infty(q) = -\frac{a(\lambda_v - \bar{\pi}^q(v))}{q_v}.$$

Proof. For any $v \in V$, $\bar{\pi}^q(v) = \frac{q_v^a}{s_a(q)}$. Thus,

$$\partial_v \bar{\pi}^q(v) = \frac{a\bar{\pi}^q(v)\sum_{w\neq v}\bar{\pi}^q(w)}{q_v},$$

and for $w \neq v$,

$$\partial_v \bar{\pi}^q(w) = -\frac{a\bar{\pi}^q(v)\bar{\pi}^q(w)}{q_v}.$$

In addition, since $\partial_v d^\infty(q) = -\sum_{w\in V}\lambda_w \partial_v \log(\bar{\pi}^q(w))$ we get

$$\begin{aligned}
\partial_v d^\infty(q) &= -\frac{a\lambda_v\bar{\pi}^q(v)}{q_v\bar{\pi}^q(v)}\sum_{w\neq v}\bar{\pi}^q(w) + \sum_{w\neq v}\frac{a\lambda_w\bar{\pi}^q(v)\bar{\pi}^q(w)}{q_v\bar{\pi}^q(w)} \\
&= \frac{a}{q_v}\left[-\lambda_v\sum_{w\neq v}\bar{\pi}^q(w) + \bar{\pi}^q(v)\sum_{w\neq v}\lambda_w\right] \\
&= \frac{a}{q_v}(-\lambda_v + \bar{\pi}^q(v))
\end{aligned}$$

The last inequality is due to the fact that in the critical case, $\sum_{w\neq v}\lambda_w = 1-\lambda_v$. □

We can easily translate this into a result for the solutions to (4.3). Recall q_c^* the solution to the ODE (4.3).

Corollary 5.3
For any initial condition $q^0 \in (0, +\infty)^V$, for any $t > 0$,

$$(d^\infty \circ q^*)'(t) \leqslant 0.$$

As long as $\min_v q_v^*(t) \geqslant \epsilon$,

$$(d^\infty \circ q^*)'(t) \leqslant -\frac{a}{\epsilon}\left\|\lambda - \bar{\pi}^q\right\|_2^2.$$

Proof. Simply notice that $(d^\infty \circ q^*)'(t) = \sum_{v\in V}(\lambda_v - \bar{\pi}^q(v))\partial_v d^\infty(q^*(t)) = -\sum_{v\in V}\frac{a(\lambda_v - \bar{\pi}^q(v))^2}{q_v}$.

□

Proof of Proposition 5.1. Using the same reasoning as in the proof of Lemma 4.4, since $s_1(q^*)(t)$ is constant and $q_v^0 > 0$ for all v, we know that for any $t \geqslant 0$,

$$q_v^*(t) \geqslant \epsilon := \min_w\left(q_w^0, \frac{\lambda_w^{1/a}s_1(q^0)}{n}\right).$$

Since $\inf_{v\in V,t\geqslant 0} q_v^*(t) \geqslant \epsilon$,

$$\min_{v\in V,t\geqslant 0} \bar{\pi}^{q^*(t)}(v) \geqslant \frac{\epsilon^a}{s_a(q^*(t))} \geqslant \frac{\epsilon^a}{ns_1(q^*(t))^a} = \frac{\epsilon^a}{ns_1(q^0)^a}.$$

We can rewrite $d_{\mathrm{TV}}(\lambda, \bar{\pi}^q)$ as $2\left\|\lambda - \bar{\pi}^q\right\|_1$. Second, $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent norms. By the last inequality of Proposition 2.8,

$$d^\infty(q) \leqslant \frac{1}{\min_{v\in V} \bar{\pi}^q(v)} \left\|\lambda - \bar{\pi}^q\right\|_1^2.$$

To sum up, using the equivalence between $\|\cdot\|_1$ and $\|\cdot\|_2$,

$$(d^\infty \circ q^*)'(t) \leqslant -\frac{a}{\epsilon}\left\|\lambda - \bar{\pi}^{q^*(t)}\right\|_2^2 \leqslant -\frac{C'a}{\epsilon}\left\|\lambda - \bar{\pi}^{q^*(t)}\right\|_1^2 \leqslant -\frac{C'a}{\epsilon}\min_{v\in V} \bar{\pi}^{q^*(t)}(v)(d^\infty \circ q^*)(t).$$

Using the bound on $\bar{\pi}^{q^*}$,

$$(d^\infty \circ q^*)'(t) \leqslant -\frac{C'a\epsilon^{a-1}}{ns_1(q^0)^a}(d^\infty \circ q^*)(t),$$

Using Corollary 5.3 and the previous inequality, we know that

$$(d^\infty \circ q^*)'(t) \leqslant -\frac{C'a\epsilon^{a-1}}{ns_1(q^0)^a}\left\|\lambda - \bar{\pi}^{q^*(t)}\right\|_2 \leqslant -\frac{C'a\epsilon^{a-1}}{ns_1(q^0)^a}(d^\infty \circ q^*)(t).$$

Let $u(t) = \exp(-\frac{C'a\epsilon^{a-1}}{ns_1(q^0)^a}t)$. Using the product rule,

$$\begin{aligned}
\left(\frac{(d^\infty \circ q^*)}{u}\right)'(t) &= \frac{(d^\infty \circ q^*)'(t)u(t) - (d^\infty \circ q^*)(t)u'(t)}{u(t)^2} \\
&\leqslant \frac{C'a\epsilon^{a-1}}{ns_1(q^0)^a u(t)^2}\left((d^\infty \circ q^*)(t)u(t) - (d^\infty \circ q^*)(t)u(t)\right) \\
&= 0
\end{aligned}$$

Hence $\frac{(d^\infty \circ q^*)}{u}$ is decreasing and bounded by its initial value.

$$\frac{(d^\infty \circ q^*)}{u}(t) \leqslant \frac{(d^\infty \circ q^*)}{u}(0) = (d^\infty \circ q^*)(0).$$

To sum up, $(d^\infty \circ q^*)(t) \leqslant (d^\infty \circ q^*)(0)\exp(-\frac{C'a}{\epsilon}t) \to 0$ as $t \to +\infty$. □

The facts that there are initial states for which the fluid limit is constant and these points are attractive for the limiting ODE suggest to look at faster time scales to see a non-trivial evolution. The convergence of the multidimensional process to a process unidimensional is called state space collapse. The goal of Section 5.4 is to prove that such a state space collapse holds for the process of queue lengths. This will play an important role for the proof of convergence for the process of queue lengths scaled by N^{1+a} in time. Because this time scale is so much faster than the previous one, the state space collapse will happen instantaneously when not starting on the manifold.

The reason behind the state space collapse property for queue lengths is simple

to understand based on the stochastic averaging principle and Proposition 5.1. Once again, an averaging principle holds and we have the approximation

$$\int_0^t F\left(Q^N(s), \sigma^N(s)\right) \mathrm{d}s \approx \int_0^t \pi^{NQ^N(s)} \left[F\left(Q^N(s), \cdot\right)\right] \mathrm{d}s. \tag{5.2}$$

In Section 5.4, we use this approximation and the method from Proposition 5.1 to prove its equivalent for queue lengths Proposition 5.10.

5.1.2 Idleness in random-access settings

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, it is fairly well-understood that what happens when servers are idle can have a significant impact on the heavy traffic behavior and performance, for instance through reflection terms in diffusion limits. 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 $\frac{1}{(1-\rho)^2}$, 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 5.1.3, and in particular explain why the N/N^{1+a} scaling emerges and the heavy traffic is deterministic.

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 indeed 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. We mention as well [MS16], where the author proved that the delay scales like $(1-\rho)^{-1}(1-\frac{1}{2n})$ as the load of a $n \times n$ switched network grows to 1. Here the behavior is completely different because of the additional idleness induced by the decentralized nature of QB-CSMA. We will argue in Remark 5.5 that because of idle time, queue lengths scale like $\rho^{-1/a}$ as the load approaches 1.

5.1.3 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 $\pi^q(v) \approx \lambda_v$. 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 $\lambda_v - \pi^q(v)$ is of the order of N^{-a}. This suggests that node $v \in V$ behaves as a near-critical $M/M/1$ queue with arrival rate λ_v and service rate $\lambda_v - N^{-a}$. 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

$$N^b \lambda_v \left(f\left(q + \frac{e^v}{N}\right) - f(q) \right) + N^b \left(\lambda_v - N^{-a}\right) \left(f\left(q - \frac{e^v}{N}\right) - f(q) \right).$$

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 $\lambda_v - \pi^q$ of the order of $1/N$, then we would see the usual N/N^2 scaling and a diffusion process in the limit.

The critical behavior of most queueing networks that we know of is akin to a functional CLT: the time scale needs to be the square of the space scale and the limit is a diffusion, typically a reflected Brownian motion. See for instance [Whi02] for a review of standard results. In the model studied here, the behavior is nonstandard in both ways: 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, and the limit is actually deterministic and governed by an ordinary differential equation (ODE). 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.

5.1.4 Beyond $a < \frac{1}{2}$

Proposition 5.9 below shows that the averaging approximation (5.2) holds for $a < 1/2$ but this is not the condition that is expected for it to hold: since the typical time scale of the slow process is N and the mixing time of the fast process N^a, 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 5.4 is the more stringent condition $a < 1/2$. To see why this condition pops up, consider the following semimartingale decomposition of Q^N:

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

The martingale term can be shown to vanish for $a < 1$, but we see that in order for the error term to also vanish we would need to show that the integral from Proposition 5.9 decreases to 0. Proposition 5.9 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 $a < 1/2$. We also note that this threshold $1/2$ corresponds to the threshold at which the lingering effect pointed out in [SBB14] kicks in. Indeed, in the near-critical case $\lambda_v = \lambda_v^0 - \gamma_v \varepsilon$ with $\varepsilon > 0$, then the steady state should be of the order of $\varepsilon^{-1/a}$. For $a > 1/2$ this would suggest a scaling $\varepsilon^{-1/a} \ll \varepsilon^{-2}$ but the lingering effect discussed in [SBB14] suggests that because of idleness consideration, one cannot go beyond this ε^{-2} scaling. This would constitute an argument against homogenization for $a > 1/2$, although it is not clear in our view that the lingering effect indeed kicks in in the case of the complete interference graph.

5.1.5 Initial state, limiting ODE

We fix throughout an initial state $q^0 \in I \setminus \{\mathbf{0}\}$ and assume that $Q^N(0) \to q^0$. Moreover, we consider $S = (S(t), t \geq 0)$ the solution to the ODE $\dot{x} = \mu x^{-a}$ with initial condition $S(0) = s_1(q^0)$, i.e.,

$$S(t) = \left(\mu(a+1)t + s_1(q^0)^{a+1}\right)^{1/(a+1)}, \ t \geq 0.$$

We also consider $\bar{q} = (\bar{q}(t), t \geq 0)$ the $\mathbb{R}_+^V$-valued function with $s_1 \circ \bar{q} = S$ and $q(t) \in I$ for all $t \geq 0$, i.e.,

$$\bar{q}_v(t) = \frac{\lambda_v^{1/a}}{s_{1/a}(\lambda)} S(t), t \geq 0, v \in V.$$

With some extra work, but without giving much more insight on the system's behavior, our 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 present then discuss this result in more details.

5.2 Main result

Recall the sequence $((Q^N, \sigma^N), N \geq 1)$ of scaled processes given by

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

In the sequel we use $\xrightarrow{\mathbb{P}}$ 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 critical case $s_1(\lambda) = 1$.

Theorem 5.4

Assume that the three following assumptions hold:

- $a < 1/2$;
- $s_1(\lambda) = 1$;
- $Q^N(0) \to q^0$ *for some* $q^0 \in I \setminus \{0\}$.

Then $Q^N \xrightarrow{\mathbb{P}} \bar{q}$ *uniformly on compact time-sets, where* $\bar{q}$ *is uniquely characterized as follows:* $\bar{q}(t) \in I$ *for every* $t \geq 0$ *and* $s_1 \circ \bar{q}$ *is the unique solution to the ODE* $\dot{x} = \mu x^{-a}$ *with initial condition* $x(0) = s_1(q^0)$ *where* $\mu = s_{1/a}(\lambda)^a$.

Note that the limiting process $\bar{q}$ actually has an explicit expression, namely for every $v \in V$ and $t \geq 0$,

$$\bar{q}_v(t) = \frac{\lambda_v^{1/a}}{s_{1/a}(\lambda)} \left(\mu(a+1)t + s_1(q^0)^{a+1}\right)^{1/(a+1)}.$$

In the rest of the chapter we assume that the conditions of this theorem are enforced, i.e., we assume throughout that $a < \frac{1}{2}$, that $s_1(\lambda) = 1$ and that $Q^N(0) \to q^0 \in I \setminus \{0\}$.

Remark 5.5
Our result could also be generalized to the near-critical case where $\lambda_v = \lambda_v^0 - \varepsilon\gamma_v$ *for* $v \in V$, *for some* $\varepsilon \in \mathbb{R}_+$ *and vectors* $\lambda^0 = (\lambda_v^0, v \in V) \in \mathbb{R}_+^V$ *with* $s_1(\lambda^0) = 1$ *and* $\gamma = (\gamma_v, v \in V) \in \mathbb{R}^V$. *Up until the time queue lengths reach an order of magnitude* $\varepsilon^{1/a}$, *the difference between the arrival rates for positive* ε *and its limit has a negligible influence on the dynamic. As long as* $\|q^\epsilon\|_\infty \ll \epsilon^{-1/a}$, *we have*

$$\epsilon \ll \pi^{q^\epsilon}(\boldsymbol{0}),$$

Which leads to increasing queue lengths . In this case, in order to see both the influence of idle time and convergence to the border of the stability region, the correct scaling is also nonstandard and given by

$$Q^\varepsilon(t) = \varepsilon^{1/a} Q(\varepsilon^{-\frac{a+1}{a}} t), \varepsilon > 0, t \geqslant 0.$$

If the queue lengths have an order of magnitude larger than $\varepsilon^{-1/a}$, *the idle time will have a negligible influence on the dynamic because as long as* $\|q^\varepsilon\|_\infty \gg \varepsilon^{-1/a}$, *we have*

$$\varepsilon \gg \pi^{q^\varepsilon}(\boldsymbol{0}),$$

which leads to decreasing queue lengths until they reach the order of magnitude $\varepsilon^{-1/a}$. *As* $\varepsilon \to 0$, *we have* $Q^\varepsilon \xrightarrow{\mathbb{P}} q$ *with* q *constrained to one-dimensional manifold* I *and such that* $s_1 \circ q$ *is the unique solution to the ODE*

$$x' = \mu x^{-a} - s_1(\gamma),$$

with $\mu \coloneqq (\sum_{v \in V} \lambda_v^{1/a})^a$. *Except in the case* $s_1(\gamma) = 0$, *there seems to be no explicit solution to this ODE. What can be proved however is that if* $s_1(\gamma) \leqslant 0$ *then* $x(t) \to \infty$ *while if* $s_1(\gamma) > 0$ *then* $x(t) \to (\mu/s_1(\gamma))^{1/a}$.

5.3 Notation and main steps of the proof

5.3.1 General notations

Since we impose that only one node can be active at a time, whenever convenient we will identify σ with the active node, or put $\sigma = \mathbf{0}$ 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 $V_0 = V \cup \{\mathbf{0}\}$ when seeing σ as the current schedule. Because a schedule is associated to 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 $q_{\mathbf{0}} = 0$. Note that with this convention, we have $\sigma_0 = 1 - \sum_{v\in V} \sigma_v$.

The generators and *carré du champ* used in this Chapter are the same ones from Section 3.3 with $\theta = 1 + a$. Said otherwise, for any $f : \frac{1}{N}\mathbb{N}^V \times V_0 \to \mathbb{R}$, $g : \frac{1}{N}\mathbb{N}^V \to \mathbb{R}$, $h : V_0 \to \mathbb{R}$, $q \in \frac{1}{N}\mathbb{N}^V$ in the respective domains, $L^N[f]$, $L_{\mathrm{f}}^{N,q}[g]$ and $L_{\mathrm{s}}^{N,\sigma}[h]$ are given for $q \in \frac{1}{N}\mathbb{N}^V$ and $\sigma \in V_0$ by

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

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 $g^N(q) = g(q/N)$. From Proposition 2.1, for a suitable f,

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

is a local martingale with increasing process

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

with Γ^N the *carré du champ* associated to L^N.

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

$$L_{\mathrm{h}}^N f(q) = N^{a+1} \sum_{v\in V} \lambda_v \left(f\left(q + \frac{e^v}{N}\right) - f(q)\right) + N^{a+1} \sum_{v\in V} \pi^{Nq}(v) \mathbb{1}_{q_v>0} \left(f\left(q - \frac{e^v}{N}\right) - f(q)\right). \quad (5.3)$$

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

5.3.2 Localization, constants

Most of the proof of Theorem 5.4 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:

$$C_+ = \max\left(2S(T), \frac{2}{S(0)}, \frac{1}{2}\right) \quad \text{and} \quad C_- = \frac{1}{C_+\mu^{1/a}} \min_v \lambda_v^{1/a}.$$

Here and in the sequel, we will treat as constants all numerical parameters that only depend on a, n, and λ as these are fixed throughout the entire chapter. Recall that C is a finite constant that only depends on a, n, T, λ and q^0 whose precise value is irrelevant and that may change from line to line.

We then define

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

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

$$U \coloneqq \left\{q \in \mathbb{R}_+^V : \frac{1}{C_+} < s_1(q) < C_+ \quad \text{and} \quad \min_i q_i > C_-\right\},$$

its intersection U^N with $\frac{1}{N}\mathbb{N}^V$

$$U^N = U \cap \frac{1}{N}\mathbb{N}^V,$$

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

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

Because jumps of Q^N are of size $1/N$, at time T^N we have $\|Q^N(T^N) - \bar{q}(T^N)\| \leqslant C_-/2 + 1/N$. The constants C_- and C_+ have been chosen such that the following result holds. The proof is similar to Lemma 4.12 in the previous chapter so we omit it here.

Lemma 5.6
We have $T^N < \tau^N$. In particular, $Q^N(t \wedge T^N) \in U^N$ for all $t \geq 0$.

5.3.3 Distance to I

In order to control the distance to the invariant manifold I given by (5.1), 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

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

When $N \to \infty$ and $(q, \sigma) \in U \times V_0$, by Lemma 2.4, since $\bar{\pi}^q = \pi^q_\infty$ for a complete interference graph, we have $\pi^{Nq}(\sigma) \to \pi^q_\infty(\sigma)$ where

$$\pi^q_\infty(\sigma) = \frac{q^a_\sigma}{s_a(q)},\ v \in V,$$

with the convention $q_{\mathbf{0}} = 0$. We thus introduce

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

Note that $d^\infty(q) = 0$ if and only if $q \in I$, so d^∞ can indeed be seen as a distance to I. As the next lemma shows, as long as q stays in U the two distances d^N and d^∞ are close to each other.

