Thèse présentée pour obtenir le grade de

DOCTEUR DE L'ÉCOLE POLYTECHNIQUE

spécialité : Mathématiques

par

Jean-Marc LASGOUTTES

Contribution à l'analyse de réseaux stochastiques et application aux systèmes de transport

soutenue le 30 Mars 1995 devant le jury composé de:

M. NEVEU	Président,
M. FAYOLLE	Directeur de thèse,
M. BOXMA	Rapporteur,
M. SCHASSBERGER	Rapporteur,
M. BACCELLI	Examinateur,
M. BRÉMAUD	Examinateur,
M. MALYSHEV	Examinateur.

Contribution à l'analyse de réseaux stochastiques et application aux systèmes de transport

Jean-Marc LASGOUTTES

Guy FAYOLLE *m'a suivi avec une rare assiduité durant toute la période de ma thèse ; il a su m'aider à exprimer et mettre en forme nombre des idées présentées ici au cours de discussions parfois mouvementées mais toujours constructives ;*

Jacques NEVEU *m'a fait le plaisir de présider mon jury de thèse et surtout, en faisant passer les probabilités pour un jeu d'enfant, m'a donné l'envie de découvrir ce qui se cache derrière ;*

Onno BOXMA *et Rolf* SCHASSBERGER *ont examiné ce travail avec soin, sans se laisser rebuter par la diversité des sujets abordés, qui pourrait, à première vue, passer pour du fouillis ;*

Pierre BRÉMAUD *a guidé patiemment mes premiers pas chancelants dans la recherche pendant mon* DEA *et le début de ma thèse ;*

François BACCELLI *et Vadim* MALYSHEV *ont influencé une partie de ce travail au travers des cours et des ouvrages dont ils m'ont fait profiter ;*

Michel PARENT *et l'équipe de* PRAXITÈLE *m'ont permis de découvrir le monde étonnant du* VUPI *et m'ont fourni une source de problèmes nouveaux et passionnants ;*

Les autres membres présents et passés du projet MÉVAL — *Franck* DELCOIGNE, *Vincent* DUMAS, *Christine* FRICKER *et Philippe* ROBERT — *m'ont abreuvé de discussions, cafés et cigarettes sans lesquels la vie serait bien morne ;*

Danny MOREAU *m'a montré ce que c'est de pouvoir compter sur quelqu'un ;*

*Le personnel de l'*INRIA *a toujours su faire preuve de disponibilité et d'efficacité aux moments où j'en avais besoin ;*

Vivek BADRINATH *a été un compagnon de route irremplaçable bien qu'éphémère ;*

Danièle et Daniel LASGOUTTES *ont suivi avec moi ce long chemin accidenté ;*

Nathalie LASGOUTTES *lit ce que je suis en train d'écrire, mais de toute façon elle lit aussi dans mes pensées et sait ce que je n'écris pas.*

A tous ceux-là et à tous les autres, je dis :

« Merci. »

Avant-propos

Il est des fois où la demande en applications industrielles va de pair avec la définition de modèles attrayants pour les mathématiciens. C'est sans doute le cas ici. D'une part, la nécessité de mettre en œuvre de nouveaux moyens de transport pour décongestionner les zones urbaines s'est imposée non seulement aux États-Unis, mais aussi en France, où l'INRIA participe au projet PRAXITÈLE : il s'agit de gérer une flotte de petits véhicules en libre service qui permette de limiter l'usage de la voiture individuelle pour des déplacements locaux. D'autre part, la problématique liée à la modélisation stochastique de telles flottes est intéressante, en ce sens qu'elle fournit des modèles nouveaux, souvent difficiles à analyser et d'une portée plus grande qu'il n'y semblait au premier abord.

Le premier problème étudié est celui d'un taxi qui se déplace dans une ville, où il sert des clients qui attendent dans des stations. Ce taxi peut être représenté par un « système à polling » (scrutin). Ces systèmes ont connu une grande popularité ces dernières années, qui peut s'expliquer par le fait que ce sont des marches aléatoires en grande dimension qu'on peut miraculeusement analyser en partie : on sait, en général, donner des conditions d'ergodicité et même calculer des temps moyens d'attente aux files. Ce qui différencie le « modèle du taxi » de ses prédécesseurs est que son déplacement dépend de l'état des files visitées. Ceci supprime les propriétés de *monotonie* qui sont cruciales pour étudier la stabilité. On donne ici des conditions d'ergodicité et de non ergodicité, et on montre qu'elles coïncident notamment quand seules les durées de trajet dépendent de l'état. La méthode repose sur l'étude du système dynamique associé à la marche aléatoire. Enfin, on montre, sous certaines hypothèses de symétrie, comment calculer le temps moyen d'attente.

Le modèle décrit dans le deuxième chapitre comprend plusieurs véhicules, mais les clients sont rejetés s'il ne peuvent être servis tout de suite. Cela conduit à un autre outil classique : le réseau fermé à « forme produit », dont la distribution à l'équilibre est connue explicitement. De nombreux réseaux de ce type sont suffisamment grands pour que leur description *microscopique* ne soit pas suffisante pour comprendre leur comportement *macroscopique*. Cette constatation, qui est d'ailleurs un des fondements de la mécanique statistique, conduit tout naturellement à regarder ce qui se passe quand le réseau « grandit », aussi bien en termes de nombre de files que de clients. Les méthodes existantes pour traiter cette question ont le défaut d'imposer trop de contraintes sur les files elles-mêmes, alors que le réseau possède des propriétés de mélange qui gomment certains comportements individuels. On développe ici une nouvelle approche, inspirée de celle du théorème de la limite centrale, qui permet de conclure à partir de conditions plus globales et donc plus générales. Le résultat principal est que le réseau se comporte asymptotiquement comme un ensemble de files indépendantes, dont certaines peuvent devenir des goulots d'étranglement. De plus, on montre que le nombre de clients et de nœuds du réseaux ne doivent pas nécessairement être liés par une relation de proportionnalité.

Le troisième chapitre de cette thèse occupe une place un peu à part, puisqu'il n'a pas été motivé par l'étude de modèles de transport. Il s'agit en fait de mesurer, lors de la simulation d'une file G/G/1, la dépendance d'une certaine famille de mesures de performance en fonction d'un paramètre du temps de service ou d'interarrivée. Pour cela, on montre comment la méthode dite d'« analyse de perturbations infinitésimale » peut être appliquée directement sur une file dont le processus est stationnaire ergodique, et cela grâce au formalisme des mesures de Palm. On donne des estimateurs des dérivées première et seconde de la performance moyenne, qui restent valables même si la grandeur mesurée n'est pas dérivable par rapport au paramètre. On montre enfin comment l'approche utilisée s'étend à certains paramètres du processus des arrivées.