Lemma 5.7
As $N \to \infty$ we have

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

Proof. For any $q \in U^N$,

$$\begin{aligned}
|d^N(q) - d^\infty(q)| &= \sum_{v \in V} \lambda_v \left| \log(\frac{\pi^q_\infty(v)}{\pi^{Nq}(v)}) \right| \\
&\leqslant \sum_{v \in V} \lambda_v \left(\left| \log(\frac{q^a_v}{(q_v + N^{-1})^a}) \right| + \left| \log(\frac{N^{-a} + \sum_{v \in V}(q_v + N^{-1})^a}{s_a(q)}) \right| \right) \\
&\leqslant \sum_{v \in V} \lambda_v \left(\left| \log(\frac{1}{(1 + \frac{1}{Nq_v})^a}) \right| + \left| \log\left(N^{-a} + \sum_{v \in V}(q_v + N^{-1})^a \right) - \log(s_a(q)) \right| \right).
\end{aligned}$$

To conclude the proof, compute Taylor expansions of $\log(\frac{1}{(1+t/q_v)^a})$ and $\log(t^a + s_a(q+t))$ for t a neighbourhood of 0 and use the fact that

$$C_- \leqslant \min_v q_v \leqslant \max_v q_v \leqslant C_+$$

to obtain an explicit bound depending only on N, C_- and C_+, that vanishes as $N \to +\infty$. □

5.3.4 Main steps

In order to be able to use the improved bounds from Corollary 3.7, we need to prove this lemma on idle time:

Lemma 5.8
We have

$$\mathbb{E}\left[\int_0^{T \wedge T^N} \sigma^N_{\mathbf{0}}(s) \mathrm{d}s \right] \leqslant CTN^{-a} + CTN^{a-1} \log(N)^3,$$

or equivalently

$$\mathbb{E}\left[\int_0^{N^{1+a}(T\wedge T^N)} \sigma_{\boldsymbol{0}}(s)\mathrm{d}s\right] \leqslant CTN + CTN^{2a}\log(N)^3,$$

Proof. Note that

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

and so since $Q_j^N(s) \geq C_-$ for $t \leqslant T^N$, we have $\pi^{NQ^N(s)}(0) \leqslant CN^{-a}$ for $s \leqslant T^N$ and so

$$\mathbb{E}\left[\int_0^{T\wedge T^N} \sigma_{\mathbf{0}}^N(s)\mathrm{d}s\right] \leqslant CTN^{-a} + \mathbb{E}\left[\int_0^{T\wedge T^N}\left(\sigma_{\mathbf{0}}^N(s) - \pi^{NQ^N(s)}(\mathbf{0})\right)\mathrm{d}s\right].$$

We now deal with the homogenization term: recall $\phi_{\mathbf{0}}(q,\cdot)$ the solution to the Poisson equation

$$L^q[\phi_{\mathbf{0}}(q,\cdot)](\sigma) = \sigma_{\mathbf{0}} - \pi^q(\mathbf{0}).$$

We get

$$\begin{aligned}
\mathbb{E}\left[\int_0^{T\wedge T^N}\left(\sigma_{\mathbf{0}}^N(s) - \pi^{NQ^N(s)}(\mathbf{0})\right)\mathrm{d}s\right] &= \mathbb{E}\left[\frac{1}{N^{1+a}}\int_0^{N^{a+1}\left(T\wedge T^N\right)} \sigma_{\mathbf{0}}(s) - \pi^{Q(s)}(\mathbf{0})\mathrm{d}s\right] \\
&= \frac{1}{N^{a+1}}\mathbb{E}\left[\phi_{\mathbf{0}}\left(Q(N^{a+1}\left(T\wedge T^N\right), \sigma(N^{a+1}\left(T\wedge T^N\right)\right)\right] \\
&- \frac{1}{N^{a+1}}\mathbb{E}\left[\int_0^{N^{a+1}\left(T\wedge T^N\right)} L_{\mathrm{s}}^{\sigma(s)}\left(\phi_{\mathbf{0}}(\,\cdot\,,\sigma(s))\right)(Q(s))\,\mathrm{d}s\right] \\
&+ \frac{1}{N^{a+1}}\phi_{\mathbf{0}}\left(Q(0),\sigma(0)\right).
\end{aligned}$$

By definition of L_{s}^{σ} we have

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

From Lemma 3.6, with U^N as localization set, we get

$$\mathbb{E}\left[\int_0^{T\wedge T^N}\left(\sigma_{\mathbf{0}}^N(s) - \pi^{NQ^N(s)}(\mathbf{0})\right)\mathrm{d}s\right] \leqslant N^{-a-1}T\left(CN^a\log(N)^{3/2} + CN^{a+1+a-1}\log(N)^3\right).$$

The last quantity is smaller than $CN^{a-1}\log(N)^3$. □

First step: homogenization

Recall that C denotes a numerical constant allowed to depend on a, n and λ. We begin by stating the homogenization result:

Proposition 5.9

If $f : U \to \mathbb{R}$ is twice continuously differentiable and $T > 1$ then

$$\mathbb{E}\left[\sup_{t\leqslant T\wedge\bar{\tau}^U(Q^N)}\left|\int_0^t\left(L_{\mathrm{s},N}^{\sigma^N(s)}-L_{\mathrm{h},N}\right)(g)(Q^N(s))\mathrm{d}s\right|\right]\leqslant C\max_v\|\partial_v g\|_{\infty,U}\,N^{a-\frac{1}{2}}\log(N)^{3/2}$$
$$+C\max_{v,w\in V}\left\|\partial^2_{v,w}g\right\|_{\infty,U}N^{2a-1}\log(N)^{3/2}.$$

Proof. This result is a restatement of Corollary 3.7. □

Second step: State Space Collapse

Using the averaging result of Proposition 5.9, the next step is to prove the following state space collapse result.

Proposition 5.10
As $N\to\infty$ we have

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

The proof proceeds in two steps: we first control 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 5.9. This result is proved in Section 5.4.

Third step: full proof

The third step of the proof consists in showing that $Q^n(\cdot\wedge T^N)\xrightarrow{\mathbb{P}}\bar{q}$. The proof proceeds in two steps: first we establish the convergence of the one-dimensional total queue length process $s_1\circ Q^N(\cdot\wedge T^N)\xrightarrow{\mathbb{P}}s_1\circ\bar{q}=S$ by using Gronwall's lemma. Together with the state space collapse property of Proposition 5.10, 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.

5.4 State space collapse

In this section we prove Proposition 5.10 through a series of lemmas. We start by writing

$$d^N(Q^N(t))=d^N(Q^N(0))+\int_0^t L_{\mathrm{h}}^N[d^N](Q^N(s))\mathrm{d}s$$
$$+\int_0^t(L^N-L_{\mathrm{h}}^N)[d^N](Q^N(s),\sigma^N(s))\mathrm{d}s+M_{d^N}^N(t).\quad(5.4)$$

In the above expression, in order to give sense to $L^N d^N$ we consider $d^N(q,\sigma) = d^N(q)$. Consider the following lemma, which we will prove later on.

Lemma 5.11
For any $q \in \frac{1}{N}\mathbb{N}^V \cap U$ we have $L^N_{\mathrm{h}}[d^N](q) \leqslant CN^{-(1-a)}$.

Then defining

$$\Xi^N = d^N(Q^N(0)) + \sup_{0\leqslant t\leqslant T\wedge T^N} \left|M^N_{d^N}(t)\right| + \frac{CT}{N^{1-a}} + \sup_{0\leqslant t\leqslant T\wedge T^N} \left|\int_0^t (L^N - L^N_{\mathrm{h}})[d^N](Q^N(s),\sigma^N(s))\mathrm{d}s\right| \quad (5.5)$$

we obtain $d^N(Q^N(t)) \leqslant \Xi^N$ for any $t \leqslant T^N$. In particular, in order to prove Proposition 5.10 we only have to prove Lemma 5.11 and that $\mathbb{E}(\Xi^N) \to 0$. We first prove Lemma 5.11 and then $\mathbb{E}(\Xi^N) \to 0$ in the next section.

Proof of Lemma 5.11. Let $q \in U^N$ and for each $v \in V$, let $\zeta^v_{\pm}$ such that

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

Then

$$\begin{aligned}
L^N_{\mathrm{h}}[d^N](q) &= N^{a+1}\sum_{v\in V}\lambda_v \Delta^N_{+,v} d^N(q) + N^{a+1}\sum_{v=v}^{n}\pi^{Nq}(v)\Delta^N_{-,v}d^N(q)\\
&= N^{a+1}\sum_{v\in V}\lambda_v\left(\frac{1}{N}\partial_v d^N(q) + \frac{1}{2N^2}\partial^2_{vv}d^N(\zeta^v_+)\right)\\
&\quad + N^{a+1}\sum_{v\in V}\pi^{Nq}(v)\left(-\frac{1}{N}\partial_v d^N(q) + \frac{1}{2N^2}\partial^2_{vv}d^N(\zeta^v_-)\right)\\
&= N^a\sum_{v\in V}\partial_v d^N(q)\delta^N_v(q)\\
&\quad + \frac{1}{2N^{1-a}}\sum_{v\in V}\left(\partial^2_{vv}d^N(\zeta^v_+)\lambda_v + \partial^2_{vv}d^N(\zeta^v_-)\pi^{Nq}(v)\right)
\end{aligned}$$

with $\delta^N_v(q) = \lambda_v - \pi^{Nq}(v)$. Similarly to Lemma 5.2t may be checked through elementary algebra that

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

and that

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

Plugging in these expressions and bounds in the previous expression for $L^N_{\mathrm{h}}[d^N]$ we

thus obtain

$$L_{\mathrm{h}}^N[d^N](q) \leqslant -aN^a \sum_{v\in V} \frac{\delta_v^N(q)^2}{q_v + \frac{1}{N}} + \frac{1}{N^{1-a}C_-^2}$$

which gives the result. □

5.4.1 Proof of $\mathbb{E}(\Xi^N) \to 0$

We now prove that $\mathbb{E}(\Xi^N) \to 0$ through a series of lemmas. As explained above, this implies Proposition 5.10. In view of the expression (5.5) of Ξ^N, we have four terms to control. The third term $CTN^{-(1-a)}$ vanishes because $a < 1$. The first term $d^N(Q^N(0))$ also vanishes, because $Q^N(0) \to q^0 \in I$ and in view of Lemma 5.7. The next two lemmas show that the other two terms also vanish, which concludes the desired proof of $\mathbb{E}(\Xi^N) \to 0$.

Lemma 5.12
We have

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

Proof. This lemma is a direct consequence of Proposition 5.9 because d^N is twice differentiable with bounded derivative. □

Lemma 5.13
We have

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

Proof. Proceeding as in the proof of Lemma 3.4 we obtain

$$\mathbb{E}\left[\sup_{0\leqslant t\leqslant T\wedge T^N} M_{d^N}^N(t)^2\right] \leqslant 4N^{a+1}\mathbb{E}\left[\int_0^{T\wedge T^N}\sum_{v\in V}(\Delta_{+,v}^N d^N(Q^N(s))^2\mathrm{d}s\right] + 4N^{a+1}\mathbb{E}\left[\int_0^{T\wedge T^N}\sum_{v\in V}(\Delta_{-,v}^N d^N(Q^N(s))^2\mathrm{d}s\right].$$

The result then follows from the same Taylor expansion as in the proof of Lemma 5.11. □

The proof of Proposition 5.10 is therefore complete.

Remark 5.14
If q^0 is not on I, $\mathbb{E}\left[\Xi^N\right]$ does not converge to 0. Using the same method as in Proposition 5.1 it is possible to prove that $\mathbb{E}\left[\left\|\delta^N(Q^N(s))\right\|_2^2\right] \to 0$ as $N \to +\infty$ for

almost every $t \geqslant 0$: indeed,

$$\int_0^{T\wedge T^N} \left\|\delta^N(Q^N(s))\right\|_2^2 \mathrm{d}s \leqslant CN^{-a}\left(d^N(Q^N(0)) - \left\|\delta^N(Q^N(t))\right\|_2^2 + CN^{a-1} + M^N_{d^N}(t)\right) + CN^{-a}\sup_{0\leqslant t\leqslant T\wedge T^N}\left|\int_0^{t\wedge T^N}(L^N - L^N_{\mathrm{h}})[d^N](Q^N(s),\sigma^N(s))\mathrm{d}s\right|.$$

If $q^0_v > 0$ for every v, $d^N(Q^N(0))$ remains bounded. This means that the right hand side converges to 0. This is enough to prove uniform convergence of $s_1(Q^N)$, but only enough for

$$\mathbb{E}\left[\int_0^T \left\|Q^N(s) - q^*(s)\right\| \mathrm{d}s\right] \to 0 \text{ as } N \to +\infty.$$

5.5 Proof of main result

To prove Theorem 5.4, 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 5.6.

Because of the N^{1+a} time scale, Q^N can have variation of order N^a even for small times. The process has a continuous limit because the homogenized service rates are close to the arrival rates, which slows down the variation speed for queue lengths. The crucial part of this argument is that $s_1 \circ Q^N$ evolves slow enough to be tight and have a continuous function as a limit. We do not need to prove this result because we will directly show that the distance between the process and its limit decreases uniformly to 0. To identify the limit, we use the fact that when $q \in I$, it is possible to write

$$L_{\mathrm{h}}[s_1](q) = \frac{\mu}{s_1(q)^a},$$

which indicates an ODE-type behavior.

One of the major disadvantages of the method from [LN13] is that it is required to approximate the dynamic of the fast variable with rates that only depend on the slow variable. For this time scale, Q^N evolves too fast for this method to be applicable. In addition, without the state space collapse, it is not possible to express the activation/deactivation rates of a specific queue using only information about the sum of queue lengths. Finally, in order to prove the collapse of the state space for the queue lengths, it is necessary to prove a homogenization result beforehand in order to replace the service rate by their steady state average in order to use the same method as in Proposition 5.1.

5.5.1 First step: convergence of the sum to S

The first step is to prove that $s_1 \circ Q^N(\cdot \wedge T^N) \xrightarrow{\mathbb{P}} S$ uniformly on $[0,T]$, which we do now. Starting from the definition of $L^{N,\sigma}_{\mathrm{s}}$ and using $\sum_{v\in V}\lambda_v = 1$ and

$\sum_{v\in V}\sigma_v = 1-\sigma_0$ we obtain

$$L_{\mathrm{s}}^{N,\sigma}[s_1](q) = N^a - N^a\sum_{v\in V}\sigma_v \mathbb{1}_{q_v>0} = N^a\sigma_0 + N^a\sum_{v\in V}\sigma_v\mathbb{1}_{q_v=0}.$$

The semimartingale decomposition of $s_1\circ Q^N$ and the fact that $S(t) = S(0) + \mu\int_0^t S(s)^{-a}\mathrm{d}s$ by definition of S then leads to

$$\begin{aligned} s_1(Q^N(t)) - S(t) = s_1(Q^N(0)) - S(0) + \int_0^t \left(N^a\sigma_0^N - \frac{\mu}{S(s)^a}\right)\mathrm{d}s \\ + \sum_{v\in V}\int_0^t \sigma_v^N(s)\mathbb{1}_{Q_v^N(s)=0}\mathrm{d}s + M_{s_1}^N(t). \quad (5.6) \end{aligned}$$

Define

$$\varepsilon^N(t) = s_1(Q^N(0)) - S(0) + \eta^N(t) + e^N(t) + h^N(t) + M_{s_1}^N(t)$$

where

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

$$e^N(t) = \int_0^{t\wedge T^N}\left(\frac{1}{s_a(Q^N(s))} - \frac{\mu}{s_1(Q^N(s))^a}\right)\mathrm{d}s$$

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 $Q_i^N(s) > 0$ for $t < T^N$, starting from (5.6) and plugging in the above expressions, we obtain

$$s_1(Q^N(t\wedge T^N)) - S(t) = \varepsilon^N(t) + \mu\int_0^t\left(\frac{1}{s_1(Q^N(s))^a} - \frac{1}{S(s)^a}\right)\mathrm{d}s$$

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

$$\left|s_1(Q^N(t\wedge T^N)) - S(t)\right| \le \left|\varepsilon^N(t)\right| + C\int_0^t\left|s_1(Q^N(s\wedge T^N)) - S(s)\right|\mathrm{d}s$$

and Gronwall's lemma from [EK86] Appendix 5 implies

$$\begin{aligned} &\sup_{0\le t\le T}\left|s_1(Q^N(t\wedge T^N)) - S(t)\right| \\ &\qquad\le \left(\left|s_1(Q^N(0)) - S(0)\right| + \bar\eta^N + \bar e^N + \bar h^N + \sup_{0\le t\le T\wedge T^N}\left|M_{s_1}^N(t)\right|\right)e^{CT} \end{aligned}$$

with

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

$$\bar e^N = \int_0^{T\wedge T^N}\left|\frac{1}{s_a(Q^N(s))} - \frac{\mu}{s_1(Q^N(s))^a}\right|\mathrm{d}s$$

and

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

By assumption we have $s_1(Q^N(0)) \to S(0)$ and so in order to prove the desired result $s_1 \circ Q^N(\,\cdot\,\wedge T^N) \xrightarrow{\mathbb{P}} S$ 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 $\bar{\eta}^N \xrightarrow{\mathbb{P}} 0$ is straightforward (using that $Q^N(t\wedge T^N) \in U$). The martingale term is handled with the exact same arguments as the previous martingale terms in Lemmas 3.4 and 5.13, the proof is omitted. The next two lemmas show that the other two terms also vanish.

Lemma 5.15
We have $\bar{e}^N \xrightarrow{\mathbb{P}} 0$.

Proof. Let $\varepsilon > 0$: since $d^\infty \circ Q^N(\,\cdot\,\wedge T^N) \xrightarrow{\mathbb{P}} 0$ according to Proposition 3.3, we only have to prove that $\mathbb{P}(\bar{e}^N \ge \varepsilon, X \le \eta) \to 0$ for $\eta > 0$ small enough, where

$$X = \sup_{0\le t\le T\wedge T^N} d^\infty(Q^n(t)).$$

Actually, we will show that if η is small enough, then $X \le \eta$ implies $\bar{e}^N < \varepsilon$. For $q \in U$ we have

$$\left| \frac{1}{s_a(q)} - \frac{\mu}{s_1(q)^a} \right| \le C \left| s_1(q)^a - \mu s_a(q) \right|.$$

Using Pinsker's inequality and elementary algebra, one can show that if $d^\infty(q) \le \eta$ with η small enough, then $|s_1(q)^a - \mu s_a(q)| \le C\eta^{1/2}$. Combining the above bounds, we see that if $X \le \eta$, then for every $t \le T^N$ we have

$$\left| \frac{1}{s_a(Q^N(s))} - \frac{\mu}{s_1(Q^N(s))^a} \right| \le C\eta^{1/2}$$

which proves the result. □

Lemma 5.16
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

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

and so Proposition 3.3 with $f = 1$ implies that

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

As $a < 1/2$ we have the result. □

5.5.2 Second step: proof of Theorem 5.4

We now conclude the proof of Theorem 5.4. Fix $v \in V$, we write

$$\sup_{0\le t\le T\wedge T^N} \left|Q_v^N(t)^a - \bar{q}_v(t)^a\right| \le \varepsilon_1^N + \varepsilon_2^N + \varepsilon_3^N$$

with

$$\varepsilon_1^N = \sup_{0\le t\le T\wedge T^N} \left|Q_v^N(t)^a - \lambda_v s_a(Q^N(t))^a\right|,$$

$$\varepsilon_2^N = \sup_{0\le t\le T\wedge T^N} \left|\lambda_v s_a(Q^N(t))^a - \frac{\lambda_v}{\mu} s_1(Q^N(t))^a\right|$$

and

$$\varepsilon_3^N = \frac{\lambda_i}{\mu} \sup_{0\le t\le T\wedge T^N} \left|s_1(Q^N(t))^a - S(t)^a\right|.$$

We have just proved that $\varepsilon_3^N \xrightarrow{\mathbb{P}} 0$. Moreover, using similar arguments as in the previous step, we can prove that $\varepsilon_2^N \xrightarrow{\mathbb{P}} 0$. The first term ε_1^N also vanishes because by Pinsker's inequality, for $t \le T^N$ we have

$$\left|Q_v^N(t)^a - \lambda_v s_a(Q^N(t))^a\right|^2 \le C d^\infty(Q^N(t))$$

and so $\varepsilon_1^N \xrightarrow{\mathbb{P}} 0$ according to Proposition 5.10. We thus have $Q^N(\cdot \wedge T^N) \xrightarrow{\mathbb{P}} \bar{q}$ uniformly on $[0,T]$.