Sommaire

Chapitre 1

Un système à polling avec routage dépendant de l'état

La problématique présentée dans ce chapitre est issue de la modélisation d'un taxi servant des clients arrivant à n stations réparties dans une zone géographique. On suppose que les clients arrivent à chaque file selon un processus de Poisson composé et choisissent leur destination selon une matrice de routage ; les distances parcourues sont décrites en termes de durées de trajet. Quand le taxi arrive à une station vide, il part en maraude en tirant au sort ses destinations successives selon une autre matrice de routage. La durée du trajet entre deux stations quelconques peut alors être complètement différente.

Dans la terminologie des files d'attente, les modèles où plusieurs files se partagent un seul serveur sont appelés « systèmes à polling » (scrutin). Le membre le plus simple de cette classe est le « réseau à jeton », où plusieurs nœuds d'un réseau de communication émettent à tour de rôle sur un canal commun. De nombreuses variations sur ce thème ont fait leur apparition ces vingt dernières années et Takagi [56] cite plus de 400 articles sur ce sujet entre 1970 et 1990 ! La plupart de ces travaux visent à analyser le temps moyen d'attente des clients ou, quand ce n'est pas possible, à en trouver une approximation. S'il existe de très nombreuses disciplines de *routage* (la façon dont se déplace le serveur) et de *service* (le nombre de clients servis à chaque visite à une file), les méthodes utilisées restent similaires dans un grand nombre de cas. Très récemment, des conditions nécessaires et suffisantes d'ergodicité de systèmes à polling ont été données, notamment dans les articles de Fricker et Jaïbi [19, 20] et Borovkov et Schassberger [3].

Les résultats présentés dans ce chapitre concernent aussi bien l'ergodicité du modèle que le calcul de temps moyen d'attente. La particularité de ce « modèle du taxi » est que les règles de déplacement du serveur dépendent de l'état des files visitées, ce qui crée des difficultés supplémentaires :

– la variable aléatoire décrivant la position du serveur n'est plus une chaîne de Markov indépendante de l'état du système ;

– la plupart des méthodes pour déterminer les conditions d'ergodicité des systèmes à polling utilisent des propriétés de monotonie stochastiques qui peuvent ne pas être vérifiées ici ;

– le calcul du temps moyen d'attente des clients est beaucoup plus complexe que dans les cas classiques et les méthodes habituelles ne sont plus adaptées.

L'étude du modèle du taxi comporte trois étapes. Dans un premier temps, on montre, lorsque le système est ergodique, comment calculer la probabilité stationnaire que le serveur soit à une file donnée et que cette file soit vide/pleine (Théorème 1.3.1). La détermination de grandeurs de ce type est en général impossible pour les modèles généraux avec routage dépendant de l'état, ce qui rend illusoire la détermination de conditions d'ergodicité explicites. Dans un deuxième temps, on utilise la théorie des systèmes dynamiques (voir par exemple le livre de Fayolle, Malyshev et Menshikov [15]) pour déterminer une condition suffisante d'ergodicité (Théorème 1.4.2). La méthode utilisée ressemble à celle développée dans Fayolle et Zamyatin [16] : on montre que le *second* champ de vecteur peut être obtenu comme solution d'un système linéaire, et la condition d'ergodicité est obtenue en appliquant une fonction de Lyapounov linéaire sur le système dynamique associé. Cette condition est aussi nécessaire dans le cas où les deux matrices de routage sont identiques. Enfin, le temps moyen d'attente est calculé dans le cas où le modèle présente une propriété de symétrie rotationnelle (Théorèmes 1.5.2 et 1.6.1). Cette hypothèse supplémentaire est indispensable pour tous les systèmes à polling où le nombre maximal de clients pouvant être servis à chaque visite à une file est limité. Toutefois, la méthode de résolution utilisée permet de garder une forme de routage Markovien et d'unifier différents résultats antérieurs.

Ces travaux ont été réalisés en collaboration avec Guy Fayolle et ont été publiés par l'*Institute for Mathematics and its Applications* [14] (voir aussi [12]). Un problème naturel non traité est celui de la convergence faible d'un tel modèle vers un polling *continu* sur un graphe lorsque le nombre de files croit vers l'infini, pour des normalisations convenables des différents paramètres. Ceci fera le sujet de recherches futures.

1.1 Introduction

Consider a taxicab in a city in which there are N stations at which clients arrive and wait for the vehicle. When their turn has come to be served, they ask for a transit to a destination (one of the other stations) where they leave the system. Whenever the taxicab finds a station empty, it goes somewhere else to look for a client. The choice of the destinations by a client or by the empty taxicab are made via the two distinct routing matrices P and $\widetilde{P}$.

This system can also be seen as a polling model with N queues at which customers arrive and one server which visits the queues according to the following rules: if the server finds a client at the current queue, it serves this client and chooses a new queue according to the routing matrix P; otherwise it selects the next queue according to $\widetilde{P}$. This polling scheme is an extension of the classical Markovian polling model, with routing probabilities depending on the state of the last visited queue. As a taxicab only takes one client at a time, the service strategy is the 1-*limited* strategy, which is known to be more difficult to analyze than either the *gated* or *exhaustive* strategies—in which the server respectively serves the clients present on its arrival or serves clients until the queue is empty.

Since the bibliography on polling models is plethoric, we refer the reader to the references given in Takagi [56] and Levy and Sidi [38]. Among the studies related to our work, we can cite Kleinrock and Levy [31], who compute waiting times in random polling models in which the destination of the server is chosen independently of its provenance, and Boxma and Weststrate [4] who give a pseudo-conservation law for a model with Markovian routing. Ferguson [18] and Bradlow and Byrd [5] study approximatively a model where the switching time is station-dependent. Srinivasan [53] analyzes a polling system with the same routing policy as in our model, and with different service policies at each queue. However, the waiting times are not computed in the case where the routing is state dependent. Another important issue concerns state-independent ergodicity conditions. In this case, necessary and sufficient conditions were obtained by Georgiadis and Szpankowski [22] and Fricker and Jaïbi [19, 20] for deterministic and Markovian routing with various service disciplines. Borovkov and Schassberger [3] and Fayolle and Zamyatin [16] give necessary and sufficient conditions for a system with Markovian routing and 1-limited service. Finally, Schassberger [50] gives a necessary condition for the ergodicity of a polling model with 1-limited service and a routing that depends on the whole state of the system.

Here, we give a new method to get the stationary distribution of the position of the server and, in the case of a fully symmetrical system, a way to compute the mean waiting time of a customer. This extends known results on 1-limited polling models, especially for the symmetrical case. However, we do not provide a pseudo-conservation law as for example Boxma and

Weststrate [4], since the computations are not as simple as in the state-independent routing case. Our method of proof allows to make only minimal assumptions on the arrival process and applies to either discrete or continuous time models. Moreover, in the symmetrical case the results allow to compare various polling strategies.

The chapter is organized as follows: in Section 1.2, we present the model and give a functional equation describing its evolution. In Section 1.3, stationary probabilities of the position of the server are obtained and Section 1.4 is devoted to the general problem of ergodicity. Section 1.5 presents formulas to compute the first moment of the queue length at polling instants. These results are used in Section 1.6 to calculate the waiting time of an arbitrary customer. Applications to known models are also given.