Let us now remove the localization and prove that $Q^N \xrightarrow{\mathbb{P}} \bar{q}$ uniformly on $[0,T]$. In order to do so, it is enough to show that $\mathbb{P}(T^N \ge T) \to 1$. By definition of T^N, we have

$$\left\|Q^N(T^N) - \bar{q}(T^N)\right\|_1 \ge \frac{C_-}{2}.$$

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

$$\mathbb{P}\left(T^N \le T\right) \le \mathbb{P}\left(\left\|Q^N(T^N \wedge T) - \bar{q}(T^N \wedge T)\right\|_1 \ge \frac{C_-}{2}\right).$$

Since we have proved that $Q^N(\cdot \wedge T^N) \xrightarrow{\mathbb{P}} \bar{q}$ uniformly on $[0,T]$, the previous probability vanishes. This concludes the proof of Theorem 5.4.

Chapter 6

Conclusion

Contents

6.1 Generalizing the fluid limits result 109
6.1.1 How/when do coordinates of solutions to the ODE reach 0? . 110
6.1.2 Homogenization when touching 0 111
6.2 Generalizing the Heavy traffic result 112
6.2.1 Additional time scales, State space collapse 112
6.2.2 Diffusive time scale 114
6.3 Generalizing the homogenization result 115
6.3.1 Metastability phenomenon 116
6.3.2 Large deviations 116

In this thesis, we presented some advances in provable approximations for QB-CSMA. To be able to prove those approximations, we developed a new method to obtain a stochastic averaging principle. Because of this phenomenon, we are able to prove convergence of $\frac{Q(N\cdot)}{N}$ in a general setting and $\frac{Q(N^{1+a}\cdot)}{N}$ in a complete interference graph for critical arrival rates as $N \to +\infty$ to deterministic processes governed by ODEs.

6.1 Generalizing the fluid limits result

For fluid limits, the limiting ODE is given by (4.2):

$$\begin{cases} f' = g(f) \\ f(0) = q^0 \end{cases},$$

with

$$\begin{array}{rl} g : \mathbb{R}_+^V \setminus \{\mathbf{0}\} & \to [-1,1]^V \\ q & \longmapsto \lambda - \overline{\pi}^q, \end{array}$$

and $\bar{\pi}^q(v) = \pi^q_\infty(\sigma_v = 1)$ and $\pi^q_\infty(\sigma) = \lim_{N\to+\infty} \pi^{Nq}(\sigma)$ for any $v \in V$ and $\sigma \in S(G)$. Remember from Chapter 4 that (4.2) has a unique maximal solution

defined up to $T_{\text{ext}}(q^0)$ the exit time of $(0,+\infty)^V$. In this section, it is important to highlight the dependence of the solution to (4.2) on the interference graph. For any graph G, let's use the notation q^G for the solution from Definition 4.2 when the interference graph is G. The dependence in the interference graph is also emphasized in β from (2.8) which actually depends on G.

The two first discussions regard the study of the solution to the ODE (4.2) while the third discussion regards problems related to homogenization when some coordinates are too small in the queueing model.

6.1.1 How/when do coordinates of solutions to the ODE reach 0?

Theorem 4.5 is limited in time by $T_{\text{ext}}(q^0)$. It would be interesting to identify situations where $T_{\text{ext}}(q^0) = +\infty$. For instance if $\lambda_v > 1$ for every v, the input rate at each q_v^G is greater than the decrease rate so no queue can reach 0. As we saw in Section 4.2.3 when $\lambda \in (1-\epsilon)\Lambda^*(S(G))$, at least one coordinate will reach 0 in finite time. It would be interesting to characterize interference graphs such that a result analogous to Lemma 4.4 can be proved. It essentially states that the only possibility for solutions of (4.2) to exit $(0,+\infty)^V$ is by approaching $\mathbf{0}$.

To formalize this question, let us define two classes of graphs:

Definition 6.1
Let $\mathcal{G}$ be the set of graphs such that for any $G \in \mathcal{G}$, $q^0, \lambda \in (0,+\infty)^V$,

$$\tau^0(q^G(\cdot, q^0)) = \tau^0(s_1 \circ q^G(\cdot, q^0)).$$

We will say that $G \in \mathcal{G}(\lambda, q^0)$ if

$$\tau^0(q^G(\cdot, q^0)) = \tau^0(s_1 \circ q^G(\cdot, q^0))$$

Lemma 4.4 states that $\mathcal{G}$ contains the complete interference graphs with any finite number of nodes. Identifying a family of graph in $\mathcal{G}$ can be a challenging question for the future. In some interference graph, having $\tau^0(q^G(\cdot, q^0)) = \tau^0(s_1 \circ q^G(\cdot, q^0))$ may also depend on the initial state or the arrival rates. We mention for instance the case of 4 nodes on a square where that may be the case. It is only possible to schedule two nodes at the same time if they are on opposit corners of the square. We will go in further details about this example at the end of the section.

Once $\mathcal{G}$ has been identified, it suffice to identify the behavior when some queues start null to prove the folowing conjecture:

Conjecture 6.2
Let $q^0 \in \mathbb{R}_+^V$, $\lambda \in \mathbb{R}_+^V$, and assume that

- *The initial condition $Q^N(0)$ converges to q^0 as $N \to +\infty$.*
- *The interference graph G is an element of $\mathcal{G}$.*
- *The condition on the spectral gap is $2a\beta(G) < 1$.*

Then

$$\frac{Q(N\cdot)}{N} \xrightarrow{\mathbb{P}} q^G(\cdot, q^0)$$

uniformly over compact time sets, as $N \to +\infty$*, with* q^G *from Definition 4.2.*

If $q^0 \in (0, +\infty)^V$*, this conjecture is proved in the same way as Theorem 5.4*

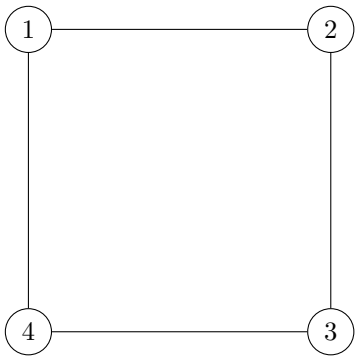

Figure 6.1: Square interference graph with 4 nodes

Depending on arrival rates and initial conditions it is possible that

$$\tau^0(q^G(\cdot, q^0)) \neq \tau^0(s_1 \circ q^G(\cdot, q^0)).$$

For the situation described in Figure 6.1, if q_1^0 is close to 0 and $\lambda_1 < \lambda_3$, it is reasonable to expect q_1^G to touch 0 before q_3^G and end up with a situation similar to three nodes on a line. This intuition comes from the fact that the derivative of q_1^G is smaller than the one of q_3^G. If $q_1^0 = q_3^0$ and $\lambda_1 = \lambda_3$ because of the strong symmetries in the network, since q_1^G and q_3^G have the same derivatives, we expect them to touch 0 at the same time. In order to keep the exit rate at each queue positive, one group of two oposite nodes cannot approach 0 without the other.

6.1.2 Homogenization when touching 0

The major flaw in the homogenization result proved in this thesis is that the upper bound goes to infinity when queue lengths are too small. In order to prove convergence for any finite time horizon, it would be helpful to have a homogenization result that is more efficient when queue lengths become small. As seen in Remark 4.8, we can prove homogenization up to the time $Q(Nt)$ becomes smaller than $N^{\max(a\beta+1/2, \frac{2a\beta}{2+a})}$. This is not sufficient for the example from Section 4.2.4 with three nodes on a line because in order to have all queues with non-zero service rate, it is required to have at least one queue of order of magnitude smaller than $\sqrt{N}$.

6.2 Generalizing the Heavy traffic result

6.2.1 Additional time scales, State space collapse

When the arrival rates are critical, we prove at the end of Chapter 4 that the solution to (4.2) converges as $t \to +\infty$ to a state where input rates and exit rates are equal. This is obviously not true in any interference topology. Recall the example of three nodes on a line from Section 4.2.4 where the outer queues will stay at 0. This is due to the difference in the size of independent sets of the interference graph. If for instance every node is in an independent set of maximum size, does there exist a configuration of queue lengths q_{eq} where $\lambda = \bar{\pi}^{q_{\text{eq}}}$? When the arrival rates are critical, we have

$$\sum_{v \in V} \lambda_v = \sum_{v \in V} \bar{\pi}^q(v) \text{ for any } q \in \mathbb{R}_+^V \setminus \{\mathbf{0}\}$$

so the sum of coordinates evolves on a slower time scale than individual queue lengths but it is unclear if there exists $q_{\text{eq}} \in (0, +\infty)^V$ such that

$$\lambda = \bar{\pi}^{q_{\text{eq}}}.$$

When they exist, are such states always attractive stable points for the fluid limits?

For any interference graph, recall $\Lambda^*(S(G))$ interior of the convex hull of $S(G)$. First, if $\lambda \in (1-\epsilon)\Lambda^*(S(G))$, we expect the queue lengths to reach $\mathbf{0}$ and stay absorded. By (4.4) and the discussion after, second if $\lambda \in (1+\epsilon)\Lambda^*(S(G))$, it is reasonable to expect $s_1(Q(N^b t))$ to be of order of magnitude N^b so if we renormalize the process of queue lengths with N in space, the limit should be infinite as soon as $b > 1$. If λ is on the border of the capacity region, with a reasoning similar to (4.4), the sum of coordinate should evolve very slowly. We saw in Chapter 5 the behavior of the queue lengths when the arrival rates are on the border of $\Lambda^*(S(G))$. In this configuration, the fluid limit converges to a deterministic value as $t \to +\infty$. The limit is a state such that arrival rates and homogenized departure rates are equal and this state is attractive. For fluid limits, we only need to consider $\pi_\infty^q = \lim_{N \to +\infty} \pi^{Nq}$ the first order approximation. As a probability measure, $\pi_\infty^q \in \Lambda^*(S(G))$. Whenever there exists q such that $\lambda_v = \pi_\infty^q(\sigma_v = 1)$, we can expect the fluid limit to converge to such a state. Since this happens in the case of a complete interference graph, we looked at a faster time scale to see the evolution of the total number of request in the system. If

$$I = \left\{ q \in \mathbb{N}^V,\ \lambda_v = \pi_\infty^q(\sigma_v = 1) \right\}$$

is attractive for the fluid limits, any fluid limit starting in I should remain in I but could potentially still evolve. There is a sequence of time scales N^{1+ka} for more and more precise approximations of the invariant measure that could entail variation for the limiting ODE for $k = 0, \dots, \Upsilon - 1$. For $k > 0$, the N^{1+ka} time scale makes it so that the schedules of size greater than $\Upsilon - k$ have a non-negligible influence on the dynamic of queue lengths. The last time scale $N^{1+a\Upsilon}$ gives the influence of idle time on the dynamic of queue lengths. Depending on the topology of the graph, the sum of coordinates might evolve in the critical case for time scales larger than N.

Let's illustrate this discussion with the example from Figure 6.1. For any $q \in \mathbb{R}_+^4$,

and $N < +\infty$,

$$\pi^{Nq}(\sigma_1 = 1) = \frac{(Nq_1+1)^a + (Nq_1+1)^a(Nq_3+1)^a}{1+(Nq_1+1)^a(Nq_3+1)^a + (Nq_4+1)^a(Nq_2+1)^a + \sum_{v\in V}(Nq_v+1)^a}$$
$$= \frac{(q_1+N^{-1})^a(q_3+N^{-1})^a}{(q_1+N^{-1})^a(q_3+N^{-1})^a + (q_4+N^{-1})^a(q_2+N^{-1})^a + O(N^{-a})} \tag{6.1}$$
$$+ \frac{(q_1+N^{-1})^a}{N^a((q_1+N^{-1})^a(q_3+N^{-1})^a + (q_4+N^{-1})^a(q_2+N^{-1})^a + O(N^{-a}))}. \tag{6.2}$$

On the fluid scale, the influence of (6.2) in the invariant measure vanishes as $N \to +\infty$. When $\lambda_1 = \lambda_3 = 1-\lambda_2 = 1-\lambda_4$, the expected state space collapse is given by

$$I = \left\{ q \in \mathbb{R}_+^V,\ \frac{(q_1q_3)^a}{(q_1q_3)^a + (q_2q_4)^a} = \lambda_1 \right\}.$$

For any value of $q_1, q_2 > 0$ it is possible to find q_3, q_4 such that $(q_v)_{v=1,\dots,4} \in I$ so the manifold is actually two dimensional. The manifold should also be attractive as well: if for instance $q \in \mathbb{R}_+^V$ is such that q_1q_3 is too small compared to the value it should be to be on the manifold, the solution to the fluid ODE with coordinates in configuration q has q_1^* and q_3^* decreasing. For any $b > 0$, assuming homogenization, the martingale problem for queue 1 on the time scale N^b gives

$$\frac{Q_1(N^bt)}{N} \approx \frac{Q_1(0)}{N} + N^{b-1}\int_0^t \lambda_1 - \left(\frac{Q_1(N^bs)Q_3(N^bs)}{N}\right)^a + O(N^{b-1-a}) + M_1^N(t).$$

As explained before, after the fluid time scale $b = 1$, the next time scale of interest is $b = 1+a$. On this time scale, the part of the invariant measure given in (6.2) will play a role in the dynamic. On the time scale $b = 1+a$, let

$$Q^N(t) = \frac{Q(N^{1+a}t)}{N}.$$

The dynamic of Q_1^N is given by

$$Q_1^N(t) \approx Q_1^N(0) + N^a\int_0^t \lambda_1 - \frac{(Q_1^N(s)Q_3^N(s))^a}{(Q_1^N(s)Q_3^N(s))^a + (Q_2^N(s)Q_4^N(s))^a}\mathrm{d}s + M_1^N(t)$$
$$+ \int_0^t \frac{Q_1^N(s)^a}{(Q_1^N(s)Q_3^N(s))^a + (Q_2^N(s)Q_4^N(s))^a}\mathrm{d}s$$

The only possibility for the limit of Q^N to be a càdlàg function is if Q^N is close to I because of the N^a in front of the integral. In order to understand the importance of this time scale, consider $Q_1^N - Q_3^N$: the martingale problem gives

$$Q_1^N(t) - Q_3^N(t) \approx Q_1^N(0) - Q_3^N(0) + \int_0^t \frac{(Q_3^N(s))^a - (Q_1^N(s))^a}{(Q_1^N(s)Q_3^N(s))^a + (Q_2^N(s)Q_4^N(s))^a}\mathrm{d}s + M_{1,3}^N(t).$$

Because of this, the difference between nodes that are in the same maximal stable set tend to become equal as $t \to +\infty$. Let's define the manifold $\widetilde{I} \subset I$ such that $q \in \widetilde{I}$ if and only if $q \in I$ and $q_1 - q_3 = q_2 - q_4 = 0$. If $Q^N \xrightarrow{\mathbb{P}} \widetilde{q}$ uniformly over compact time sets and $\widetilde{q}$ is continuous, the influence of the stable sets of size 1

suggests that

$$|\widetilde{q}_1(t) - \widetilde{q}_3(t)| \to 0 \text{ as } \to +\infty.$$

The additional time scales due to successive approximations of the invariant measure seem to lead to state space collapses on manifold of decreasing dimensions.

6.2.2 Diffusive time scale

Our second time scale is not the usual N^2 from heavy traffic results. In fact, with this time scale because N^{1+a} is slower than N^2, the queueing process do not "diffuse". The martingale terms in the evolution of queue lengths all vanish as $N \to +\infty$. What happens on the N^2 time scale is linked to the asymptotic behavior of the ODE for the fluid limits. If $s_1(\lambda) = 1$ in a complete interference graph, $\bar{q}(t) \to +\infty$ as $t \to +\infty$ so $Q(N^2 t)$ may not be $O(N)$. The standard heavy traffic assumption is that there exists λ on the border of the stability region, $\gamma \in \mathbb{R}_+^V$ and $(\rho_N)_{N\in\mathbb{N}}$ a sequence converging to 0 such that

$$\frac{1}{\rho_N}\left(\lambda_v - \lambda_v^N\right) \to \gamma_v.$$

For any $b > 0$, looking at the generator for $\frac{Q_v(N^b \cdot)}{N}$ in a complete interference graph, assuming the distance to the state space collapse is N^{-a} (which can be justified formally), we get

$$\begin{aligned} &N^b \lambda_v^N \left(f\left(q + \frac{e^v}{N}\right) - f(q)\right) + N^b\left(\lambda_v - N^{-a}\right)\left(f\left(q - \frac{e^v}{N}\right) - f(q)\right) \\ &= N^b(\lambda_v - \rho_N\gamma_v)\left(f\left(q + \frac{e^v}{N}\right) - f(q)\right) + N^b\left(\lambda_v - N^{-a}\right)\left(f\left(q - \frac{e^v}{N}\right) - f(q)\right). \end{aligned}$$

Using a second order Taylor expansion with integral remainder,

$$f(q + \frac{e^v}{N}) - 2f(q) + f(q - \frac{e^v}{N}) = \frac{1}{N^2}\int_0^1 \partial_{v^2} f(q + \frac{ue^v}{N}) + \partial_{v^2} f(q - \frac{ue^v}{N})\mathrm{d}u.$$

If the second derivative of f is bounded and $b < 2$, the generator can be approximated by

$$N^b \rho_N \gamma_v \left(f\left(q + \frac{e^v}{N}\right) - f(q)\right) + N^{b-a}\left(f\left(q - \frac{e^v}{N}\right) - f(q)\right).$$

Thus it is reasonable to take $b = -a - 1$ and $N^{b-1}\rho_N = O(1)$. Recall from 5.1.3 that if $\gamma_v > 0$ for all $v \in V$ the corresponding ODE for queue lengths has a unique stable point q^s given by $q^s \in I$ and $s_1(q^s) = (\frac{\mu}{s_1(\gamma)})^{1/a}$ with $\mu = (\sum_{v\in V} \lambda_v^{1/a})^a$. When starting with q^s, the limit is constant. Because this state is attractive, we can expect that $Q(N^{1+\theta}) = q^s$ as long as $a < \theta < 1$. When $\theta = 1$, $N^{1+\theta}$ is actually the usual diffusion time scale. On the N^2 time scale, we expect the limit to be some kind of semi-martingale reflected Brownian motion with a drift towards the attracting set. For any a, Corollary 3.7 is not enough to prove homogenization occurs on this time scale so we need to improve the homogenization result as well. After the result from Chapter 5, it is natural to pose a conjecture on the N^2 time scale: let $\bar{Q}^N(t) = \frac{Q(N^2 t)}{N}$.

Conjecture 6.3
Assume $Q^N(0) \to q^0 \in (0, +\infty)$, G is a complete interference graph and $a < 1/2$. Then $\bar{Q}^N$ converges to a Gaussian process with state dependent drift and variance.

6.3 Generalizing the homogenization result

Looking at Chapter 3 and the proof of the homogenization result, we see that it is easy to generalize it to other types of queueing models. Assume that L is the irreducible generator of a queueing model such that for any $q \in \mathbb{N}^V$, and $\sigma \in S^0$, with S^0 finite,

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

Assume that L_{f}^q has a unique invariant measure π^q. The bounds we obtain on the regularity of solution to Poisson equations in their parameters are essentially functions of L_{f}^q, its invariant measure and its Log-Sobolev constant:

- The log-Sobolev constant
$$\alpha^q,$$
- the derivative of the transition rates of the fast process
$$d_l(q) \coloneqq \max_{v\in V,\sigma,\sigma'\in S^0} \partial_v L_{\mathrm{s}}^q(\sigma,\sigma');$$
- and the derivative of its invariant measure
$$d_\pi(q) \coloneqq \max_{v\in V,\sigma\in S^0} \partial_v \pi^q(\sigma).$$

Redefining a localization set U and denoting $\phi_g(q,\cdot)$ the solution to the Poisson equation

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

we get that for any $q \in U$,

$$\max_{\sigma\in S^0} |\phi_g(q \pm e^v,\sigma) - \phi_g(q,\sigma)| \leqslant C \|g\|_\infty \left(\|\alpha^q\|_{\infty,U} \|d_\pi\|_{\infty,U} + \|\alpha^q\|_{\infty,U}^2 \|d_l\|_{\infty,U} \right).$$

If the size of jumps are bounded, with this we can provide the next bound:

$$\begin{aligned}
\mathbb{E}\left[\int_0^{t\wedge\bar{\tau}^U} \left(L_{\mathrm{s}}^{\sigma(s)} - \pi^{Q(s)}[L_{\mathrm{s}}^{\sigma(s)}] \right)[f](Q(s))\mathrm{d}s \right] &\leqslant Ct \max_{v\in V} \|\partial_v g\|_{\infty,U} \|\alpha^q\|_{\infty,U} \|d_\pi\|_{\infty,U} \\
&+ Ct \max_{v\in V} \|\partial_v g\|_{\infty,U} \|\alpha^q\|_{\infty,U}^2 \|d_l\|_{\infty,U} \\
&+ C\sqrt{t} \|\alpha^q\|_{\infty,U} \left(\max_{v\in V} \|\partial_v g\|_{\infty,U} + \sqrt{t} \max_{v,w\in V} \left\|\partial_{v,w}^2 g\right\|_{\infty,U} \right),
\end{aligned}$$

which when scaled properly, can give a homogenization result. We now informally present two potential future research projects expanding on this homogenization argument.

6.3.1 Metastability phenomenon

The hypothesis on the generator is a bit restrictive: because of the separation between the two components, we are not able to handle cases where σ and q jump at the same time. For instance in wireless networks, we want a queue to deactivate at the same time as it finishes its last job. Such transitions are not allowed with our decomposition because the schedule and queue lengths cannot jump at the same time. If this type of transition does not occur often, our result should still holds, provided this type of jump does not make the invariant measure change too much. If those simultaneous jumps significantly change the invariant measure (for instant by being a bottleneck for transitions between two regions of the state space), some interesting metastable phenomena can occur. For instance, let (Q^N, σ^N) be a sequence of Markov processes whose generators are given for any $q, \sigma \in \mathbb{N}^V \times S^0$, by

$$L^N[f](\sigma, q) \coloneqq NL_{\mathrm{f}}^q[f(\cdot, q)](\sigma) + L_{\mathrm{s}}^\sigma[f(\sigma, \cdot)](q) + L_{\mathrm{m},N}[f](\sigma, q).$$

Assume that L_{f}^q is reducible and $L_{\mathrm{m},N}$ contains transitions between fractions of the state space that would not be connected otherwise with all rates uniformly bounded and bounded away from zero. Assume the state space of σ can be decomposed in irreducible classes $S_1, \dots, S_k$ such that L_{f}^q restricted to each S_i is irreducible. For each $i \leqslant k$, there exists a unique invariant measure π_i invariant for L_{f}^q. If there is a time scale separation for the Markov process of generator $NL_{\mathrm{f}}^q[f(\cdot, q)](\sigma) + L_{\mathrm{s}}^\sigma[f(\sigma, \cdot)](q)$ on all S_i, $1 \leqslant i \leqslant k$, the dynamic we can expect for Q^N, σ^N is a mixture between the homogenized processes on each S_i. As $N \to +\infty$, σ^N will randomly switch between fractions of S^0 at a bounded rate and thus the limiting dynamic randomly switches its "driving invariant measure" depending on which fraction of the state space the fast process is.

6.3.2 Large deviations

The convergence to a deterministic process indicates that the distribution of the trajectory concentrates on a single element of the space of continuous trajectories. It is also interesting to understand just how the distribution concentrates. One natual future research consists in establishing a large deviation result for QB-CSMA in any of the two regimes considered in this thesis. It would be interesting as well to have a better understanding of large deviation results for homogenized process in a fully coupled model. Freidlin and Wentzell proved in [FW12] a homogenization and a large deviation result for the solution to the problem

$$\begin{cases} (X^\epsilon)'(t) = b(X^\epsilon(t), \zeta^\epsilon(t)), & X^\epsilon(0) = x \\ (\zeta^\epsilon)'(t) = \epsilon^{-1} b_0(X^\epsilon(t), \zeta^\epsilon(t)), & \zeta^\epsilon(0) = y \end{cases}, \tag{6.3}$$

or

$$\mathrm{d}\zeta^\epsilon(t) = b(X^\epsilon(t), \zeta^\epsilon(t))\mathrm{d}t + \frac{1}{\sqrt{\epsilon}} g(\zeta^\epsilon(t))\mathrm{d}W(t),$$

with $W(t)$ a Brownian motion, under the conditions of the existence of $\bar{b}$ such that for any $x \in \mathcal{X}$ and $\delta > 0$,

$$\sup_{t \geqslant 0} \lim_{T \to +\infty} \mathbb{P}\left(\left| \frac{1}{T} \int_t^{t+T} b(x, \zeta(s))\mathrm{d}s - \bar{b}(x)\mathrm{d}s \right| > \delta \right) = 0,$$

and of a function H such that for any step-functions ϕ_s and α_s,

$$\lim_{\epsilon\to 0} \epsilon \log\left(M \exp\left(\frac{1}{\epsilon}\int_0^T \langle \alpha_s, b(\phi_s, \zeta(s/\epsilon))\rangle \,\mathrm{d}s\right)\right) = \int_0^T H(\phi_s, \alpha_s)\mathrm{d}s,$$

with M a numerical constant. They prove the existence of a function D on the space of continuous functions such that for any $A \subset C((0,T), \mathbb{R}_+^n)$ not containing the limit,

$$\lim_{\epsilon\to 0} -\epsilon \log(\mathbb{P}\left(X^\epsilon \in A\right)) = \sup_{f\in A} D(f).$$

In a time scale separation framework, both components could deviate from their theoretical limit. Is it possible to identify the influence of the fast and the slow variables in the rate function if a large deviation result hold. Does the mixing time, the log-Sobolev constant and other quantities related to the fast variable play a role in the rate function or even the speed at which $\mathbb{P}\left(X^\epsilon \in A\right) \to 0$? For instance for QB-CSMA with polynomial rates, how does the parameter a influence the result.

Appendix A

Asymptotic approximation of service rates

Proof of Lemma 2.4: Fix $v \in V$, for any $q \in \frac{1}{N}\mathbb{N}^V$ such that $C_- \leqslant \min_w q_w \leqslant \max_w q_w \leqslant C_+$, we get

$$\begin{aligned}\pi^{Nq}(\sigma_v = 1) &= \frac{\sum_{\sigma\in S} \sigma_v \prod_{w\in V} (Nq_w+1)^{a\sigma_w}}{1+\sum_{\sigma\in S}\prod_{w\in V}(Nq_w+1)^{a\sigma_w}} \\ &= \frac{\sum_{\rho\in S^*} \rho_v \prod_{w\in V}(q_w+\frac{1}{N})^{a\rho_w} + \sum_{\sigma\in S\setminus M} N^{a(|\sigma|-v)}\sigma_v \prod_{w\in V}(q_w+\frac{1}{N})^{a\sigma_w}}{N^{-av}+\sum_{\sigma\in S} N^{a(|\sigma|-v)}\prod_{w\in V}(q_w+\frac{1}{N})^{a\sigma_w}}\end{aligned}$$

Since every coordinate of q is bounded by M and bounded away from 0,

$$\frac{\sum_{\sigma\in S\setminus M} N^{a(|\sigma|-v)}\sigma_v \prod_{w\in V}(q_w+\frac{1}{N})^{a\sigma_w}}{N^{-av}+\sum_{\sigma\in S} N^{a(|\sigma|-v)}\prod_{w\in V}(q_w+\frac{1}{N})^{a\sigma_w}} \leqslant N^{-a}\frac{(2C_+)^{a(\Upsilon-1)}}{C_-^{a\Upsilon}}.$$

If v is not in a stable set of maximum size, $\pi^{Nq}(\sigma_v=1) \leqslant CN^{-a}$ and $\overline{\pi}^q(v)=0$ so the convergence holds. Assume now that there is $\rho \in S^*$ such that $\rho_v = 1$. With the previous computation, we know that

$$|\pi^{Nq}(\sigma_v=1) - \overline{\pi}^q(v)| \leqslant \left| \frac{\sum_{\rho\in S^*}\rho_v\prod_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}}{N^{-av}+\sum_{\sigma\in S}N^{a(|\sigma|-v)}\prod_{w\in V}(q_w+\frac{1}{N})^{a\sigma_w}} - \frac{\sum_{\rho\in S^*}\rho_v\prod_{w\in V}q_w^{a\rho_w}}{\sum_{\rho\in S^*}\prod_{w\in V}q_w^{a\rho_w}}\right| + CN^{-a}.$$

So we are left to deal with

$$e_v^N(q) \coloneqq \left| \frac{\sum_{\rho\in S^*}\rho_v\prod_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}}{N^{-av}+\sum_{\sigma\in S}N^{a(|\sigma|-v)}\prod_{w\in V}(q_w+\frac{1}{N})^{a\sigma_w}} - \frac{\sum_{\rho\in S^*}\rho_v\prod_{w\in V}q_w^{a\rho_w}}{\sum_{\rho\in S^*}\prod_{w\in V}q_w^{a\rho_w}}\right|.$$

In order to obtain convergence, we now add and subtract the intermediary quantity:

$$\frac{\sum\limits_{\rho\in S^*}\rho_v\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}}{\sum\limits_{\rho\in S^*}\prod\limits_{w\in V}q_w^{a\rho_w}}.$$

Using the triangular inequality we get

$$\begin{aligned}
e_v^N(q) \leqslant & \left|\frac{\sum\limits_{\rho\in S^*}\rho_v\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}}{N^{-a\upsilon}+\sum\limits_{\sigma\in S}N^{a(|\sigma|-\upsilon)}\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\sigma_w}}-\frac{\sum\limits_{\rho\in S^*}\rho_v\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}}{\sum\limits_{\rho\in S^*}\prod\limits_{w\in V}q_w^{a\rho_w}}\right| \\
& + \left|\frac{\sum\limits_{\rho\in S^*}\rho_v\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}}{\sum\limits_{\rho\in S^*}\prod\limits_{w\in V}q_w^{a\rho_w}}-\frac{\sum\limits_{\rho\in S^*}\rho_v\prod\limits_{w\in V}q_w^{a\rho_w}}{\sum\limits_{\rho\in S^*}\prod\limits_{w\in V}q_w^{a\rho_w}}\right| \\
\leqslant & \left|\frac{\sum\limits_{\rho\in S^*}\rho_v\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}}{\sum\limits_{\rho\in S^*}\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}+\sum\limits_{\sigma\in S\setminus M}N^{a(|\sigma|-\upsilon)}\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\sigma_w}}-\frac{\sum\limits_{\rho\in S^*}\rho_v\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}}{\sum\limits_{\rho\in S^*}\prod\limits_{w\in V}q_w^{a\rho_w}}\right| \\
& + \frac{1}{\sum\limits_{\rho\in S^*}\prod\limits_{w\in V}q_w^{a\rho_w}}\sum_{\rho\in S^*}\rho_v\left|\prod_{w\in V}q_w^{a\rho_w}(1+\frac{1}{q_wN})^{a\rho_w}-\prod_{w\in V}q_w^{a\rho_w}\right| \\
\leqslant & \sum_{\rho\in S^*}\rho_v\left|\frac{\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}}{\sum\limits_{\rho\in S^*}\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}+\sum\limits_{\sigma\in S\setminus M}N^{a(|\sigma|-\upsilon)}\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\sigma_w}}-\frac{\prod\limits_{w\in V}(q_w+\frac{1}{N})^{a\rho_w}}{\sum\limits_{\rho\in S^*}\prod\limits_{w\in V}q_w^{a\rho_w}}\right| \\
& + \frac{\sum\limits_{\rho\in S^*}\rho_v\prod\limits_{w\in V}q_w^{a\rho_w}\left|\prod\limits_{w\in V}(1+\frac{1}{q_wN})^{a\rho_w}-1\right|}{\sum\limits_{\rho\in S^*}\prod\limits_{w\in V}q_w^{a\rho_w}}
\end{aligned}$$

The idea to conclude the proof is to write Taylor expansions for

$$\sum_{w\in V}a\rho_w\log(1+\frac{t}{q_w})$$

and

$$\frac{1}{t^a+\sum\limits_{\rho\in S^*}\prod\limits_{w\in V}(q_w+t)^{a\rho_w}},$$

for t in a neighbourhood of 0, when $\min_w q_w > C_-$. □

Appendix B

Positivity of queue lengths for small times

Contents

B.1 General remarks 121
B.2 Main steps of the proof 122
B.3 Proofs using coupling 128

The only thing left is to prove Lemma 4.14. It will be helpful to accurately describe a coupling for the different activation time/duration. We present in the last part of this appendix an exact construction of the process. This construction will enable us to formally prove that all queues are positive for positive times.

B.1 General remarks

For all practical matters, it is possible to replace the Poisson processes for arrivals and departures by their "compensated" quantity. Recall Proposition 2.1 and Lemma 4.10. Because of theses results, regardless of homogenization, for any $T < +\infty$ and $v \in V$,

$$\mathbb{P}\left(\sup_{t\leqslant T}\left|Q_v^N(t) - Q_v^N(0) - \lambda_v t + \int_0^t \sigma_v^N(s)\mathbb{1}_{Q_v^N(s)>0}\mathrm{d}s\right| > N^{-1/3}\right)$$
$$= \mathbb{P}\left(\sup_{t\leqslant T}|M_v^N(t)| > N^{-1/3}\right) \leqslant N^{1/3-1/2} \to 0 \text{ as } N \to +\infty. \quad \text{(B.1)}$$

In particular, since

$$\sup_{v\in V,\, t\leqslant T} Q_v^N(0) + \lambda_v t - \int_0^t \sigma_v^N(s)\mathbb{1}_{Q_v^N(s)>0}\mathrm{d}s \leqslant \left\|Q^N(0)\right\|_\infty + \|\lambda\|_\infty T,$$

we get that for any $v \in V$,

$$\mathbb{P}\left(\sup_{t\leqslant T} Q_v^N(t) > \left\|q^0\right\|_\infty + \|\lambda\|_\infty T + N^{-1/3}\right) \to 0 \text{ as } N \to +\infty.$$

All of the next discussion will be in the event

$$\left\{\sup_{t\leqslant T}|M^N(t)| \leqslant N^{-1/3}\right\},$$

and the $N^{-1/3}$ term will be omitted in the computations. This choice does not change the result and is only made to alleviate notations. For the previous probability, we would for instance state that

$$\mathbb{P}\left(\sup_{t\leqslant T} Q_v^N(t) > \left\|q^0\right\|_\infty + \|\lambda\|_\infty T\right) \to 0 \text{ as } N \to +\infty.$$

The proof of Lemma 4.14 will be provided in a series of lemmas using a coupling argument. We will only deal with the case where $\left\|q^0\right\|_\infty > 0$ because the case $s_1(\lambda) > 1$ can be handled similarly: if $\left\|q^0\right\|_\infty = 0$ and $s_1(\lambda) > 1$, for any $t > 0$ there is a queue bounded below. Since $s_1(\sigma^N(t)) \leqslant 1$, by summing over coordinates in (B.1),

$$\mathbb{P}\left(s_1(Q^N(t)) \geqslant s_1(Q^N(0)) + (s_1(\lambda) - 1)t - nN^{-1/3}\right) \to 1 \text{ as } N \to +\infty.$$

In order for $s_1(q)$ to be greater than ϵ, there is at least one v such that $q_v \geqslant \frac{\epsilon}{n}$. Necessarily,

$$\mathbb{P}\left(\max_{v\in V} Q_v^N(t) \geqslant \frac{(s_1(\lambda) - 1)t}{n}\right) \to 1 \text{ as } N \to +\infty.$$

After any positive time, there is at least one queue with bounded below queue length. Recall the hitting time

$$T_-^\epsilon(Q_{v^0}^N) = \inf\left\{t > 0,\, Q_{v^0}^N(t) \geqslant \epsilon\right\}.$$

In terms of hitting times, if $s_1(\lambda) > 1$ and $\left\|q^0\right\|_\infty = 0$,

$$\mathbb{P}\left(T_-^\epsilon(\max_v Q_v^N) \leqslant \frac{n}{s_1(\lambda) - 1}\epsilon\right) \to 1 \text{ as } N \to +\infty.$$

With the strong Markov property, the reasoning in the next sections can be handled given which queue is above ϵ at time $T_-^\epsilon(\max_v Q_v^N)$ on the shifted process, and then integrate over $v \in V$.

B.2 Main steps of the proof

Introduce

$$V' \coloneqq \{q \in \mathbb{R}_+^V,\, q_v^0 > 0\} \text{ and } \epsilon_0 \coloneqq \min_{v\in V'} q_v^0.$$

Lemma 4.14 will be proved by induction on the nodes not in V'. From now on assume $V' \neq \emptyset$ and $V \setminus V' \neq \emptyset$. Fix $v^0 \in V \setminus V'$ and introduce

$$K_{\text{time}} := \frac{8^{1+a} n \|q^0\|_\infty^a}{\min_v \lambda_v \epsilon_0^a}.$$

Without loss in generality assume that $\min_v \lambda_v \leqslant 1$. Otherwise, on the fluid scale, all queue lengths are increasing and any level ϵ is trivially reached in a time linear in ϵ. With the convention $1/0 = +\infty$, let $\epsilon_1 > 0$ be such that

$$\epsilon_1 < \frac{\epsilon_0}{2} \min \left[1, \frac{1}{2^{1+1/a}} \left(\min_v \lambda_v \right)^{1/a}, \frac{2 \left\| q^0 \right\|_\infty}{\|\lambda\|_\infty \epsilon_0 K_{\text{time}}}, \frac{1}{K_{\text{time}}(1 - \min_v \lambda_v)} \right].$$

In this section, we first present the technical Lemma B.1 that allows us to prove Lemma 4.14. It provides a bound on crossing time of ϵ_1 for $Q^N_{v^0}$. The probability that the hitting time exceeds a linear function of ϵ_1 goes to 0 as $N \to +\infty$. If the hitting time is small enough, queues that are in V' do not have time to reach 0 on the fluid scale before $Q^N_{v^0}$ reaches ϵ_1. After briefly justifying Lemma B.1, we present the proof of Lemma 4.14 using Lemma B.1. The proof relies on an induction argument on the nodes in $V \setminus V'$. Next, we present some auxiliary lemmas for the proof of Lemma B.1. Once the three auxiliary lemmas are stated the proof of Lemma B.1 is provided using a reasoning on events. Lemmas B.2, B.3 and B.4 rely on a coupling argument.