1.2 Description of the model

The system consists of N stations at which clients arrive according to a stationary process. Let $\mathcal{S} \stackrel{\text{def}}{=} \{1, \ldots, N\}$ be the set of stations. Assume that the server arrives after $n-1$ moves at station $i \in \mathcal{S}$ where a client is waiting. Then the server loads this client and goes to station $j \in \mathcal{S}$ with probability $p_{i,j}$, such that $p_{i,1} + \cdots + p_{i,N} = 1$. Conversely, if station i is empty, the server polls station j with probability $\tilde{p}_{i,j}$. The service policy is known as *1-limited* policy, since at most 1 customer is served each time a station is visited. The two transition matrices will be denoted respectively by

$$P = (p_{i,j})_{i,j \in \mathcal{S}}, \quad \widetilde{P} = (\tilde{p}_{i,j})_{i,j \in \mathcal{S}}.$$

The number of new customers arriving to station $q \in \mathcal{S}$ between the polling of stations i and j when there have been a service (resp. no service) at station i is $B_{i,j;q}(n)$ (resp. $\widetilde{B}_{i,j;q}(n)$). The vectors $B_{i,j}(n) = (B_{i,j;1}(n), \ldots, B_{i,j;N}(n))$ (resp. $\widetilde{B}_{i,j}(n)$) are i.i.d. for different n, but their components may be dependents and not identically distributed. In the classical polling terminology, $B_{i,j}(n)$ is the number of arrivals during *both* the service and the switchover and $\widetilde{B}_{i,j}(n)$ is the number of arrivals during a switchover. However, our model covers a wider class of applications. In particular, the service time may depend on both i and j. In addition, the switchover time distribution may depend on the state of the last visited queue (empty or not). Remark also that, unlike in Srinivasan [53], we can also have $\mathbb{E}B_{i,j}(n) < \mathbb{E}\widetilde{B}_{i,j}(n)$, which means that serving a customer can actually take less time than just moving. This would amount to *negative* service times in the usual polling formulation, which are not easy to handle!

It is worth noting that our setting is valid for both discrete-time and continuous-time models with or without batch arrivals. We do not need a separate analysis for each case, until the computation of waiting times, where we will have to give precise definitions of the vectors $B_{i,j}(n)$ and

$\widetilde{B}_{i,j}(n)$. We do not describe yet the exact arrival process and the time taken to travel from one station to another, since we only need the distribution of the number of clients arriving during such a move. However, due to the assumption of independence between successive arrivals, the model applies mainly to the following situations:

- discrete-time evolution, when a batch of customers arrives at each station at the beginning of each time slot; batches are i.i.d. with respect to the time slots but they can be correlated between stations;

- continuous time evolution, when the arrival process is a compound Poisson process: customers arrive at the system in batches at the epochs of a Poisson process and the batches have the same properties as in the previous case.

Let $X_i(n)$ be the number of clients waiting at station i at polling instant n and $S(n)$ be the corresponding position of the server. If we define $X(n) \stackrel{\text{def}}{=} (X_1(n), \ldots, X_N(n))$, then $\mathcal{L} \stackrel{\text{def}}{=} (S, X)$ forms a Markov chain. Throughout this chapter, we will assume that this chain is irreducible. For any vector $\vec{z} = (z_1, \ldots, z_N) \in \mathcal{D}^N$ (where $\mathcal{D}$ denotes the unit disc in the complex plane), we note $\vec{z}^B = z_1^{B_1} \cdot z_2^{B_2} \cdots z_N^{B_N}$ and define the following generating functions:

$$
\begin{aligned}
a_{i,j}(\vec{z}) &\stackrel{\text{def}}{=} \mathbb{E}[\vec{z}^{B_{i,j}(n)}], \\
\tilde{a}_{i,j}(\vec{z}) &\stackrel{\text{def}}{=} \mathbb{E}[\vec{z}^{\widetilde{B}_{i,j}(n)}], \\
F_i(\vec{z}; n) &\stackrel{\text{def}}{=} \mathbb{E}[\vec{z}^{X(n)} \mathbb{1}_{\{S(n)=i\}}], \\
\widetilde{F}_i(\vec{z}; n) &\stackrel{\text{def}}{=} \mathbb{E}[\vec{z}^{X(n)} \mathbb{1}_{\{S(n)=i, X_i(n)=0\}}], \\
&= F_i(z_1, \ldots, z_{i-1}, 0, z_{i+1}, \ldots, z_N; n), \\
F_i(\vec{z}) &\stackrel{\text{def}}{=} \lim_{n \to \infty} F_i(\vec{z}; n), \\
\widetilde{F}_i(\vec{z}) &\stackrel{\text{def}}{=} \lim_{n \to \infty} \widetilde{F}_i(\vec{z}; n),
\end{aligned}
$$

where $\mathbb{1}_{\mathcal{E}}$ is as usual the indicator function of the set $\mathcal{E}$. The following result holds:

Theorem 1.2.1 *Let $A(\vec{z})$, $\widetilde{A}(\vec{z})$ and $\Delta(\vec{z})$ be matrices defined as follows: for all $i, j \in \mathcal{S}$, the elements of row i and column j are given by*

$$
\begin{aligned}
[A(\vec{z})]_{i,j} &= p_{j,i} a_{j,i}(\vec{z}), \\
[\widetilde{A}(\vec{z})]_{i,j} &= \tilde{p}_{j,i} \tilde{a}_{j,i}(\vec{z}), \\
[\Delta(\vec{z})]_{i,j} &= \frac{1}{z_i} \mathbb{1}_{\{i=j\}}.
\end{aligned}
$$

Then the vectors $F(\vec{z}) = (F_1(\vec{z}), \ldots, F_N(\vec{z}))$ and $\widetilde{F}(\vec{z}) = (\widetilde{F}_1(\vec{z}), \ldots, \widetilde{F}_N(\vec{z}))$ are related by the functional equation

$$
[I - A\Delta(\vec{z})] F(\vec{z}) = [\widetilde{A}(\vec{z}) - A\Delta(\vec{z})] \widetilde{F}(\vec{z}), \tag{1.2.1}
$$

where $A\Delta(\vec{z})$ *stands for* $A(\vec{z})\Delta(\vec{z})$.

Proof We have, for all $i, j \in \mathcal{S}$ and $n > 0$,

$$\begin{aligned}
&\mathbb{E}[\vec{z}^{X(n+1)}\mathbb{1}_{\{S(n)=i,S(n+1)=j\}}] \\
&\qquad = \mathbb{E}[\vec{z}^{X(n)+B_{i,j}(n)-\vec{e}_i}\mathbb{1}_{\{S(n)=i,S(n+1)=j,X_i(n)>0\}}] \\
&\qquad\quad +\mathbb{E}[\vec{z}^{X(n)+\widetilde{B}_{i,j}(n)}\mathbb{1}_{\{S(n)=i,S(n+1)=j,X_i(n)=0\}}] \\
&\qquad = p_{i,j}a_{i,j}(\vec{z})\mathbb{E}[\vec{z}^{X(n)-\vec{e}_i}\mathbb{1}_{\{S(n)=i,X_i(n)>0\}}] \\
&\qquad\quad +\tilde{p}_{i,j}\tilde{a}_{i,j}(\vec{z})\mathbb{E}[\vec{z}^{X(n)}\mathbb{1}_{\{S(n)=i,X_i(n)=0\}}] \\
&\qquad = p_{i,j}a_{i,j}(\vec{z})\frac{F_i(\vec{z};n)-\widetilde{F}_i(\vec{z};n)}{z_i}+\tilde{p}_{i,j}\tilde{a}_{i,j}(\vec{z})\widetilde{F}_i(\vec{z};n).
\end{aligned}$$

The second equality above uses the independence of the routing and the arrivals with respect to the past. Summing over all possible i, we get

$$F_j(\vec{z};n+1)=\sum_{i=1}^{N}p_{i,j}a_{i,j}(\vec{z})\frac{F_i(\vec{z};n)-\widetilde{F}_i(\vec{z};n)}{z_i}+\sum_{i=1}^{N}\tilde{p}_{i,j}\tilde{a}_{i,j}(\vec{z})\widetilde{F}_i(\vec{z};n),$$

so that, letting $n\to\infty$,

$$F_j(\vec{z})-\sum_{i=1}^{N}\frac{p_{i,j}a_{i,j}(\vec{z})}{z_i}F_i(\vec{z})=\sum_{i=1}^{N}\tilde{p}_{i,j}\tilde{a}_{i,j}(\vec{z})\widetilde{F}_i(\vec{z})-\sum_{i=1}^{N}\frac{p_{i,j}a_{i,j}(\vec{z})}{z_i}\widetilde{F}_i(\vec{z}),$$

which is equivalent to (1.2.1). ■

Defining, when it exists,

$$D(\vec{z})\stackrel{\text{def}}{=}[I-A\Delta(\vec{z})]^{-1}[\widetilde{A}(\vec{z})-A\Delta(\vec{z})], \tag{1.2.2}$$

one sees that (1.2.1) can be rewritten as

$$F(\vec{z})=D(\vec{z})\widetilde{F}(\vec{z}). \tag{1.2.3}$$

This functional equation contains all the information sufficient to characterize $F(\vec{z})$ and $\widetilde{F}(\vec{z})$. Although its solution for $N\geq 3$ is still an open question, partial results can be derived as shown in the following sections.

1.3 The stationary distribution of the position of the server

What renders the polling model presented in the previous section difficult to analyze is, among other things, the fact that the movements of the server depend on the state of the visited stations. In particular, $\{S(n)\}_{n\geq 0}$ is not a Markov process, as it would be when $p_{i,j}=\tilde{p}_{i,j}$ for all i and j. Some computations are needed to get the stationary probabilities of the position of the server, that is

$$\begin{aligned} F_i(\vec{e}) &= P(S=i) \\ \widetilde{F}_i(\vec{e}) &= P(S=i, X_i=0). \end{aligned}$$

It appears that these stationary probabilities depend not only on the transition probabilities, but also on the following mean values:

$$\alpha_{i;q} \stackrel{\text{def}}{=} \sum_{j=1}^{N} p_{i,j} \mathbb{E} B_{i,j;q}(n), \tag{1.3.1}$$

$$\tilde{\alpha}_{i;q} \stackrel{\text{def}}{=} \sum_{j=1}^{N} \tilde{p}_{i,j} \mathbb{E} \widetilde{B}_{i,j;q}(n). \tag{1.3.2}$$

Here $\alpha_{i;q}$ (resp. $\tilde{\alpha}_{i;q}$) is the mean number of new clients at station q between the arrival of the server at station i which was non-empty (resp. empty) and its arrival at the next (arbitrary) polled station. Since the arrival process is stationary, it is possible to define for all $q \in \mathcal{S}$ the mean number λ_q of arrivals at station q per unit of time and write, with obvious definitions for τ_i and $\tilde{\tau}_i$,

$$\alpha_{i;q} = \lambda_q \tau_i, \tag{1.3.3}$$

$$\tilde{\alpha}_{i;q} = \lambda_q \tilde{\tau}_i. \tag{1.3.4}$$

Throughout this chapter we will sometimes write F (resp. $\widetilde{F}$) instead of $F(\vec{e})$ (resp. $\widetilde{F}(\vec{e})$) in order to shorten the notation. Moreover, ${}^*F$ denotes the transpose of the vector F.

Theorem 1.3.1 *Define the matrices $\mathcal{A} \stackrel{\text{def}}{=} (\alpha_{i;q})$ and $\widetilde{\mathcal{A}} \stackrel{\text{def}}{=} (\tilde{\alpha}_{i;q})$. When the Markov chain $\mathcal{L}$ is ergodic, F and $\widetilde{F}$ satisfy the following linear system of equations:*

$${}^*F\, \vec{e} = 1, \tag{1.3.5}$$

$${}^*F\, [I-P] = {}^*\widetilde{F}\, [\widetilde{P} - P], \tag{1.3.6}$$

$${}^*F\, [I-\mathcal{A}] = {}^*\widetilde{F}\, [I - \mathcal{A} + \widetilde{\mathcal{A}}]. \tag{1.3.7}$$

Define

$$\hat{\rho} \stackrel{\text{def}}{=} \sum_{i=1}^{N} \lambda_i (\tau_i - \tilde{\tau}_i).$$

When $\hat{\rho} \neq 1$, Equations (1.3.6) and (1.3.7) can be rewritten as

$$F_j = \sum_{i \in \mathcal{S}} \tilde{p}_{i,j} F_i + \bar{\tau} \sum_{i \in \mathcal{S}} \lambda_i (p_{i,j} - \tilde{p}_{i,j}), \quad \textit{for all } j \in \mathcal{S}, \tag{1.3.8}$$

$$\bar{\tau} = \frac{1}{1-\hat{\rho}} \sum_{j \in \mathcal{S}} F_j \tilde{\tau}_j, \tag{1.3.9}$$

and $\widetilde{F}$ is given by

$$\widetilde{F}_j = F_j - \lambda_j \bar{\tau}. \tag{1.3.10}$$

Moreover the mean time between two consecutive visits to queue j is

$$\mathbb{E}T_{j,j} = \frac{\bar{\tau}}{F_j} \tag{1.3.11}$$

The quantity $\bar{\tau}$ defined in (1.3.9) can be seen as the mean time between two polling instants. Note that, with the above definitions, $\hat{\rho}$ is in general different from the classical ρ defined in queueing theory. However, they coincide if $P = \widetilde{P}$ and the time between two polling instants can be decomposed into a state-independent switchover time and a service time which only depends on the station where the customer is.