The time $\epsilon_1 K_{\text{time}}$ is constructed in such a way that because of (B.1), by definition of ϵ_0,

$$\min_{v \in V'} Q^N_v(t) \geqslant \epsilon_0 - (1 - \min_{v \in V} \lambda_v)t + M^N_v(t).$$

Since

$$\epsilon_1 K_{\text{time}} < \frac{\epsilon_0}{2(1 - \min_{v \in V} \lambda_v)},$$

we get

$$\mathbb{P}\left(\inf_{v \in V', s \leqslant \epsilon_1 K_{\text{time}}} Q^N_v(s) \geqslant \frac{\epsilon_0}{2}, \sup_{v \in V,\, s \leqslant \epsilon_1 K_{\text{time}}} Q^N_v(s) \leqslant \left\| q^0 \right\|_\infty + \|\lambda\|_\infty \epsilon_1 K_{\text{time}} \right) \to 1 \text{ as } N \to +\infty.$$

Introduce the events

$$E^N_-(t) := \left\{ \inf_{v \in V', s \leqslant t} Q^N_v(s) \geqslant \frac{\epsilon_0}{2} \right\}, \tag{B.2}$$

and

$$E^N_+(t) := \left\{ \sup_{s \leqslant t,\, v \in V} Q^N_v(s) \leqslant 2 \left\| q^0 \right\|_\infty \right\}. \tag{B.3}$$

Since $\|\lambda\|_\infty \epsilon_1 K_{\text{time}} < \left\| q^0 \right\|_\infty$,

$$\mathbb{P}\left(E^N_-(\epsilon_1 K_{\text{time}}) \cap E^N_+(\epsilon_1 K_{\text{time}}) \right) \to 1 \text{ as } N \to +\infty. \tag{B.4}$$

Lemma B.1
With the parameters defined before this lemma,

$$\mathbb{P}\left(\min_{v \in V' \cup \{v^0\}} Q^N_v(T^{\epsilon_1}_-(Q^N_{v^0})) \geqslant \epsilon_1 \right) \to 1 \text{ as } N \to +\infty,$$

and

$$\mathbb{P}\left(T_{-}^{\epsilon_1}(Q_{v^0}^N) < \epsilon_1 K_{\text{time}}\right) \to 1 \text{ as } N \to +\infty.$$

The proof of this lemma relies heavily on a coupling argument that will be explained in the next section. The general idea is to observe the evolution of node v^0 over $N^{1-a}K_{\text{act}}$ activations of v^0. The proof of Lemma B.1 will be split in three auxiliary lemmas:

- The first lemma ensures that the large number of activations can occur in a time smaller than $N\epsilon_1 K_{\text{time}}$ so that a node that was in V' does not have time to reach 0 on the fluid scale. In addition bounding this time ensures that all queue lengths will remain bounded before it. This is necessary to be able to iterate the procedure and prove that the crossing time of a level ϵ converges to 0 as $\epsilon \to 0$.

- The second auxiliary lemma states that over this large number of activations of v^0, the fraction of the time this node was active is smaller than half its arrival rate, at least until it reaches ϵ_1, provided all queues in V' keep their queues lengths above $\frac{\epsilon_0}{2}$.

- The third lemma states that this large number of activations occurs in a time of order of magnitude N. This means that over a non-negligible period of time queue v^0 will be increasing, at least until it reaches ϵ_1.

Before formally stating these lemmas, we provide the proof of Lemma 4.14 using Lemma B.1.

Proof of Lemma 4.14. By Lemma B.1,

$$\mathbb{P}\left(T_{-}^{\epsilon_1}(Q_{v^0}^N) < \epsilon_1 K_{\text{time}}\right) \to 1 \text{ as } N \to +\infty,$$

so $T_{-}^{\epsilon_1}(Q_{v^0}^N)$ is a stopping time asymptitocally finite. By the strong Markov property, the process $Q^N(\cdot + T_{-}^{\epsilon_1}(Q_{v^0}^N))$ is a Markov process with the same dynamic as Q^N and starting point $Q^N(T_{-}^{\epsilon_1}(Q_{v^0}^N))$. Again by Lemma B.1, we get

$$\mathbb{P}\left(\min_{v \in V' \cup \{v^0\}} Q^N(T_{-}^{\epsilon_1}(Q_{v^0}^N)) \geqslant \epsilon_1\right) \to 1 \text{ as } N \to +\infty,$$

and

$$\mathbb{P}\left(\left\|Q^N(T_{-}^{\epsilon_1}(Q_{v^0}^N))\right\|_\infty > 2\left\|q^0\right\|_\infty\right) \to 0 \text{ as } N \to +\infty,$$

because $T_{-}^{\epsilon_1}(Q_{v^0}^N)) \leqslant \epsilon_1 K_{\text{time}}$ and $\epsilon_1 K_{\text{time}} < \frac{\left\|q^0\right\|_\infty}{\|\lambda\|_\infty}$.

As long as $V \setminus V' \cup \{v^0\} \neq \emptyset$, it is possible to iterate the procedure with V' replaced by $V' \cup \{v^0\}$ and ϵ_0 replaced by ϵ_1 for the shifted process $Q^N(\cdot + T_{-}^{\epsilon_1}(Q_{v^0}^N))$. Since there is a finite number of nodes, the time it takes for the minimum of queue lengths to reach a positive threshold ϵ vanishes with probability close to one as $N \to +\infty$ and $\epsilon \to 0$.

Recall the definition of K_{time}:

$$K_{\text{time}} = \frac{4^{1+a} \left\|q^0\right\|_\infty}{n\epsilon_0^a \left\|\lambda\right\|_\infty}.$$

When iterating the procedure at step $k > 0$, we need to chose a constant:

$$K_{\text{time}}(k) = \frac{8^{1+a} n^{1+(k-1)a} \|q^0\|_\infty^a}{\min_v \lambda_v \epsilon_{k-1}^a} \leqslant \frac{8^{1+a} n^{1+(n-1)a} \|q^0\|_\infty^a}{\min_v \lambda_v \epsilon_k^a},$$

with a sequence $(\epsilon_k)_{k=0,\dots,|V\setminus V'|}$ such that: $\epsilon_0 = \min_{v\in V'} q_v^0$,

$$\epsilon_{k+1} < \frac{\epsilon_k}{2} \min\left[1,\ \frac{1}{2^{1+1/a}} \left(\min_v \lambda_v\right)^{1/a},\ \frac{2^{1+k} \left\|q^0\right\|_\infty}{\|\lambda\|_\infty \epsilon_k K_{\text{time}}(k)},\ \frac{1}{K_{\text{time}}(k)(1-\min_v \lambda_v)}\right],$$

and $\epsilon \leqslant \epsilon_{|V\setminus V'|}$, which is possible for ϵ small enough. In this case,

$$\mathbb{P}\left(T_-^\epsilon(Q^N) \geqslant \left(\sum_{k=1}^{|V\setminus V'|} K_{\text{time}}(k)\epsilon_k\right)\right) \to 0 \text{ as } N \to +\infty.$$

When $\epsilon \to 0$, it is possible to choose $\sum_{k=1}^{|V\setminus V'|} \epsilon_k^{1-a}$ as small as desired. □

We now present the three lemmas used to prove Lemma B.1. Their proofs are provided in the next section. Once those three lemmas are formally stated, we present a proof of Lemma B.1 at the end of the section. Introduce the constant:

$$K_{\text{act}} \coloneqq \frac{\epsilon_1 K_{\text{time}}}{n4^{1+a} \|q^0\|_\infty^a},$$

and the stopping time

$$d(N^{1-a}K_{\text{act}}) = \text{ Time of the } N^{1-a}K_{\text{act}}^{th} \text{ deactivation for } v^0,$$

see (B.8) for more details.

Lemma B.2
As $N \to +\infty$,

$$\mathbb{P}\left(d(N^{1-a}K_{\text{act}})) \leqslant N\epsilon_1 K_{\text{time}}\right) \to 1.$$

Consequently,

$$\mathbb{P}\left(E_-^N(\frac{d(N^{1-a}K_{\text{act}})}{N})\right) = \mathbb{P}\left(\inf_{v\in V', Ns\leqslant d(N^{1-a}K_{\text{act}}))} Q_v^N(s) \geqslant \frac{\epsilon_0}{2}\right) \to 1 \text{ as } N \to +\infty,$$

and

$$\mathbb{P}\left(E_+^N(\frac{d(N^{1-a}K_{\text{act}})}{N})\right) = \mathbb{P}\left(\sup_{v\in V,\, Ns\leqslant d(N^{1-a}K_{\text{act}}))} Q_v^N(s) \leqslant 2\left\|q^0\right\|_\infty\right) \to 1 \text{ as } N \to +\infty.$$

The idea is to reason in the event

$$E_+^N(\epsilon_1 K_{\text{time}}) = \left\{ \sup_{v\in V,\, s\leqslant \epsilon_1 K_{\text{time}}} Q_v^N(s) \leqslant 2\left\|q^0\right\|_\infty \right\},$$

and use the fact that its probability goes to 1. When proving that $d(N^{1-a}K_{\text{act}})$ is smaller than $N\epsilon_1 K_{\text{time}}$, only the value of queue lengths before $N\epsilon_1 K_{\text{time}}$ is important. In $E_+^N(\epsilon_1 K_{\text{time}})$, it is possible to bound each activation duration using the coupling described in the next section. The second part of the result is a direct consequence of (B.4).

Introduce the event

$$E_{v^0}^N := \left\{T_-^{\epsilon_1}(Q_{v^0}^N) > d(N^{1-a}K_{\text{act}})\right\} \cap E_-^N(d(N^{1-a}K_{\text{act}})). \tag{B.5}$$

The next lemma uses this event:

Lemma B.3
As $N \to +\infty$, we get

$$\mathbb{P}\left(\int_0^{d(N^{1-a}K_{\text{act}})} \sigma_{v^0}(s)\mathrm{d}s > d(N^{1-a}K_{\text{act}})\frac{\lambda_{v^0}}{2}, E_{v^0}^N\right) \to 0.$$

In particular, because of Lemma B.2

$$\mathbb{P}\left(T_-^{\epsilon_1}(Q_{v^0}^N) \leqslant d(N^{1-a}K_{\text{act}}) \text{ or } Q_{v^0}^N(d(N^{1-a}K_{\text{act}})) \geqslant \frac{\lambda_{v^0}}{2} d(N^{1-a}K_{\text{act}})\right) \to 1 \text{ as } N \to +\infty.$$

This is quite intuitive: as long as queue v^0 is small enough and all queues in V' are large, queue v^0 cannot be active for a large fraction of the time compared to nodes in V'. This entails that $Q_{v^0}^N$ must be increasing during this period until it reaches ϵ_1 with probability close to one. Because of (B.1), we can bound the probability that v^0 is small provided all queues in V' stay above $\frac{\epsilon_0}{2}$. In the first case, Lemma B.2 is enough to prove Lemma B.1. Otherwise, we need to prove that

$$\mathbb{P}\left(d(N^{1-a}K_{\text{act}}) \geqslant \frac{2N\epsilon_1}{\lambda_{v^0}}\right) \to 1 \text{ as } N \to +\infty.$$

This ensures that $Q_{v^0}^N$ has had time to cross the level ϵ_1 before $d(N^{1-a}K_{\text{act}})$.

Lemma B.4
With the constants described at the beginning of the section,

$$\mathbb{P}\left(d(N^{1-a}K_{\text{act}}) < \frac{2N\epsilon_1}{\lambda_{v^0}}, E_-^N(d(N^{1-a}K_{\text{act}}))\right) \to 0 \text{ as } N \to +\infty.$$

In particular, because of Lemma B.2,

$$\mathbb{P}\left(d(N^{1-a}K_{\text{act}}) \geqslant \frac{2N\epsilon_1}{\lambda_{v^0}}\right) \to 1 \text{ as } N \to +\infty.$$

Given these three lemmas, the proof of Lemma B.1 can be based on the study of events whose probability goes to one:

Proof of Lemma B.1: In terms of event,

$$\left\{T_-^{\epsilon_1}(Q_{v^0}^N) \leqslant d(N^{1-a}K_{\text{act}})\right\} \cap \left\{d(N^{1-a}K_{\text{act}}) \leqslant N\epsilon_1 K_{\text{time}}\right\} \subset \left\{T_-^{\epsilon_1}(Q_{v^0}^N) \leqslant N\epsilon_1 K_{\text{time}}\right\}. \tag{B.6}$$

Similarly,

$$\left\{\int_0^{d(N^{1-a}K_{\text{act}})} \sigma_{v^0}(s)\mathrm{d}s \leqslant d(N^{1-a}K_{\text{act}})\frac{\lambda_{v^0}}{2}\right\} \subset \left\{Q_{v^0}^N(\frac{d(N^{1-a}K_{\text{act}})}{N}) \geqslant \frac{\lambda_{v^0}d(N^{1-a}K_{\text{act}})}{2N}\right\},$$

and

$$\left\{Q_{v^0}^N(\frac{d(N^{1-a}K_{\text{act}})}{N}) \geqslant \frac{\lambda_{v^0}d(N^{1-a}K_{\text{act}})}{2N}\right\} \cap \left\{\frac{2N\epsilon_1}{\lambda_{v^0}} \leqslant d(N^{1-a}K_{\text{act}}) \leqslant N\epsilon_1 K_{\text{time}}\right\} \subset \left\{T_-^{\epsilon_1}(Q_{v^0}^N) \leqslant N\epsilon_1 K_{\text{time}}\right\}. \tag{B.7}$$

By (B.6) and (B.7), $\left\{T_-^{\epsilon_1}(Q_{v^0}^N) \leqslant N\epsilon_1 K_{\text{time}}\right\}$ contains

$$\left(\left\{NT^{\epsilon_1(Q_{v^0}^N)} \leqslant d(N^{1-a}K_{\text{act}})\right\} \cup \left\{Q_{v^0}^N(\frac{d(N^{1-a}K_{\text{act}})}{N}) \geqslant \frac{\lambda_{v^0}d(N^{1-a}K_{\text{act}})}{2N}\right\}\right) \cap \left\{\frac{2N\epsilon_1}{\lambda_{v^0}} \leqslant d(N^{1-a}K_{\text{act}}) \leqslant N\epsilon_1 K_{\text{time}}\right\}.$$

In conclusion,

$$\mathbb{P}\left(T^{\epsilon_1}(Q_{v^0}^N) \leqslant N\epsilon_1 K_{\text{time}}\right) \geqslant \mathbb{P}\Big(\left\{\frac{2N\epsilon_1}{\lambda_{v^0}} \leqslant d(N^{1-a}K_{\text{act}}) \leqslant N\epsilon_1 K_{\text{time}}\right\}$$
$$\cap\left(\left\{NT^{\epsilon_1(Q_{v^0}^N)} \leqslant d(N^{1-a}K_{\text{act}})\right\} \cup \left\{\int_0^{d(N^{1-a}K_{\text{act}})} \sigma_{v^0}(s)\mathrm{d}s \leqslant d(N^{1-a}K_{\text{act}})\frac{\lambda_{v^0}}{2}\right\}\right)\Big).$$

We now justify why the probability on the right hand side converges to 1. As we already mentionned, because of Lemma B.3,

$$\mathbb{P}\left(\left\{NT_-^{\epsilon_1}(Q_{v^0}^N) \leqslant d(N^{1-a}K_{\text{act}})\right\} \cup \left\{Q_{v^0}^N(\frac{d(N^{1-a}K_{\text{act}})}{N}) \geqslant \frac{\lambda_{v^0}d(N^{1-a}K_{\text{act}})}{2N}\right\}\right) \to 1 \text{ as } N \to +\infty.$$

By Lemmas B.2 and B.4,

$$\mathbb{P}\left(\frac{2N\epsilon_1}{\lambda_{v^0}} \leqslant d(N^{1-a}K_{\text{act}}) \leqslant N\epsilon_1 K_{\text{time}}\right) \to 1 \text{ as } N \to +\infty.$$

Since both probabilities converge to 1, the probability of the intersection converges to 1 as well.

For the second part, in the event

$$\left\{T_-^{\epsilon_1}(Q_{v^0}^N) \leqslant \epsilon_1 K_{\text{time}}\right\},$$

we get

$$\inf_{v\in V', s\leqslant T_-^{\epsilon_1}(Q_{v^0}^N)} Q_v^N(s) \geqslant \inf_{v\in V', s\leqslant \epsilon_1 K_{\text{time}}} Q_v^N(s).$$

Since $\epsilon_1 < \frac{\epsilon_0}{2}$ and $\mathbb{P}\left(E_-^N(\epsilon_1 K_{\text{time}})\right) \to 1$, we get by definition of $E_-^N(t)$ that

$$\mathbb{P}\left(\inf_{v\in V', s\leqslant T_-^{\epsilon_1}(Q_{v^0}^N)} Q_v^N(s) \geqslant \epsilon_1\right) \to 1 \text{ as } N \to +\infty.$$

By definition of the stopping time, $Q_{v^0}^N(T_-^{\epsilon_1}(Q_{v^0}^N)) \geqslant \epsilon_1$, which proves the second part of the result. □

B.3 Proofs using coupling

Lemma B.2 is proved independently from the rest using only the coupling described below. Recall the definition of the events from (B.2) and (B.3)

$$E_-^N(t) = \left\{\inf_{v\in V', s\leqslant t} Q_v^N(s) \geqslant \frac{\epsilon_0}{2}\right\},$$

and

$$E_+^N(t) = \left\{\sup_{v\in V,\, s\leqslant t} Q_v^N(s) \leqslant 2\left\|q^0\right\|_\infty\right\}.$$

The results in Lemmas B.3 and B.4 are proved using the coupling and Lemma B.2 to state that

$$\mathbb{P}\left(E_-^N(d(N^{1-a}K_{\text{act}})) \cap E_+^N(d(N^{1-a}K_{\text{act}}))\right) \to 1 \text{ as } N \to +\infty.$$

Let $\mathcal{E}(\lambda)$ denote the exponential distribution with parameter λ and $\mathcal{G}(p)$ the geometric distribution with success probability p. We use the symbol $\overset{(d)}{=}$ to denote equality in distribution. Elementary computations show that if $G \overset{(d)}{=} \mathcal{G}(p)$ and $(E_k)_{k\in\mathbb{N}}$ an *i.i.d.* family of exponential variables with parameter λ are independent,

$$\sum_{k=1}^{G} E_k \overset{(d)}{=} \mathcal{E}(p\lambda), \quad \min_{k=1,\dots,n} E_k \overset{(d)}{=} \mathcal{E}(n\lambda) \text{ and } E_k \overset{(d)}{=} (\lambda)^{-1}\mathcal{E}(1).$$

Given the schedule, the dynamic of Q is easy to construct. Arrivals at each node $v \in V$ happen at the points of independent Poisson processes on $\mathbb{R}_+$ of intensities λ_v: let $(P_v)_{v\in V}$ be n independent Poisson processes on $\mathbb{R}_+$ of intensities $(\lambda_v)_{v\in V}$. They will serve as arrival processes. Similarly, let $(R_v)_{v\in V}$ also be independent unit intensity Poisson processes. Departures from queue $v \in V$ happen at the points of $\widetilde{R}_v(t) \coloneqq R_v(\int_0^t \mathbb{1}_{Q_v(s)>0}\sigma_v(s)\mathtt{d}s)$. All P_v and R_v are independent.