Proof of Theorem 1.3.1 Although this theorem could be proved analytically using the functional equation (1.2.1), we give here a simple probabilistic interpretation of (1.3.6) and (1.3.7). Indeed, when the system is ergodic, (1.3.6) follows directly from

$$P(S=j) = \sum_{i=1}^{N} \Big[P(S=i, X_i > 0) p_{i,j} + P(S=i, X_i = 0) \tilde{p}_{i,j} \Big].$$

Moreover, writing the equality of the outgoing and ingoing flows at station q gives

$$\begin{aligned} P(S=q, X_q > 0) \quad = \quad & \sum_{i=1}^{N} \sum_{j=1}^{N} \Big[P(S=i, X_i > 0) p_{i,j} \mathbb{E}B_{i,j;q}(n) \\ & + P(S=i, X_i = 0) \tilde{p}_{i,j} \mathbb{E}\widetilde{B}_{i,j;q} \Big], \end{aligned}$$

which is equivalent to (1.3.7) if we take into account (1.3.1) and (1.3.2). It is straightforward to see, using (1.3.3) and (1.3.4), that, when $\hat{\rho} \neq 1$,

$$[I - \mathcal{A}][I - \mathcal{A} + \widetilde{\mathcal{A}}]^{-1} = I - \frac{\widetilde{\mathcal{A}}}{1 - \hat{\rho}}.$$

With this relation and $\bar{\tau}$ as in (1.3.9), Equations (1.3.8) and (1.3.10) follow easily from (1.3.6) and (1.3.7). Let $N_{j,j;j}$ be the number of arrivals to a queue j between two consecutive visits of the server. The equality of flows reads

$$P(X_j > 0 \mid S = j) = \mathbb{E}N_{j,j;j}.$$

Since the last equation can be rewritten as

$$\frac{F_j - \widetilde{F}_j}{F_j} = \lambda_j \mathbb{E}T_{j,j} \ ,$$

we obtain (1.3.11) and the proof of the theorem is concluded. ∎

The set of equations (1.3.5), (1.3.8), (1.3.9) and (1.3.10) provides a convenient way to compute F and $\widetilde{F}$. The following proposition gives simple conditions under which its rank is full.

Proposition 1.3.2 *The linear system given by (1.3.5), (1.3.8) and (1.3.9) has a unique solution when $\widetilde{P}$ has exactly one essential class. When $\widetilde{P}$ has $K > 1$ essential classes $\mathcal{E}_1, \ldots, \mathcal{E}_K$, the Markov chain $\mathcal{L}$ is never ergodic unless*

$$\sum_{j\in\mathcal{E}_m}\sum_{i\in\mathcal{S}} \lambda_i(p_{i,j} - \tilde{p}_{i,j}) = 0, \quad \textit{for all } m \le K. \tag{1.3.12}$$

Proof Define $\psi_i \stackrel{\text{def}}{=} F_i/\bar{\tau}$. Then the system (1.3.8)–(1.3.9) reads

$$\psi_j - \sum_{i\in\mathcal{S}} \tilde{p}_{i,j}\psi_i = \sum_{i\in\mathcal{S}} \lambda_i(p_{i,j} - \tilde{p}_{i,j}), \quad \text{for all } j \in \mathcal{S}, \tag{1.3.13}$$

$$\sum_{i\in\mathcal{S}} \psi_i\tilde{\tau}_i = 1 - \hat{\rho}. \tag{1.3.14}$$

The system (1.3.13) have rank $N-1$ unless the stochastic matrix $\widetilde{P}$ has several essential classes. Then, for each $m \le K$, $i \in \mathcal{E}_m$ and $j \notin \mathcal{E}_m$, $\tilde{p}_{i,j} = 0$. In this case, (1.3.13) has solutions only when (1.3.12) holds. ■

The conditions (1.3.12) represent in some sense "zero drift" relationships which, as usual, imply more involved derivations. However, one can show that, when the arrivals form independent compound Poisson processes at each queue, the system is never ergodic! The proof is of analytic nature. It requires Taylor expansions of second order in equation (1.2.1), similar to those extensively used in Section 1.5. In general, when the batches are correlated, it is difficult to conclude and we suspect that all situations might occur (ergodicity, null recurrence or transience). Since this proof is rather intricate and does not give any insight on the behaviour of the system, it is omitted here.

It is worth noting the special role played by $\widetilde{P}$. In particular, when the arrivals are Poisson, the Markov chain is *never* ergodic if $\widetilde{P}$ admits several essential classes (obviously, P has then to be chosen to ensure the irreducibility of $\mathcal{L}$). In our opinion, this result is not intuitive and cannot be explained just by waving hands.

1.4 Conditions for ergodicity

The purpose of this section is to classify the process $\mathcal{L}$, viewed as a random walk on $\mathcal{S} \times \mathbb{Z}_+^N$, in terms of ergodicity and non-ergodicity.

1.4.1 Necessary condition

We give a necessary condition for ergodicity which is a simple consequence of the results of Theorem 1.3.1. When the system is ergodic, we have $\widetilde{F}_i > 0$ for all $i \in \mathcal{S}$ and Equation (1.3.10) implies the following.

Theorem 1.4.1 *If the Markov chain $\mathcal{L}$ is ergodic, then*

$$\hat{\rho} < 1, \tag{1.4.1}$$

$$\lambda_i \bar{\tau} < F_i, \quad \text{for all } i \in \mathcal{S}. \tag{1.4.2}$$

These conditions extend in the state-dependent case the results obtained in [3]. When they hold, we prove another useful inequality. Indeed, instantiating (1.4.2) in (1.3.9) yields

$$\bar{\tau} > \frac{\sum_{j \in \mathcal{S}} \bar{\tau} \lambda_j \tilde{\tau}_j}{1 - \hat{\rho}},$$

or, equivalently,

$$1 - \sum_{j \in \mathcal{S}} \lambda_j \tau_j > 0. \tag{1.4.3}$$

1.4.2 Sufficient condition

Let us emphasize that this part could be skipped by readers not strongly interested (if any!) in ergodicity. The mathematical understanding requires the reading of Appendix 1.A.1, which summarizes deep works of [42], [15] and [16]. The approach relies on the study of a dynamical system associated to $\mathcal{L}$. For the sake of readability, we recall hereafter two main notions.

- *Induced chain $\mathcal{L}^\wedge$.* It is a Markov chain corresponding to a polling system in which the queues belonging to the *face* $\wedge \subset \mathcal{S}$ are kept saturated. From a purely notational point of view, the original system would correspond to the case $\wedge = \emptyset$. The behaviours of $\mathcal{L}$ and $\mathcal{L}^\wedge$ are not *directly* connected to each other.

- *Second vector field $\vec{v}^\wedge$.* One can imagine that the random walk starts from a point which is close to $\wedge$, but sufficiently far from all other faces $\wedge'$, with $\wedge \not\subset \wedge'$. After some time—sufficiently long, but less than the minimal distance from $\wedge'$—, the stationary regime in the induced chain will be installed. In this regime, one can ask about the mean drift along $\wedge$: it is defined exactly by $\vec{v}^\wedge$.