For the scheduler, we will construct the dynamic between each activation of the node fixed at the beginning of the previous section $v^0 \in V \setminus V'$. Let $(b(l))_{l\in\mathbb{N}}$ and $(d(l))_{l\in\mathbb{N}}$ the successive activation and deactivation times of node v^0: assume $\sigma(0) = \mathbf{0}$, then $d(0) = 0$ and for all $k > 0$,

$$b(k) = \inf\{t > d(k-1),\, \sigma(t) = v^0\},\ d(k) = \inf\{t > b(k),\, \sigma(t) = \mathbf{0}\}. \tag{B.8}$$

Between each time the schedule is empty, n inhomogeneous exponential variables compete to activate their nodes. For each $v \in V$, the activation rate is $\dfrac{1}{(Q_v(t)+1)^{-a}+1}$.

Let's define $d_0^k = d(k)$, and for any $m \geqslant 1$,

$$b_m^k \coloneqq \inf\left\{t > d_{m-1}^k,\, \sigma(t) \neq \mathbf{0}\right\} \text{ and } d_m^k \coloneqq \inf\left\{t > b_m^k,\, \sigma(s) = \mathbf{0}\right\}.$$

More formally, let $(A_v^k(m))_{k\in\mathbb{N},m\in\mathbb{N},v\in V}$ be a family of unit parameter independent exponential variables. The m^{th} time a node deactivates after $d(k)$, all nodes enter competition for the next activation. If this deactivation occured before the $k+1^{th}$ activation of v^0, the deactivation happened at time d_m^k and the next activation occurs at time

$$d_m^k + \min_{v\in V} \inf\left\{t > 0,\, A_v^k(m+1) \leqslant \int_0^t \frac{\mathrm{d}s}{1 + (Q_v(d_m^k + s) + 1)^{-a}}\right\}, \tag{B.9}$$

where for every w, before any activation,

$$Q_w(d_m^k + .) = Q_w(d_m^k) + P_w(d_m^k + .) - P_w(d_m^k). \tag{B.10}$$

The queue that activates is the one realising the minimum in (B.9). Let $G(k) \geqslant 0$ be the number of times a node different from v^0 activates between the k^{th} and $k+1^{th}$ activation of v^0. For any $k, m \in \mathbb{N}$, the variables $(A_v^k(m))_{v\in V}$ are only used to determine the idle period and the active queue before the m^{th} activation of a node after the k^{th} activation of v^0 provided v^0 has not activated $k+1$ times. An important remark that we will use multiple times in this section is that the activation rates of any node can be bounded regardless of the state of the network: for any $Q \in \mathbb{R}_+$,

$$1/2 \leqslant \frac{1}{1 + (Q+1)^{-a}} \leqslant 1.$$

Similarly, for any $m > 0$, if v is the m^{th} node to activate after the k^{th} activation of v^0, this activation occurs at time b_m^k, it will have an activation duration noted $d^k(m) - b^k(m)$ that can be expressed as another "inhomogeneous exponential" variable:

$$d_m^k - b_m^k = \inf\left\{t > 0,\, D_m^k \leqslant \int_0^t \frac{\mathrm{d}s}{1 + (Q_v(b_m^k + s) + 1)^{a}}\right\}, \tag{B.11}$$

with $(D_m^k)_{k\in\mathbb{N},m\in\mathbb{N}}$ unit intensity *i.i.d.* exponential variables, discarded after each deactivation. For consistency purposes, the activation duration of node v^0 is constructed similarly with $(D_0^k)_{k\in\mathbb{N}}$ a family of *i.i.d.* exponential variables with unit parameter independent from the rest: for the k^{th} activation period of v^0,

$$d(k) - b(k) = \inf\left\{t > 0,\, D_0^k \leqslant \int_0^t \frac{\mathrm{d}s}{1 + (Q_{v^0}(b(k) + s) + 1)^{a}}\right\}.$$

Before any deactivation, the evolution of the active node is given by

$$Q_v(b_m^k + t) = Q_v(b_m^k) + P_v(b_m^k + t) - P_v(b_m^k) - \widetilde{R}_v(b_m^k + t) + \widetilde{R}_v(b_m^k).$$

Each inactive node evolves as in (B.10). Once node v^0 activates $k+1$ times, discard all $(A_v^k(m))_{v\in V,m\in\mathbb{N}}$ and $(D_m^k)_{m\in\mathbb{N}}$.

Proof of Lemma B.2. The first step is to consider the intersection between $E_+^N(\epsilon_1 K_{\text{time}})$

and $\left\{d(N^{1-a}K_{\text{act}}) \leqslant N\epsilon_1 K_{\text{time}}\right\}$. In $E_+^N(\epsilon_1 K_{\text{time}})$, if the m^{th} activation of a node between the k^{th} and $k+1^{th}$ activations of node v^0 is such that

$$d_m^k \leqslant N\epsilon_1 K_{\text{time}},$$

by (B.11), $d_m^k - b_m^k$ is smaller than

$$\inf\left\{t>0,\, D_m^k \leqslant \frac{t}{(N4\left\|q^0\right\|_\infty)^a}\right\} = (N4\left\|q^0\right\|_\infty)^a D_m^k.$$

Similarly, for any $k \leqslant N^{1-a}K_{\text{act}}$,

$$d(k) - b(k) \leqslant \inf\left\{t>0,\, D_0^k \leqslant \frac{t}{(N4\left\|q^0\right\|_\infty)^a}\right\} = (N4\left\|q^0\right\|_\infty)^a D_0^k.$$

There are exactly $N^{1-a}K_{\text{act}}$ activations of node v^0 before $d(N^{1-a}K_{\text{act}})$. Between each activation of node v^0, the number of times a node other than v^0 activates can be bounded by a geometric variable with parameter $\frac{1}{2n-1}$ because of (B.9): recall $G(k)$ the number of times a node different from v^0 activates between the k^{th} and $k+1^{th}$ activation of v^0. By (B.9), $G(k)$ is smaller than the number of times a queue different from v^0 would activate if the activation rate of node v^0 was $\frac{1}{2}$ and the activation rates of all other queues were 1. More formally, $G(k)$ is smaller than $\bar{G}(k)$: the number of $p>0$ such that

$$\min\left[\min_{v\neq v^0} A_v^k(p), 2A_{v^0}^k(p)\right] \neq 2A_{v^0}^k(p),$$

before the first time

$$\min\left[\min_{v\neq v^0} A_v^k(p), 2A_{v^0}^k(p)\right] = 2A_{v^0}^k(p).$$

Queue v^0 will necessarily activate if

$$\min\left[\min_{v\neq v^0} A_v^k(p), 2A_{v^0}^k(p)\right] = 2A_{v^0}^k(p),$$

but the activation of v^0 may occur for a smaller p', thus $G(k) \leqslant \bar{G}(k)$. By independence of $(A_v^m(p))_{m,p\in\mathbb{N},v\in V}$,

$$\bar{G}(k) \overset{(d)}{=} \mathcal{G}(\frac{1}{2n-1}) \tag{B.12}$$

and is independent from $(D_v^k)_{k\in\mathbb{N},v\in V}$. For any $p>0$, the queue realizing the minimum of the p^{th} competition is independent from $\min\left[\min_{v\neq v^0} A_v^k(p), 2A_{v^0}^k(p)\right]$ and from the result of the competition for a different k or p. This is due to independence of $(A_v^k(p))_{k,p\in\mathbb{N},v\in V}$ and the fact that

$$\mathbb{P}\left(\min\left[\min_{v\neq v^0} A_v^k(p), 2A_{v^0}^k(p)\right] = A_{v^1}^k,\, A_{v^1}^k(p) > t\right) = \frac{2}{2n-1}\exp(-t\frac{2n-1}{2}), \tag{B.13}$$

and

$$\mathbb{P}\left(\min\left[\min_{v\neq v^0} A_v^k(p), 2A_{v^0}^k(p)\right] = 2A_{v^0}^k(p),\ 2A_{v^0}^k(p) > t\right) = \frac{1}{2n-1}\exp(-(\frac{2n-1}{2})t). \tag{B.14}$$

Since the activation rate is greater than $\frac{1}{2}$, the associated idle period is smaller than

$$\widetilde{A}^k(p) \coloneqq 2\min\left[\min_{v\neq v^0} A_v^k(p), 2A_{v^0}^k(p)\right] \overset{(d)}{=} \frac{4}{2n-1}\mathcal{E}(1),$$

which is independent from $\bar{G}(k)$. The independence between $\bar{G}(k)$ and the duration of each idle period comes from (B.13), (B.14) and independence of $(A_v^k(p))_{k,p\in\mathbb{N},v\in V}$. Let us define

$$\bar{D}^N \coloneqq \sum_{k=1}^{N^{1-a}K_{\text{act}}}\sum_{p=0}^{\bar{G}(k)}\left(\widetilde{A}^k(p) + (4N\left\|q^0\right\|_\infty)^a D_p^k\right),$$

with the convention that $\widetilde{A}_0^k \coloneqq \widetilde{A}_{\bar{G}(k)}^k$, which is still independent from $(\bar{G}(k))_{k\in\mathbb{N}}$. The term with $p=0$ handles the activation duration of node v^0 and the last idle period before v^0 activates. Because of the use of different $(A_v^k(m))_{v\in V,m\in\mathbb{N}}$ after the $k+1^{th}$ activation of v^0, $(\bar{G}(k))_{k\in\mathbb{N}}$ forms an *i.i.d.* family of geometric variables. Hence, $\bar{D}^N$ is a sum of *i.i.d.* variables.

Recall the event

$$E_+^N(\epsilon_1 K_{\text{time}}) = \left\{\sup_{v\in V,\, s\leqslant\epsilon_1 K_{\text{time}}} Q_v^N(s) \leqslant 2\left\|q^0\right\|_\infty\right\}.$$

The coupling ensures that

$$\begin{aligned}\mathbb{P}\left(d(N^{1-a}K_{\text{act}}) \leqslant N\epsilon_1 K_{\text{time}},\, E_+^N(\epsilon_1 K_{\text{time}})\right) &\geqslant \mathbb{P}\left(\bar{D}^N \leqslant N\epsilon_1 K_{\text{time}},\, E_+^N(\epsilon_1 K_{\text{time}})\right)\\ &= \mathbb{P}\left(\frac{1}{NK_{\text{act}}}\bar{D}^N \leqslant \frac{\epsilon_1 K_{\text{time}}}{K_{\text{act}}},\, E_+^N(\epsilon_1 K_{\text{time}})\right)\end{aligned}$$

By construction of the coupling, $(\widetilde{A}^k(p))_{k,p\in\mathbb{N}}$ and $(D_p^k)_{k,p\in\mathbb{N}}$ from two *i.i.d.* families, independent from $\bar{G}(k)$ for any $k>0$. By independence of $(\bar{G}(k))_{k\in\mathbb{N}}$, $(\widetilde{A}_p^k)_{k,p\in\mathbb{N}}$,$(D_p^k)_{k,p\in\mathbb{N}}$ and independence between those variables, by the law of large numbers

$$\frac{1}{N^{1-a}K_{\text{act}}}\sum_{k=1}^{N^{1-a}K_{\text{act}}}\sum_{p=0}^{\bar{G}(k)}\left(\frac{\widetilde{A}_p^k}{N^a} + (4\left\|q^0\right\|_\infty)^a D_p^k\right) \to n2^{1+2a}\left\|q^0\right\|_\infty^a \text{ almost surely as } N\to+\infty. \tag{B.15}$$

By definition of K_{act},

$$K_{\text{act}} = \frac{\epsilon_1 K_{\text{time}}}{4^{1+a}n\left\|q^0\right\|_\infty^a} < \frac{\epsilon_1 K_{\text{time}}}{2^{1+2a}n\left\|q^0\right\|_\infty^a},$$

and so

$$n2^{1+2a}\left\|q^0\right\|_\infty^a < \frac{\epsilon_1 K_{\text{time}}}{K_{\text{act}}}.$$

From (B.4), we get

$$\mathbb{P}\left(E_+^N(N\epsilon_1 K_{\text{time}})\right) \to 1 \text{ as } N \to +\infty,$$

and so,

$$\mathbb{P}\left(d(N^{1-a}K_{\text{act}}) \leqslant N\epsilon_1 K_{\text{time}}\right) \to 1 \text{ as } N \to +\infty.$$

The second part of the result comes from the fact that on $\{d(N^{1-a}K_{\text{act}}) \leqslant N\epsilon_1 K_{\text{time}}\}$, we get

$$\sup_{v\in V,\, t\leqslant d(N^{1-a}K_{\text{act}})} Q_v^N(t) \leqslant \sup_{v\in V,\, t\leqslant N\epsilon_1 K_{\text{act}}} Q_v^N(t) \leqslant 2\left\|q^0\right\|_\infty,$$

and

$$\inf_{v\in V',\, t\leqslant d(N^{1-a}K_{\text{act}})} Q_v^N(t) \geqslant \inf_{v\in V',\, t\leqslant N\epsilon_1 K_{\text{act}}} Q_v^N(t) \geqslant \frac{\epsilon_0}{2}.$$

In this situation, (B.4) and the definition of the events in (B.2) and (B.3) are sufficient to state that

$$\mathbb{P}\left(E_+^N(\frac{d(N^{1-a}K_{\text{act}})}{N}) \cap E_-^N(\frac{d(N^{1-a}K_{\text{act}})}{N}) \cap \{d(N^{1-a}K_{\text{act}}) \leqslant N\epsilon_1 K_{\text{time}}\}\right) \to 1 \text{ as } N \to +\infty.$$

□

We now turn to the proof of Lemma B.3. Recall the definition of the event:

$$E_{v^0}^N = \left\{\sup_{Ns\leqslant d(N^{1-a}K_{\text{act}})} Q_{v^0}^N(s) \leqslant \epsilon_1\right\} \cap E_-^N(d(N^{1-a}K_{\text{act}})).$$

Proof of Lemma B.3. By definition of the activation periods,

$$\int_0^{d(N^{1-a}K_{\text{act}})} \sigma_{v^0}(s)\mathrm{d}s = \sum_{k=1}^{N^{1-a}K_{\text{act}}} \left(d(k) - b(k)\right).$$

In $E_{v^0}^N$, for any $k \leqslant N^{1-a}K_{\text{act}}$,

$$d(k) - b(k) \leqslant \inf\left\{t > 0,\, D_0^k \leqslant \frac{t}{(2\epsilon_1)^a}\right\} = (2\epsilon_1)^a D_0^k,$$

by (B.11), because in the event $E_{v^0}^N$, $Q_{v^0}^N(t)$ is bounded above by ϵ_1 for $t \leqslant d(N^{1-a}K_{\text{act}})$. By independence of $(D_v^k)_{k\in\mathbb{N},v\in V}$ and the deterministic nature of the parameter, the bound on activation durations form an *i.i.d.* family of exponential variables. Similarly, it is possible to lower bound $d(N^{1-a}K)$ using once again (B.11). In $E_{v^0}^N$, every node in V' has queue lengths uniformly bounded below by $\frac{\epsilon_0}{2}$ until $d(N^{1-a}K_{\text{act}})$. For any $k < N^{1-a}K_{\text{act}}$ and $m < G(k)$, such that $\sigma_v(b_m^k) = 1$ for some $v \in V'$,

$$d_m^k - b_m^k \geqslant (\frac{N\epsilon_0}{2})^a D_m^k.$$

There are $N^{1-a}K_{\text{act}}$ activations for node v^0 before $d(N^{1-a}K_{\text{act}})$. Inbetween activations for node v^0, there is a number of times where a node in V' activates. Because of (B.9), the activation rate of a node in V' is greater than $\frac{1}{2}$. The maximum activation rate of any queue is 1. After any activation, the probability that a node in V' is the next to activate is greater than $\frac{|V'|}{2n-|V'|}$ regardless of the evolution of the network, each try being independent from the others because of the competition

between new $A_v^k(m)$. The number of times a queue in V' activates between the k^{th} and $k+1^{th}$ activations of v^0 is greater than $\check{G}(k)$ constructed similarly to (B.12): a geometric variable counting the number of times a queue in V' activates before a queue not in V' activates:

$$\check{G}(k) \overset{(d)}{=} \mathcal{G}(\frac{|V'|}{2n-|V'|}).$$

The independence between $(A_v^k(m))_{k,m\in\mathbb{N},\, v\in V}$ and $(D_m^k)_{k,m\in\mathbb{N}}$ and deterministic nature of the intensities ensures that $\check{G}(k)$ is independent from $(D_m^k)_{k,m\in\mathbb{N}}$. This discussion amounts to the bounds

$$\int_0^{d(N^{1-a}K_{\text{act}})} \sigma_{v^0}(s)\mathrm{d}s \leqslant (2N\epsilon_1)^a \sum_{k=1}^{N^{1-a}K_{\text{act}}} D_0^k,$$

and

$$d(N^{1-a}K_{\text{act}}) \geqslant \sum_{k=1}^{N^{1-a}K_{\text{act}}} \sum_{m=1}^{\check{G}(k)} (\frac{N\epsilon_0}{2})^a D_m^k. \tag{B.16}$$

Let's call

$$\overline{D}_k = \sum_{m=1}^{\check{G}(k)} (\frac{N\epsilon_0}{2})^a D_m^k.$$

In the event $E_{v^0}^N$, we get

$$\begin{aligned}\mathbb{P}\left(\sum_{k=1}^{N^{1-a}K_{\text{act}}} d(k)-d(k) > d_v(N^{1-a}K_{\text{act}})\frac{\lambda_{v^0}}{2},\, E_{v^0}^N\right) &\leqslant \mathbb{P}\left((2N\epsilon_1)^a \sum_{k=1}^{N^{1-a}K_{\text{act}}} D_0^k > \sum_{k=1}^{N^{1-a}K_{\text{act}}} \overline{D}_k\frac{\lambda_{v^0}}{2},\, E_{v^0}^N\right)\\ &\leqslant \mathbb{P}\left(\frac{1}{N^{1-a}K_{\text{act}}} \sum_{k=1}^{N^{1-a}K_{\text{act}}} (2\epsilon_1)^a D_0^k - \frac{\lambda_{v^0}\overline{D}_k}{2} > 0\right)\end{aligned}$$

By construction, since we use new $(A_v^{k+1}(p))_{p\in\mathbb{N},v\in V}$ and $(D_m^{k+1})_{m\in\mathbb{N}}$ after the $k+1^{th}$ activation of v^0, $((2\epsilon_1)^a D_0^k - \frac{\lambda_{v^0}\overline{D}_k}{2})_{k\in\mathbb{N}}$ is an *i.i.d.* family of random variables whose mean is given by

$$\mathbb{E}\left[(2\epsilon_1)^a D_{v^0}^k - \frac{\lambda_{v^0}\overline{D}_k}{2}\right] = (2\epsilon_1)^a - \frac{\lambda_{v^0}\epsilon_0^a(2n-|V'|)}{2^{1+a}|V'|}.$$

Notice that

$$\frac{\epsilon_0}{2}\left(\frac{\lambda_{v^0}(2n-|V'|)}{2^{1+a}|V'|}\right)^{1/a} > \frac{\epsilon_0}{2^{2+1/a}}\left(\min_v \lambda_v\right)^{1/a} > \epsilon_1,$$

because $|V'| \leqslant n$. By the law of large numbers

$$\mathbb{P}\left(\frac{1}{N^{1-a}K_{\text{act}}} \sum_{k=1}^{N^{1-a}K_{\text{act}}} \left((2\epsilon_1)^a D_{v^0}^k - \frac{\lambda_{v^0}\overline{D}_k}{2}\right) > 0\right) \to 0 \text{ as } N \to +\infty,$$

which proves the first part of the result.

For any $t > 0$, since by Lemma B.2

$$\mathbb{P}\left(\inf_{v\in V', Ns\leqslant d(N^{1-a}K_{\mathrm{act}})} Q_v^N(s) > \frac{\epsilon_0}{2}\right) \to 1 \text{ as } N \to +\infty,$$

we get as a consequence of the first part that

$$\mathbb{P}\left(\int_0^{d(N^{1-a}K)} \sigma_{v^0}(s)\mathrm{d}s > d(N^{1-a}K)\frac{\lambda_{v^0}}{2}, \sup_{Ns\leqslant d(N^{1-a}K)} Q_{v^0}^N(s) < \epsilon_1\right) \to 0 \text{ as } N \to +\infty. \tag{B.17}$$

With probability close to one, either

$$T_-^{\epsilon_1}(Q_{v^0}^N) \leqslant d(N^{1-a}K_{\mathrm{act}}),$$

or

$$\int_0^{d(N^{1-a}K_{\mathrm{act}})} \sigma_{v^0}(s)\mathrm{d}s \leqslant d(N^{1-a}K_{\mathrm{act}})\frac{\lambda_{v^0}}{2}.$$

In the latter case, by (B.1),

$$Q_{v^0}^N(\frac{b(N^{1-a}K_{\mathrm{act}})}{N}) \geqslant \frac{d(N^{1-a}K_{\mathrm{act}})}{N}(\lambda_{v^0} - \frac{\lambda_{v^0}}{2}).$$

More formally,

$$\begin{aligned}
&\mathbb{P}\left(Q_{v^0}^N(\frac{d(N^{1-a}K_{\mathrm{act}})}{N}) > d(N^{1-a}K_{\mathrm{act}})\frac{\lambda_{v^0}}{2}, \sup_{Ns\leqslant d(N^{1-a}K_{\mathrm{act}})} Q_{v^0}^N(s) < \epsilon_1\right) \\
&\leqslant \mathbb{P}\left(\int_0^{d(N^{1-a}K_{\mathrm{act}})} \sigma_{v^0}(s)\mathrm{d}s > d(N^{1-a}K_{\mathrm{act}})\frac{\lambda_{v^0}}{2}, \sup_{Ns\leqslant d(N^{1-a}K_{\mathrm{act}})} Q_{v^0}^N(s) < \epsilon_1\right) \to 0 \text{ as } N \to +\infty
\end{aligned}$$

□

Proof of Lemma B.4. This proof relies on the coupling argument given in (B.16) (see the discussion there for the justification). In the event

$$\left\{\inf_{v\in V', Ns\leqslant d(N^{1-a}K_{\mathrm{act}})} Q_v^N(s) \geqslant \frac{\epsilon_0}{2}\right\},$$

it states that

$$d(N^{1-a}K_{\mathrm{act}}) \geqslant \sum_{k=1}^{N^{1-a}K_{\mathrm{act}}} \bar{D}_k = \sum_{k=1}^{N^{1-a}K_{\mathrm{act}}} \sum_{m=1}^{\check{G}(k)} (\frac{N\epsilon_0}{2})^a D_m^k,$$

with $\check{G}(k)$ *i.i.d.* and independent from $(D_m^k)_{k,m\in\mathbb{N}}$, of common distribution $\mathcal{G}(\frac{|V'|}{2n-|V'|})$. Obviously, because of that,

$$\begin{aligned}
&\mathbb{P}\left(d(N^{1-a}K_{\mathrm{act}}) < \frac{2N\epsilon_1}{\lambda_{v^0}}, \inf_{v\in V', Ns\leqslant d(N^{1-a}K_{\mathrm{act}})} Q_v^N(s) \geqslant \frac{\epsilon_0}{2}\right) \\
&\qquad \leqslant \mathbb{P}\left(\sum_{k=1}^{N^{1-a}K_{\mathrm{act}}} \overline{D}_k < \frac{2N\epsilon_1}{\lambda_{v^0}}\right) = \mathbb{P}\left(\frac{1}{N^{1-a}K_{\mathrm{act}}} \sum_{k=1}^{N^{1-a}K_{\mathrm{act}}} \overline{D}_k < \frac{2\epsilon_1}{K_{\mathrm{act}}\lambda_{v^0}}\right)
\end{aligned}$$

By independence of the $\overline{D}_k$ and the law of large numbers, $\frac{1}{N^{1-a}K_{\text{act}}} \sum_{k=1}^{N^{1-a}K_{\text{act}}} \overline{D}_k \to \mathbb{E}\left[\overline{D}_1\right]$ almost surely. Elementary computations give

$$\mathbb{E}\left[\overline{D}_k\right] = \frac{\epsilon_0^a(2n - |V'|)}{2^a|V'|},$$

and thus since

$$K_{\text{act}} = \frac{\epsilon_1 K_{\text{time}}}{n4^{1+a}\left\|q^0\right\|_\infty^a} = \epsilon_1 \frac{1}{n^2\epsilon_0^a\left\|\lambda\right\|_\infty}.$$

Notice that

$$\frac{2\epsilon_1}{\lambda_{v^0}K_{time}} = \frac{\min_v \lambda_v}{\lambda_{v^0}} \frac{\epsilon_0^a(2n-n)}{2^a n} < \frac{\epsilon_0^a(2n-|V'|)}{2^a|V'|} = E[\overline{D}_k],$$

where the last inequality is due to the fact that the right hand side is decreasing in $|V'|$. We get

$$\mathbb{P}\left(\frac{1}{N^{1-a}K_{\text{act}}} \sum_{k=1}^{N^{1-a}K_{\text{act}}} \overline{D}_k < \frac{2\epsilon_1}{K_{\text{act}}\lambda_{v^0}}\right) \to 0 \text{ as } N \to +\infty.$$

This entails

$$\mathbb{P}\left(d(N^{1-a}K_{\text{act}}) < \frac{2N\epsilon_1}{\lambda_{v^0}},\ E_-^N\left(\frac{d(N^{1-a}K_{\text{act}}))}{N}\right)\right) \to 0 \text{ as } N \to +\infty.$$

By Lemma B.2,

$$\mathbb{P}\left(E_-^N\left(\frac{d(N^{1-a}K_{\text{act}}))}{N}\right)\right) \to 1 \text{ as } N \to +\infty$$

and the result is proved. □

Bibliography

[Abr70] N. Abramson. "The ALOHA System: another alternative for computer communications". In: *Proc. of the fall joint computer conference*. ACM. 1970, pp. 281–285 (cit. on p. 7).

[AC19] R. Atar and A. Cohen. "Serve the shortest queue and walsh brownian motion". In: *Ann. Appl. Probab.* 29.1 (2019), pp. 613–651 (cit. on p. 25).

[Ald81] D. J. Aldous. "Some inequalities for reversible Markov chains". In: *J. London Math. Soc.* 2.25 (1981), pp. 564–576 (cit. on p. 36).

[Ald87] D. Aldous. "Ultimate instability of exponential back-off protocol for acknowledgment-based transmission control of random access communication channels". In: *Proc. IEEE Trans. Inf. Theory*. Vol. 33. 2. 1987, pp. 219–223 (cit. on p. 7).

[Bar88] A. D. Barbour. "Stein's method and Poisson process convergence". In: *J. Appl. Probab.* 25 A celebration of Applied Probability (1988), pp. 175–184 (cit. on p. 18).

[Bar90] A. D. Barbour. "Stein's method for diffusion approximations". In: *Probab. Theory Relat. Fields* 84.3 (1990), pp. 297–322 (cit. on p. 18).

[BB09] F. Baccelli and B. Blaszczyszyn. "Stochastic Geometry and Wireless Networks, number II - Applications". In: *Foundations and Trends® in Networking* 4.1-2 (2009), pp. 1–312 (cit. on p. 3).

[BBL11] N. Bouman, S. C. Borst, and J. van Leeuwaarden. "Achievable delay performance in CSMA networks". In: *Proc. of 49th Annual Allerton Conference*. 2011, pp. 384–391 (cit. on pp. v, 10).

[BD17] A. Braverman and J. G. Dai. "Stein's method for steady-state diffusion approximations of $M/Ph/n+M$ systems". In: *Ann. Appl. Probab.* 27.1 (2017), pp. 550–581 (cit. on p. 19).

[BDF16] A. Braverman, J. G. Dai, and J. Feng. "Stein's method for steady-state diffusion approximations: An introduction through the Erlang-A and Erlang-C models". In: *Stoch. Syst.* 6.2 (2016), pp. 301–366 (cit. on p. 19).

[Bil99] P. Billingsley. *Convergence of probability measures*. Second edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York, 1999, pp. x+277 (cit. on pp. 75, 78, 85).

[Bir+12] B. Birand, M. Chudnovsky, B. Ries, P. Seymour, G. Zussman, and Y. Zwols. "Analyzing the Performance of Greedy Maximal Scheduling via Local Pooling and Graph Theory". In: *Proc. IEEE/ACM Trans. Netw.* Vol. 20. 1. 2012, pp. 163–176 (cit. on p. 6).

[BP02] T. Bonald and A. Proutière. "Insensitivity in processor-sharing networks". In: *Performance Evaluation* 49.1-4 (2002) (cit. on p. 4).

[Bra11] M. Bramson. "Stability of join the shortest queue networks". In: *Ann. Appl. Probab.* 21.4 (2011), pp. 1568–1625 (cit. on p. 13).

[Bra94] M. Bramson. "Instability of FIFO queueing networks". In: *Ann. Appl. Probab.* 4.2 (1994), pp. 414–431 (cit. on p. 23).

[Bra98] M. Bramson. "State space collapse with application to heavy traffic limits for multiclass queueing networks". In: *Queueing Syst.* 30.1-2 (1998), pp. 89–148 (cit. on pp. 19, 20, 25).

[Cas+20] E. Castiel, S. Borst, L. Miclo, F. Simatos, and P. Whitting. In: *arXiv* (2020). arXiv id: 1904.03980, `https://arxiv.org/pdf/1904.03980.pdf` (cit. on pp. xxiv, xxviii).

[Cec+16] F. Cecchi, S. C. Borst, J. S. H. van Leeuwaarden, and P. A. Whiting. "Mean-field limits for large-scale random-access networks." In: *arXiv* (2016). arXiv id: 1611.09723, `https://arxiv.org/pdf/1611.09723.pdf` (cit. on p. 12).

[Che75] L. H. Y. Chen. "Poisson approximation for dependent trials". In: *Ann. Prob.* 3.3 (1975), pp. 534–545 (cit. on p. 18).

[CPR95] E. G. Coffman, A. A. Puhalskii, and M. I. Reiman. "Polling systems with zero switchover times: a heavy-traffic averaging principle". In: *Ann. Appl. Probab.* 5.3 (1995), pp. 681–719 (cit. on p. 14).

[CPR98] E. G. Coffman, A. A. Puhalskii, and M. I. Reiman. "Polling systems in heavy traffic: a Bessel process limit". In: *Math. Oper. Res.* 23.2 (1998), pp. 257–304 (cit. on p. 14).

[Dai95] J. G. Dai. "On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models". In: *Ann. Appl. Probab.* 5.1 (1995), pp. 49–77 (cit. on pp. vii, 22, 23).

[DBT03] O. Dousse, F. Baccelli, and P. Thiran. "Impact of interferences on connectivity in ad hoc networks". In: *Proc. IEEE INFOCOM.* Vol. 3. 2003, pp. 1724–1733 (cit. on p. 3).

[Dem16] J.-P. Demailly. *Analyse numérique et équations différentielles (French).* Fourth. Grenoble Sciences. EDP Sciences, Les Ulis, 2016, pp. vii+368 (cit. on p. 65).

[DM95] J. G. Dai and S. P. Meyn. "Stability and convergence of moments for multiclass queueing networks via fluid limit models". In: *Proc. IEEE Trans. Automat. Control.* Vol. 40. 11. 1995, pp. 1889–1904 (cit. on p. 23).

[Dor+15] J. P. L. Dorsman, S. C. Borst, O. J. Boxma, and M. Vlasiou. "Markovian polling systems with an application to wireless random-access networks." In: *Perform. Eval.* 85–86 (2015), pp. 33–51 (cit. on p. 12).

[Dou+06] O. Dousse, M. Franceschetti, N. Macris, R. Meester, and P. Thiran. "Percolation in the signal to interference ratio graph". In: *J. Appl. Probab.* 43.2 (2006), pp. 552–562 (cit. on p. 3).

[DSC96] P. Diaconis and L. Saloff-Coste. "Logarithmic Sobolev inequalities for finite Markov chains". In: *Ann. Appl. Probab.* 6.3 (1996), pp. 695–750 (cit. on p. 35).

[DW06] A. Dimakis and J. Walrand. "Sufficient conditions for stability of longest-queue-first scheduling: second-order properties using fluid limits". In: *Adv. in Appl. Probab.* 38.2 (2006), pp. 505–521 (cit. on p. 6).

[EK86] S. N. Ethier and T. G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Characterization and convergence. John Wiley & Sons, Inc., New York, 1986, pp. x+534 (cit. on pp. xxi, 82, 105).

[ES12] A. Eryilmaz and R. Srikant. "Asymptotically tight steady-state queue length bounds implied by drift conditions". In: *Queueing Syst.* 72.3 (2012), pp. 311–359 (cit. on p. 20).

[FPR10] M. Feuillet, A. Proutiere, and P. Robert. "Random capture algorithms fluid limits and stability". In: *Proc. of ITA workshop*. 2010, pp. 1–4 (cit. on pp. vii, 23, 24).

[FR14] M. Feuillet and P. Robert. "A scaling analysis of a transient stochastic network". In: *Adv. in Appl. Probab.* 46.2 (2014), pp. 516–535 (cit. on pp. 14, 15).

[FW12] M. I. Freidlin and A. D. Wentzell. *Random perturbations of dynamical systems*. Third. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] 260. Translated from the 1979 Russian original by Joseph Szücs. Springer, Heidelberg, 2012, pp. xxviii+458 (cit. on pp. v, 14, 15, 116).

[GBW14] J. Ghaderi, S. C. Borst, and P. A. Whiting. "Queue-based random-access algorithms: fluid limits and stability issues". In: *Stoch. Syst.* 4.1 (2014), pp. 81–156 (cit. on pp. vii, 10, 23, 24).

[Gil59] E. N. Gilbert. "Random Graphs". In: *Ann. Math. Statist.* 30.4 (1959), pp. 1141–1144 (cit. on p. 3).

[Gil61] E. N. Gilbert. "Random Plane Networks". In: *Journal of the Society for Industrial and Applied Mathematics (SIAM)* 9.4 (1961), pp. 533–543 (cit. on p. 3).

[Gri14] D. J. Griffiths. *Introduction to quantum mechanics (second edition)*. Pearson Prentice Hall, 2014 (cit. on pp. v, 13).

[GS10] J. Ghaderi and R. Srikant. "On the design of efficient CSMA algorithms for wireless networks". In: *Proc. IEEE Conf. Decis. Control.* 2010, pp. 954–959 (cit. on pp. 9, 44).

[GSS19] F. Götze, H. Sambale, and A. Sinulis. "Higher order concentration for functions of weakly dependent random variables". In: *Electron. J. Probab.* 24 (2019), 19 pp. (Cit. on p. 35).

[Gur14] Itai Gurvich. "Diffusion models and steady-state approximations for exponentially ergodic Markovian queues". In: *Ann. Appl. Probab.* 24.6 (Dec. 2014), pp. 2527–2559 (cit. on p. 19).

[GW19] R. E. Gaunt and N. Walton. "Stein's Method for the Single Server Queue in Heavy Traffic". In: *to appear in Statistics and Probability Letters* (2019+). arXiv id: 1805.08003 `https://arxiv.org/pdf/1805.08003.pdf` (cit. on p. 19).

[HK94] P. J. Hunt and T. G. Kurtz. "Large loss networks". In: *Stochastic Process. Appl.* 53.2 (1994), pp. 363–378 (cit. on pp. 14, 15, 64).

[HR81] J. M. Harrison and M. I. Reiman. "Reflected Brownian motion on an orthant". In: *Ann. Probab.* 9.2 (1981), pp. 302–308 (cit. on p. 20).

[HW81] S. Halfin and W. Whitt. "Heavy-Traffic limits for queues with many exponential servers". In: *Math. Oper. Res.* 29.3 (1981), pp. 417–629 (cit. on p. 24).

[HW96] J. M. Harrison and R. J. Williams. "A Multiclass Closed Queueing Network with Unconventional Heavy Traffic Behavior". In: *Ann. Appl. Probab.* 6.1 (1996), pp. 1–47 (cit. on p. 25).

[Jac57] J. R. Jackson. "Networks of Waiting Lines". In: *Math. Oper. Res.* 5.4 (1957), pp. 518–521 (cit. on p. 22).

[Jen10] O. B. Jennings. "Averaging principles for a diffusion-scaled, heavy-traffic polling station with K job classes". In: *Math. Oper. Res.* 35.3 (2010), pp. 669–703 (cit. on p. 14).

[JLS09] C. Joo, X. Lin, and N. B. Shroff. "Greedy Maximal Matching: Performance Limits for Arbitrary Network Graphs Under the Node-Exclusive Interference Model". In: *Proc. IEEE Trans. Autom. Control.* Vol. 54. 12. 2009, pp. 2734–2744 (cit. on p. 6).

[JS03] J. Jacod and A. N. Shiryaev. *Limit theorems for stochastic processes.* Second. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] 288. Springer-Verlag, Berlin, 2003, pp. xx+661 (cit. on pp. 29, 84, 85).

[JW10] L. Jiang and J. Walrand. "A distributed CSMA algorithm for throughput and utility maximization in wireless networks". In: *Proc. IEEE/ACM Trans. Netw.* Vol. 18. 3. 2010 (cit. on p. 8).

[Kal02] O. Kallenberg. *Foundations of modern probability.* Second. Probability and its Applications (New York). Springer-Verlag, New York, 2002, pp. xx+638 (cit. on pp. vi, vii, 20, 21).

[Kin61] J. Kingman. "The single server queue in heavy traffic". In: *Proc. Cambridge Philos. Soc.* Vol. 57. 4. 1961, pp. 902–904 (cit. on pp. vii, 24).

[Kin62] J. Kingman. "On Queues in Heavy Traffic". In: *J. Royal Stat. Soc. Series B* 24.2 (1962), pp. 383–392 (cit. on pp. vii, 19, 24).

[Kur92] T. G. Kurtz. "Averaging for martingale problems and stochastic approximation". In: Springer, Berlin, 1992, pp. 186–209 (cit. on pp. v, 14, 15, 64).

[KW12] N. W. Kang and J. R. Williams. "Diffusion Approximation for an Input-queued Switch Operating under a Maximum Weight Matching Policy". In: Stoch. Syst. 2.2 (2012), pp. 277–321 (cit. on p. 20).

[LBX11] B. Li, C. Boyaci, and Y. Xia. "A Refined Performance Characterization of Longest-Queue-First Policy in Wireless Networks". In: *Proc. IEEE/ACM Trans. Netw.* Vol. 19. 5. 2011, pp. 1382–1395 (cit. on p. 6).

[LG08] A. Leon-Garcia. *Probability, statistics, and random processes for electrical engineering.* Prentice Hall., 2008 (cit. on p. 3).