As in Theorem 1.3.1, the stationary position of the server for any ergodic induced chain can be obtained as the solution of a linear system consisting of $N+1$ equations. Indeed, for all $t \in \mathcal{S}$, $j \notin \wedge$, we have

$$\begin{aligned}
\pi^\wedge(t) &= \sum_{s\in\wedge} \pi^\wedge(s) p_{s,t} + \sum_{s\notin\wedge} \pi^\wedge(s)\tilde{p}_{s,t} \\
&\quad + \sum_{s\notin\wedge} \pi^\wedge(s, x_s > 0)(p_{s,t} - \tilde{p}_{s,t}), && (1.4.4)\\
\pi^\wedge(j, x_j > 0) &= \sum_{s\in\wedge} \pi^\wedge(s)\lambda_j \tau_s + \sum_{s\notin\wedge} \pi^\wedge(s)\lambda_j\tilde{\tau}_s \\
&\quad + \sum_{s\notin\wedge} \pi^\wedge(s, x_s > 0)\lambda_j(\tau_s - \tilde{\tau}_s), && (1.4.5)\\
\sum_{s\in\mathcal{S}} \pi^\wedge(s) &= 1. && (1.4.6)
\end{aligned}$$

Let us introduce the quantities

$$\begin{aligned}
\hat{\rho}^\wedge &\stackrel{\text{def}}{=} \sum_{s\notin\wedge} \lambda_s(\tau_s - \tilde{\tau}_s), \\
\bar{\tau}^\wedge &\stackrel{\text{def}}{=} \frac{1}{1-\hat{\rho}^\wedge}\Bigl[\sum_{s\in\wedge} \pi^\wedge(s)\tau_s + \sum_{s\notin\wedge} \pi^\wedge(s)\tilde{\tau}_s\Bigr].
\end{aligned}$$

Note that when $\wedge$ is ergodic, $\pi^\wedge(j, x_j > 0) = \lambda_j \bar{\tau}^\wedge$, for all $j \notin \wedge$, and $\bar{\tau}^\wedge$ can be interpreted as the mean time between two polling instants of the induced chain. Then the system defined by (1.4.4) and (1.4.5) can be replaced by the forthcoming $N+1$ equations: for all $t \in \mathcal{S}$,

$$\begin{aligned}
\pi^\wedge(t) &= \sum_{s\in\wedge} \pi^\wedge(s)p_{s,t} + \sum_{s\notin\wedge} \pi^\wedge(s)\tilde{p}_{s,t} + \bar{\tau}^\wedge \sum_{s\notin\wedge} \lambda_s(p_{s,t} - \tilde{p}_{s,t}), && (1.4.7)\\
\bar{\tau}^\wedge &= \frac{1}{1-\hat{\rho}^\wedge}\Bigl[\sum_{s\in\wedge} \pi^\wedge(s)\tau_s + \sum_{s\notin\wedge} \pi^\wedge(s)\tilde{\tau}_s\Bigr]. && (1.4.8)
\end{aligned}$$

When $\wedge = \emptyset$, the system (1.4.6)–(1.4.8) coincides with (1.3.5), (1.3.8) and (1.3.9). In addition, under (1.4.3), we have

$$\hat{\rho}^\wedge < 1, \text{ for all } \wedge\text{'s}. \tag{1.4.9}$$

An easy flow computation shows that the components of the drift vector $\vec{M}(s, \vec{x})$ can be expressed as

$$M_j(s, \vec{x}) = \lambda_j \tau_s \mathbb{1}_{\{x_s>0\}} + \lambda_j \tilde{\tau}_s \mathbb{1}_{\{x_s=0\}} - \mathbb{1}_{\{x_s>0, s=j\}}.$$

Then the computation of the second vector field becomes easy. For $j \in \wedge$,

$$v_j^\wedge = \sum_{\substack{x\in C^\wedge \\ s\in\mathcal{S}}} \pi^\wedge(s, \vec{x}) M_j(s, \vec{x})$$

$$
\begin{aligned}
&= \sum_{\substack{x\in C^\wedge\\ s\notin\wedge}} \pi^\wedge(s,\vec{x})M_j(s,\vec{x}) + \sum_{\substack{x\in C^\wedge\\ s\in\wedge}} \pi^\wedge(s,\vec{x})M_j(s,\vec{x})\\
&= \lambda_j \sum_{s\notin\wedge}\Big[\pi^\wedge(s,x_s>0)\tau_s + \pi^\wedge(s,x_s=0)\tilde{\tau}_s\Big]\\
&\quad + \lambda_j\sum_{s\in\wedge}\pi^\wedge(s)\tau_s - \pi^\wedge(j)\\
&= \lambda_j\bar{\tau}^\wedge - \pi^\wedge(j).
\end{aligned}
$$

In order to apply Theorem 1.A.1, it will be convenient to introduce

$$
\begin{aligned}
f_i(\vec{x}) &\stackrel{\text{def}}{=} \Big[{}^*\vec{x}\,[I-\mathcal{A}+\widetilde{\mathcal{A}}]^{-1}\Big]_i\\
&= x_i + \frac{\lambda_i\sum_{j=1}^N x_j(\tau_j-\tilde{\tau}_j)}{1-\hat{\rho}}. \qquad (1.4.10)
\end{aligned}
$$

This function is not as outlandish as it could seem at first sight. It is indeed directly related to flow conservation equations (see [16]). Let $\wedge$ be any ergodic face. For $i\in\wedge$, we get after a straightforward computation

$$
\begin{aligned}
f_i(\vec{v}^\wedge) &= v_i^\wedge + \frac{\lambda_i}{1-\hat{\rho}}\sum_{j=1}^N(\tau_j-\tilde{\tau}_j)v_j^\wedge\\
&= \frac{\lambda_i\sum_{j=1}^N\pi^\wedge(j)\tilde{\tau}_j}{1-\hat{\rho}} - \pi^\wedge(i).
\end{aligned}
$$

For $i\notin\wedge$, and from the very definition of $f_i(\vec{x})$, we have

$$
f_i(\vec{v}^\wedge) - \sum_{j=1}^N f_j(\vec{v}^\wedge)\lambda_i(\tau_j-\tilde{\tau}_j) = v_i^\wedge = 0,
$$

or

$$
f_i(\vec{v}^\wedge) = \frac{\lambda_i\sum_{j\in\wedge}f_j(\vec{v}^\wedge)(\tau_j-\tilde{\tau}_j)}{1-\hat{\rho}^\wedge}.
$$

Theorem 1.4.2 *Assume that (1.4.1)–(1.4.2) hold and that, for any ergodic face $\wedge$,*

$$
f_i(\vec{v}^\wedge) \equiv \frac{\lambda_i\sum_{j=1}^N\pi^\wedge(j)\tilde{\tau}_j}{1-\hat{\rho}} - \pi^\wedge(i) < 0, \quad \textit{for all } i\in\wedge. \qquad (1.4.11)
$$

Then the random walk $\mathcal{L}$ is ergodic. In particular, when $P=\widetilde{P}$, the conditions (1.4.1) and (1.4.2) are necessary and sufficient for the random walk $\mathcal{L}$ to be ergodic.