[Lie+09] S. C. Liew, C. Kai, J. Leung, and B. Wong. "Back-of-the-Envelope Computation of Throughput Distributions in CSMA Wireless Networks". In: *Proc. IEEE Conf. Commun.* Vol. 9. 9. 2009, pp. 1–6 (cit. on p. 7).

[Liu+08] J. Liu, Y. Yi, A. Proutière, M. Chiang, and H. Vincent-Poor. "Maximizing Utility via Random Access Without Message Passing". In: *technical report* (2008). `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.140.8388` (cit. on p. 8).

[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". In: *Ann. Appl. Probab.* 23.3 (2013), pp. 957–986 (cit. on pp. v, 14, 16–18, 64, 104).

[LPW17] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times*. American Mathematical Society Providence, RI, 2017 (cit. on pp. 36, 37).

[LV99] M. Luby and E. Vigoda. "Fast convergence of the Glauber dynamics for sampling independent sets". In: *Random Structures Algorithms* 15.3-4 (1999), pp. 229–241 (cit. on p. 39).

[LY93] S. L. Lu and H. Yau. "Spectral gap and logarithmic Sobolev inequality for Kawasaki and Glauber dynamics". In: *Communications in Mathematical Physics* 156.2 (1993), pp. 399–433 (cit. on pp. xii, 38).

[McK95] N. McKeown. "Scheduling algorithm for Input Queued Switches". PhD thesis. Univ. of Cal. at Berkeley., 1995 (cit. on p. 6).

[Mit01] M. Mitzenmacher. "The power of two choices in randomized load balancing". In: *Proc. IEEE Tran. Parallel and Distrib. Syst.* Vol. 12. 10. 2001, pp. 1094–1104 (cit. on p. 13).

[Mit92] M. Mitzenmacher. "The power of two choices in randomized load balancing". PhD thesis. Harvard University, 1992 (cit. on p. 13).

[MM79] V. Malyshev and M.V. Men'shikov. "Ergodicity, continuity, and analyticity of countable Markov chains". In: *Trans. Moscow Math. Soc.* 39 (1979), pp. 3–48 (cit. on pp. vii, 22).

[MS16] S. T. Maguluri and R. Srikant. "Heavy traffic queue length behavior in a switch under the maxweight algorithm". In: *Stoch. Syst.* 6.1 (2016), pp. 211–250 (cit. on pp. iv, 20, 25, 93).

[Nev71] J. Neveu. "Potentiel markovien récurrent pour une chaîne de Harris". In: *Ann. Inst. Fourier* 22.2 (1971), pp. 85–130 (cit. on p. 17).

[NTS09] J. Ni, B. Tan, and R. Srikant. "Q-CSMA: Queue-Length-Based CSMA/CA Algorithms for Achieving Maximum Throughput and Low Delay in Wireless Networks". In: *Proc. IEEE/ACM Trans. Netw.* Vol. 20. 3. 2009, pp. 825–836 (cit. on p. 11).

[PSV77] G. C. Papanicolaou, D. Stroock, and S. R. S. Varadhan. "Martingale approach to some limit theorems". In: *Statistical mechanics dynamical systems and the Duke Turbulence Conference Ser., Vol. III* (1977), ii+120 pp. (Cit. on pp. v, 14, 17).

[Puh14] A. L. Puha. "Diffusion limits for shortest remaining processing time queues under nonstandard spatial scaling". In: *Ann. Appl. Probab.* 25.6 (2014), pp. 3381–3404 (cit. on p. 26).

[PV01] E. Pardoux and A. Y. Veretennikov. "On the Poisson equation and diffusion approximation. I". In: *Ann. Probab.* 29.3 (2001), pp. 1061–1085 (cit. on p. 17).

[PV03] E. Pardoux and A. Y. Veretennikov. "On the Poisson equation and diffusion approximation. II". In: *Ann. Probab.* 31.3 (2003), pp. 1166–1192 (cit. on p. 17).

[PV05] E. Pardoux and A. Y. Veretennikov. "On the Poisson equation and diffusion approximation. III". In: *Ann. Probab.* 33.3 (2005), pp. 1111–1133 (cit. on p. 17).

[PW13] O. Perry and W. Whitt. "A fluid limit for an overloaded X model via a stochastic averaging principle." In: *Math. Oper. Res.* 38.2 (2013), pp. 294–349 (cit. on pp. 14, 15).

[Rei84a] M. I. Reiman. "Open Queueing Networks in Heavy Traffic". In: *Math. Oper. Res.* 9.3 (1984), pp. 441–458 (cit. on pp. 19, 25).

[Rei84b] M. I. Reiman. "Some diffusion approximations with state space collapse". In: *Modelling and Performance Evaluation Methodology*. Vol. 60. Berlin, Heidelberg: Springer Berlin Heidelberg, 1984, pp. 207–240 (cit. on p. 19).

[Rev84] D. Revuz. *Markov chains*. Second. Vol. 11. North-Holland Mathematical Library. North-Holland Publishing Co., Amsterdam, 1984, pp. xi+374 (cit. on p. 17).

[Rob03] P. Robert. *Stochastic networks and queues*. Applications of Mathematics (New York) 52. Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2003, pp. xx+398 (cit. on pp. vii, 22).

[Ros11] N. Ross. "Fundamentals of Stein's method". In: *Probab. Surv.* 8 (2011), pp. 210–293 (cit. on p. 19).

[RS92] A. N. Rybko and A. L. Stolyar. "On the ergodicity of random processes that describe the functioning of open queueing networks." In: *translation in Problems Inform. Transmission* 28.3 (1992), pp. 199–220 (cit. on pp. vii, 22).

[RSS08] S. Rajagopalan, D. Shah, and J. Shin. "Aloha that works". In: *submitted, Nov* (2008) (cit. on p. 11).

[RSS09] S. Rajagopalan, D. Shah, and J. Shin. "Network adiabatic theorem: An efficient randomized protocol for contention resolution". In: *Proc. SIGMETRICS/Performance*. Vol. 37. 1. 2009, pp. 133–144 (cit. on pp. v, 9, 14, 44).

[RT98] D. Randall and P. Tetali. "Analyzing Glauber dynamics by comparison of Markov chains". In: *LATIN'98: Theoretical Informatics*. 1998, pp. 292–304 (cit. on pp. xii, 39).

[Rud76] W. Rudin. *Principles of mathematical analysis*. Third. International Series in Pure and Applied Mathematics. McGraw-Hill Book Co., New York-Auckland-Düsseldorf, 1976, pp. x+342 (cit. on p. 79).

[SBB14] F. Simatos, N. Bouman, and S. C. Borst. "Lingering volumes in distributed scheduling." In: *Queueing Syst.* 77.2 (2014), pp. 243–273 (cit. on pp. xxiv, 93, 95).

[SC97] L. Saloff-Coste. "Lectures on finite Markov chains". In: *Lectures on probability theory and statistics (Saint-Flour)* 1665 (1997), pp. 301–413 (cit. on pp. 36, 39).

[Sko61a] A. V. Skorokhod. "Stochastic Equations for Diffusion Processes in a Bounded Region I". In: *Theory Probab. Appl.* 6.3 (1961), pp. 264–274 (cit. on p. 20).

[Sko61b] A. V. Skorokhod. "Stochastic Equations for Diffusion Processes in a Bounded Region II". In: vol. 7. 1. 1961, pp. 3–23 (cit. on p. 20).

[SS12] D. Shah and J. Shin. "Randomized scheduling algorithm for queueing networks". In: *Ann. Appl. Probab* 22.1 (2012), pp. 128–171 (cit. on pp. v, xii, xiii, 9–11, 14, 39–41, 44, 93).

[Ste72] C. Stein. "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables". In: *Proc. Sixth Berkeley Symp. on Math. Statist. and Prob.,* vol. 2. 1972, pp. 583–602 (cit. on p. 18).

[Sto04] A. L. Stolyar. "Max Weight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic." In: *Ann. Appl. Probab.* 14.1 (2004), pp. 1–53 (cit. on p. 93).

[SW12] D. Shah and D. Wischik. "Switched networks with maximum weight policies: fluid approximation and multiplicative state space collapse". In: *Ann. Appl. Probab.* 22.1 (2012), pp. 70–127 (cit. on p. 20).

[Tan79] H. Tanaka. "Stochastic differential equations with reflecting boundary condition in convex regions". In: *Hiroshima Math. J.* 9.1 (1979), pp. 163–177 (cit. on p. 20).

[TE92] L. Tassiulas and A. Ephremides. "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks". In: *Proc. IEEE Trans. Automat. Control.* Vol. 37. 12. 1992, pp. 1936–1948 (cit. on pp. iv, 5, 6).

[VDK96] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. "Queueing sys-tem with selection of the shortest of two queues: An asymptotic approach". In: *Probl. Inf. Transm.* 32.1 (1996), pp. 15–27 (cit. on p. 13).

[Ven+10] P.M. van de Ven, S. C. Borst, J. S. H. van Leeuwaarden, and A. Proutière. "Insensitivity and stability of random-access networks". In: *Performance Evaluation* 67.11 (2010), pp. 1230 –1242 (cit. on pp. 4, 8).

[Ver00] A. Y. Veretennikov. "On large deviations for SDEs with small diffusion and averaging". In: *Stochastic Process. Appl.* 89.1 (2000), pp. 69–79 (cit. on p. 15).

[Ver13] A. Y. Veretennikov. "On large deviations in the averaging principle for SDE's with a "full dependence", revisited [MR1681106]". In: *Discrete Contin. Dyn. Syst. Ser. B* 18.2 (2013), pp. 523–549 (cit. on p. 15).

[Vig01] E. Vigoda. "A note on the Glauber dynamics for sampling independent sets". In: *Electron. J. Combin.* 8.1 (2001), 8 pp. (electronic) (cit. on p. 39).

[Whi02] W. Whitt. *Stochastic-process limits.* Springer Series in Operations Research and Financial Engineering. An introduction to stochastic-process limits and their application to queues. Springer, New York, 2002, pp. xxiv+602 (cit. on pp. vii, 25, 94).

[Wil98] R. J. Williams. "Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse". In: *Queueing Syst.* 30.1-2 (1998), pp. 27–88 (cit. on pp. 19, 20, 25).

[Yun+12] S.-Y. Yun, Y. Yi, J. Shin, and D. Y. Eun. “Optimal CSMA: A survey”. In: *Proc. of ICCS ’12*. 2012, pp. 199–204 (cit. on pp. v, 9).

Résumé — L'objet de cette thèse est de présenter une analyse rigoureuse d'un algorithme de communication. Elle est menée dans le cadre mathématique de la théorie des files d'attente, l'étude des phénomènes de congestion. Le modèle que nous présentons est un raffinement d'un algorithme classique d'ordonnancement. QB-CSMA (Queue-Based Carrier Sense Multiple Access) est un algorithme simple et distribué, créé pour ordonnancer des files d'attente placées sur un graphe. De nouvelles requêtes arrivent aux points d'un processus de Poisson dans les files d'attente des serveurs. Elles requièrent un temps de service exponentiel avant de quitter le réseau. Des nœuds voisins sur le graphe ne peuvent pas servir leurs requêtes simultanément sans interférence. Les termes "Carrier Sense" indique que les serveurs sont capables d'écouter le canal pour voir si leurs voisins sont actifs et évitent de transmettre quand c'est le cas. L'amélioration du CSMA vient de "Queue Based", signifiant que les taux auxquels les serveurs commencent et arrêtent leurs transmissions dépendent de l'état du réseau.

L'algorithme CSMA est simple: chaque noeud a une "fugacité" fixée. Quand il est actif, un noeud se désactive après un temps exponentiel avec paramètre fonction de la fugacité. Les nœuds qui ne sont pas actifs laissent tourner une horloge exponentielle qui s'arrête quand des voisins s'activent. Un nœud s'active quand l'horloge sonne et le taux d'activation est une fonction de la fugacité. Avec QB-CSMA, la fugacité de chaque nœud dépend du nombre de requêtes que le serveur doit traiter. Pour chaque taille des files, il y a une dynamique et une mesure invariante différente pour l'ordonnanceur. Contrairement aux algorithmes CSMA classiques, QB-CSMA n'a pas besoin d'information sur le graphe d'interférence où les taux d'arrivé pour produire de bons choix d'ordonnancement. Dans certains cas, cet algorithme adaptatif peut être utilisé pour approcher Max-Weight, un algorithme coûteux à mettre en place d'un point de vue complexité mais avec de bonnes performances.

Dans le Chapitre 3, nous prouvons des bornes explicites sur la différence entre les moyennes en temps des décisions d'ordonnancement et leurs moyennes pour l'équilibre de la dynamique associée à la taille courante des files. Pour chaque valeur des files d'attente, il y a une dynamique pour l'ordonnanceur et une unique mesure invariante associée à cette dynamique. C'est cette distribution évoluant avec le temps que la mesure d'occupation de l'ordonnanceur approche. Pour prouver ce genre de résultat, ce manuscrit utilise des notions d'analyse fonctionnelle. Les concepts au cœur du raisonnement sont l'équation de Poisson et sa solution. Les solutions de ces équations agissent comme un inverse pour le générateur de processus markoviens et donnent une façon alternative d'écrire la différence entre une fonction et sa moyenne à l'équilibre. Pour borner la moyenne temporelle de la différence entre le taux de service et sa moyenne, il est suffisant de borner la norme des solutions de l'équation de Poisson et connaître leurs régularités par rapport à la taille des files. La norme de la solution à une équation de Poisson est bornée par la constante de log-Sobolev du générateur correspondant. Cette quantité est reliée au temps de mélange de la dynamique associée à ce générateur. Nous prouvons également dans le Chapitre 3 une borne sur la régularité de ces solutions dans la taille des files. Elle dépend essentiellement de la régularité des taux de transitions et de la mesure invariante par rapport à la taille des files.

Dans le Chapitre 4, nous prouvons un théorème limite fonctionnel. La renormalisation fluide usuelle du processus de file d'attente converge vers une fonction

déterministe solution d'une equation différentielle ordinaire. L'étape principale de la preuve repose sur un résultat d'homogénéisation, conséquence des bornes du chapitre précédent. Le problème est que nos bornes se comportent mal quand une taille de file est trop faible. Nous ne sommes pas capables de prouver la convergence que jusqu'au temps où une des files atteint 0 sur l'échelle fluide. La difficulté principale du Chapitre 4 est l'étude des limites fluides pour n'importe quel horizon fini dans le cas du graphe complet (GC). Dans ce cas, il est possible de gérer les réflexions en 0 séparément.

Dans le chapitre 5, pour un GC et un taux d'arrivée critique, nous prouvons un deuxième théorème limite fonctionnel. La limite fluide converge en temps long temps vers un état invariant ou la limite reste constante car les taux d'arrivée et les taux de service moyennés sont en équilibre. A cause de la nature distribuée de QB-CSMA, il y a forcément un temps non nul où aucun des serveurs n'est actif. Le but de l'échelle de temps du Chapitre 5 est d'étudier l'influence de ce temps de repos sur la dynamique des files d'attente. Sur cette deuxième échelle de temps, la limite est également déterministe, ce qui peut être surprenant. Le processus de files d'attente s'effondre instantanément sur une variété de dimension 1 et l'évolution de la somme des coordonnées est donnée par une EDO déterminée par la fraction du temps où aucune file n'est active.

Mots clés : Convergence et comportement en temps long de chaînes et processus de Markov, théorèmes limite fonctionnels, réseaux stochastiques, ensembles indépendants, homogénéisation

---

Abstract — The subject of this thesis is to provide a rigorous analysis of a communication scheme. This analysis is carried out in the mathematical context of queueing theory, the study of congestion phenomena. The model that we study is a refinement of a classical distributed scheduling algorithm. The Queue-Based Carrier Sense Multiple Access (QB-CSMA) algorithm is a distributed algorithm designed to schedule queues on a graph. From a complexity standpoint, its execution is simple and requires little cooperation between nodes. Jobs arrive at the waiting area of queues, also called servers or nodes, along a Poisson process. Each job has an exponential service time before exiting the network. Two or more nodes that are neighbors on the interference graph cannot provide service to their jobs simultaneously without interfering. The "Carrier Sensing" means that nodes can sense when their neighbors are already transmitting, they refrain from using the channel when that happens. The "Queue Based" element is the refinement from the classical CSMA algorithm: it means that the rates at which queues start and end transmissions actually evolve over time and depend on the current state of the network.

The classical CSMA algorithm is quite simple: each node has a fixed "fugacity". When it is active, a node deactivates after an exponential time with a parameter function of the fugacity. Nodes that are not active let an exponential clock run when none of their neighbors are active and stop the clock when their neighbors activate. An activation occurs when the exponential clock ticks and the activation rate determined by the fugacity. With QB-CSMA the fugacity of each server actually depends

on the number of jobs they have to process. For each value of the queue lengths, there is a different dynamic for the schedule, with a unique generator and an invariant measure associated to it. Contrary to the classical, this adaptive CSMA does not require any prior knowledge on the interference graph/arrival rates to be able to produce good service decisions. In some cases, it can be used to distributively approximate the Max-Weight algorithm, a procedure that is onerous to put in place from a complexity point of view but celebrated for its good performance.

In Chapter 3, we prove some explicit bounds on the difference between the time average of the occupation measure of the schedule and its steady state average associated with the current value of the queue lengths. For each value of queue lengths, there is a dynamic for the schedule, this dynamic has a unique invariant measure. This is the time evolving steady state average that the occupation measure approaches. In order to prove this, this manuscript uses tools from functional analysis. The main concepts that we use are Poisson equations and their solutions. The solutions to Poisson equations act as an inverse application for the generator of a Markov process and give an alternate way to write the difference between a function and its average with respect to the invariant measure of the generator. To bound the time average of the difference between the service rate and its steady state average, it is sufficient to bound solutions to Poisson equations and understand their regularity in the size of queue lengths. The norm of solutions to Poisson equations can be bounded by the log-Sobolev constant of the generator of the schedule with fixed queue lengths. This quantity is closely related to the mixing time of this dynamic. Some bounds on the regularity of solution to Poisson equations are also proved in Chapter 3. The regularity of solutions essentially depends on the regularity in the size of the queues for the transition rates of the schedule and for the invariant measure.

In Chapter 4, we prove a functional limit theorem. The usual fluid limit scaling of the queue lengths converges to a deterministic process governed by an ODE. The main part of the proof is a homogenization result proven from the bounds of Chapter 3. The problem is that this bound does not behave well when some coordinates are too small. Because of that the result on a general interference graph is only proved up to the time one of the queue reaches 0 in the fluid scale. The main difficulty in Chapter 4 lies on the study of the Complete Interference Graph (CIG) over any finite time interval. In this case the possible reflections at 0 are treaded separately. Three cases are distinguished between sub-critical, super-critical and critical arrival rates.

In Chapter 5, in the case of a CIG with critical arrival rates a second functional limit theorem is proved. The fluid limit converges for long time to an invariant state where the process is constant because the arrival rates and the averaged departure rates coincide. Because of the distributed nature of the algorithm, there is always some non-null time where no queue is active between two activations. The idea behind the time scale in Chapter 5 is to investigate the influence of idle time to the dynamic. On this second time scale, the limit is also deterministic which is surprising. The process of queue lengths instantaneously collapses to a one dimensional manifold and the evolution of the sum of coordinates is given by an ODE determined by the idle time.

Keywords: convergence and long time behavior of Markov processes, functional limit theorems, homogenization, stochastic networks, distributed scheduling,

independent sets

ISAE-Supaero, DISC 10 avenue Edouard Belin
Université Paul Sabatier, IMT, équipe Statistiques et Probabilités 118 Route de Narbonne
31400, Toulouse