Proof We shall apply Theorem 1.A.1 (quoted in Appendix 1.A.1), which in itself contains a principle of gluing Lyapounov functions together. Most of the time, these functions are piecewise linear. Since the $f_i(\vec{v}^\wedge)$ enjoy "nice" properties, it is not unnatural to search for a linear combination like

$$f(\vec{x}) \stackrel{\text{def}}{=} \sum_{i=1}^{N} u_i f_i(\vec{x}),$$

where $\vec{u} = (u_1, \ldots, u_N)$ is a positive vector to be properly determined. Then, for any ergodic $\wedge$,

$$f(\vec{v}^\wedge) = \sum_{i\in\wedge} \Big[u_i + \frac{\sum_{j\notin\wedge}\lambda_j u_j}{1-\hat{\rho}^\wedge}(\tau_i - \tilde{\tau}_i)\Big] f_i(\vec{v}^\wedge).$$

The basic constraint for $\vec{u}$ is to ensure the positivity of f. Using (1.4.3) and (1.4.9), it appears that a suitable choice is $u_i = \max(\tilde{\tau}_i - \tau_i, \varepsilon)$, for some ε, positive and sufficiently small. Then one can directly check that $f(\vec{x}) > 0$, for any $\vec{x} \in \mathbb{R}_+^N$, and $f(\vec{v}^\wedge) < 0$, for any ergodic face $\wedge$. ■

Although the conditions of Theorem 1.4.2 are sufficient for ergodicity, they might well be implied by conditions (1.4.1)–(1.4.2) only. We did not check this fact, since it is difficult to compare algebraically the solutions of system (1.4.7)–(1.4.8), for different $\wedge$'s. Let us simply formulate the conjecture that (1.4.1)–(1.4.2) are also sufficient for ergodicity when $P \neq \widetilde{P}$.

Theorem 1.4.3 *Assume that $P = \widetilde{P}$ and that one of the following hold:*

(i) there exists $j \in \mathcal{S}$, such that $\lambda_j\bar{\tau} - F_j > 0$;

(ii) $\hat{\rho} > 1$.

Then the random walk $\mathcal{L}$ is transient.

Proof We assume that the queues are numbered according to the following order:

$$\frac{\lambda_1}{F_1} \leq \frac{\lambda_2}{F_2} \leq \ldots \leq \frac{\lambda_N}{F_N}. \tag{1.4.12}$$

Consider all faces $\wedge_i \stackrel{\text{def}}{=} \{i, \ldots, N\}$, for $i = 1, \ldots, N$. The idea is to show the transience of $\mathcal{L}$ by visiting the ordered set $\wedge_1, \wedge_2, \ldots \wedge_N$. In the usual terminology of dynamical systems, this amounts to finding a set of trajectories going to infinity with positive probability. To this end, the following algebraic relationship will be useful: for any $\wedge$ and $k \notin \wedge$, we have

$$(\lambda_k\bar{\tau}^{\wedge+\{k\}} - F_k)(1-\hat{\rho}^{\wedge+\{k\}}) = (\lambda_k\bar{\tau}^\wedge - F_k)(1-\hat{\rho}^\wedge). \tag{1.4.13}$$

By definition, the face $\wedge_1 \equiv \mathcal{S}$ is ergodic (see Appendix 1.A.1). Assume now $\wedge_i$ is ergodic, for some fixed $i \in \mathcal{S}$. If $v_i^{\wedge_i} < 0$, then one can prove the inequality

$$1 - \sum_{s=1}^{i} \lambda_s \tau_s > \frac{1}{\bar{\tau}^{\wedge_i}} \sum_{s=i+1}^{N} F_s \tau_s,$$

which in turn yields $1 - \hat{\rho}^{\wedge_{i+1}} > 0$. By using (1.4.13), (1.4.12) and Theorem 1.4.2, it follows that the face $\wedge_{i+1}$ is ergodic. Conversely, if $v_i^{\wedge_i} > 0$, then $\vec{v}^{\wedge_i} > \vec{0}$ and the random walk $\mathcal{L}$ is transient (see [15]).

Thus, by induction, we have shown that either $\mathcal{L}$ is transient or the face $\wedge_N \equiv \{N\}$ is ergodic. In the latter case, the assumptions of the theorem, together with (1.4.13) and (1.4.12), yield $v_N^{\wedge_N} \equiv \lambda_N \bar{\tau}^{\wedge_N} - F_N > 0$, so that $\mathcal{L}$ is transient. ■

1.4.3 Remarks and problems

(i) For general random walks, the second vector field requires to compute invariant measures of Markov chains in dimensions strictly smaller than N (the induced chains). This is rarely possible for $N > 3$, but there are some miracles comparable to what happens here, the most noticeable ones concerning Jackson networks and some polling systems [16].

(ii) No stochastic monotonicity argument seems to work, when the $\tau_k - \tilde{\tau}_k$'s are not all of the same sign, even for $P = \widetilde{P}$. This monotonicity exists and is crucial in [3], [19] and [20].

(iii) The analysis of the vector field $\vec{v}^{\wedge}$ is interesting in itself and will be the subject of a future work. Let us just quote one negative property: when $\wedge$ is ergodic and $\vec{v}_k^{\wedge} < 0$, then $\wedge_1 = \wedge \setminus \{k\}$ can be non-ergodic, contrary to what happens in Jackson networks (see [16]). In the case $P = \widetilde{P}$, we conjecture nonetheless that the dynamical system formed from the velocity vectors $\vec{v}^{\wedge}$ is strongly acyclic (in the sense that the same face is not visited twice).

1.5 The first moment of the queue length

This section shows that, when the system enjoys some symmetry properties, it becomes possible to derive the stationary mean queue length $\mathbb{E}[X_m \mid S = m]$ seen by the server at polling instants. To this end, ergodicity of $\mathcal{L}$ will be assumed, as well as the existence of second moments of all random variables of interest.

Assumption A1 *The system has a* rotational *symmetry*

(*i*) *for all* $i, j \in \mathcal{S}$, $p_{i,j} = p_{j-i}$, $\tilde{p}_{i,j} = \tilde{p}_{j-i}$, $a_{i,j}(\vec{z}) = a_{j-i}(\vec{z})$ *and* $\tilde{a}_{i,j}(\vec{z}) = \tilde{a}_{j-i}(\vec{z})$*;*

(*ii*) *for all* $1 \leq s, d \leq N$,

$$\begin{aligned} a_d(z_1, \ldots, z_N) &= a_d(z_{1+s}, \ldots, z_{N+s}), \\ \tilde{a}_d(z_1, \ldots, z_N) &= \tilde{a}_d(z_{1+s}, \ldots, z_{N+s}). \end{aligned}$$

Note that this assumption is less restrictive than the ones appearing in the literature, even for state-independent situations. It allows in particular cyclic and random polling strategies. From **A1**, we directly get, for all $1 \leq i, s \leq N$,

$$F_i(\vec{z}) = F_{i+s}(z_{1+s}, \ldots, z_{N+s}), \tag{1.5.1}$$
$$\widetilde{F}_i(\vec{z}) = \widetilde{F}_{i+s}(z_{1+s}, \ldots, z_{N+s}). \tag{1.5.2}$$

Consequently, $P(S = i) = F_i(\vec{e}) = 1/N$. In addition, $A(\vec{z})$ and $\widetilde{A}(\vec{z})$ are *circulant* matrices—see Appendix 1.A.2 for the results and notation which will be used from now on. It implies in particular that, for $k \in \mathcal{S}$, $\vec{v}_k$ is an eigenvector of $A(\vec{z})$ and $\widetilde{A}(\vec{z})$ with respective eigenvalues

$$\begin{aligned} \mu_k(\vec{z}) &\stackrel{\text{def}}{=} \sum_{d=1}^{N} p_d a_d(\vec{z}) \omega_k^d, \\ \tilde{\mu}_k(\vec{z}) &\stackrel{\text{def}}{=} \sum_{d=1}^{N} \tilde{p}_d \tilde{a}_d(\vec{z}) \omega_k^d. \end{aligned}$$

Note that $\mu_N(\vec{z})$ (resp. $\tilde{\mu}_N(\vec{z})$) is the generating function of the number of arrivals during the transportation of one client (resp. a move without client). Moreover, $\mu_k \stackrel{\text{def}}{=} \mu_k(\vec{e})$ and $\tilde{\mu}_k \stackrel{\text{def}}{=} \tilde{\mu}_k(\vec{e})$ are eigenvalues of P and $\widetilde{P}$. For *a priori* technical reasons we shall also need

Assumption A2 *For any* $k < N$, $\tilde{\mu}_k \neq 1$.

These conditions are not surprising in view of Proposition 1.3.2. Since the system is symmetrical, the mean number of customers arriving during a travel of the taxicab $\alpha_{i;q}$ and $\tilde{\alpha}_{i;q}$ as defined by (1.3.1)–(1.3.2) does not depend on i and q; we note them α and $\tilde{\alpha}$ and remark that

$$\begin{aligned} \alpha &= \frac{\partial \mu_N}{\partial z_m}(\vec{e}), \\ \tilde{\alpha} &= \frac{\partial \tilde{\mu}_N}{\partial z_m}(\vec{e}). \end{aligned}$$

The second derivatives of $\mu_N(\vec{z})$ and $\tilde{\mu}_N(\vec{z})$ are defined as

$$\alpha_{q,r}^{(2)} \stackrel{\text{def}}{=} \frac{\partial^2 \mu_N(\vec{z})}{\partial z_q \partial z_r}(\vec{e}), \tag{1.5.3}$$

$$\tilde{\alpha}_{q,r}^{(2)} \stackrel{\text{def}}{=} \frac{\partial^2 \tilde{\mu}_N(\vec{z})}{\partial z_q \partial z_r}(\vec{e}) \tag{1.5.4}$$

Note that these values depend only of the unsigned distance between q and r and that we can write $\alpha_{q,q+d}^{(2)} = \alpha_{q+d,q}^{(2)} \stackrel{\text{def}}{=} \alpha_d^{(2)}$, $\tilde{\alpha}_{q,q+d}^{(2)} = \tilde{\alpha}_{q+d,q}^{(2)} \stackrel{\text{def}}{=} \tilde{\alpha}_d^{(2)}$. It will be also convenient to introduce the first and second moments of the number of arrivals between two polling instants, that is

$$\bar{\alpha} \stackrel{\text{def}}{=} P(X_m > 0 \mid S = m)\alpha + \mathrm{P}(X_m = 0 \mid S = m)\tilde{\alpha}, \tag{1.5.5}$$

$$\bar{\alpha}_d^{(2)} \stackrel{\text{def}}{=} P(X_m > 0 \mid S = m)\alpha_d^{(2)} + \mathrm{P}(X_m = 0 \mid S = m)\tilde{\alpha}_d^{(2)}. \tag{1.5.6}$$

Using the symmetry of the system and according to the properties of generating functions, we have for any $m \in \mathcal{S}$

$$\mathbb{E}[X_m \mid S = m] = N \frac{\partial F_m}{\partial z_m}(\vec{e}). \tag{1.5.7}$$

By symmetry, this expression is independent from m. In order to derive $\mathbb{E}[X_m \mid S = m]$, we use the fact that the matrices $I - A\Delta(\vec{z})$ and $\widetilde{A}(\vec{z}) - A\Delta(\vec{z})$ are not invertible when $\vec{z} = \vec{e}$. This implies that the matrix $D(\vec{z})$ as defined in (1.2.2) is not continuous in a neighborhood of $\vec{z} = \vec{e}$ and provides a mean to compute the derivatives of $F(\vec{z})$ and $\widetilde{F}(\vec{z})$ for $\vec{z} = \vec{e}$. However in our setting, even the existence of $D(\vec{z})$ in a neighborhood of $\vec{e}$ is difficult to prove and the computations become really involved. The approach that we present here avoids the theoretical problems and simplifies the computations. In the next lemma, we show how it is possible to choose $\vec{z}$ such that the left member of Equation (1.2.1) can be rendered of "small" order w.r.t. the right one.

Lemma 1.5.1 *Let $t \to \vec{z}_t$ be a function from $\mathbb{R}_- \to \mathcal{D}^N$ such that in a neighborhood of $t = 0$, we can write $\vec{z}_t = \vec{e} + t\dot{\vec{z}} + \frac{1}{2}t^2\ddot{\vec{z}} + \vec{o}(t^2)$, where $\dot{\vec{z}} = \dot{z}_1\vec{e}_1 + \cdots + \dot{z}_N\vec{e}_N = \dot{\zeta}_1\vec{v}_1 + \cdots + \dot{\zeta}_N\vec{v}_N \in \mathbb{C}^N$ and $\ddot{\vec{z}} = (\ddot{z}_1, \ldots, \ddot{z}_N) \in \mathbb{C}^N$. Assume that $\dot{z}_1 + \cdots + \dot{z}_N = 0$. Then there exists for $t > 0$ a vector $\vec{u}_t$ such that, for $1 \le k < N$,*

$$^*\vec{u}_t\,[I - A\Delta(\vec{z}_t)]\vec{v}_k = o(t), \tag{1.5.8}$$

$$^*\vec{u}_t\,[\widetilde{A}(\vec{z}_t) - A\Delta(\vec{z}_t)]\vec{v}_k = t\frac{1-\tilde{\mu}_k}{1-\mu_k}\dot{\zeta}_{N-k} + o(t), \tag{1.5.9}$$

and

$$^*\vec{u}_t\,[I - A\Delta(\vec{z}_t)]\vec{e} = -t^2 \sum_{l=1}^{N-1} \dot{\zeta}_l \dot{\zeta}_{N-l} \left[\frac{1}{1-\mu_l} + \frac{N}{2}\sum_{d=1}^{N} \alpha_d^{(2)} \omega_l^d\right]$$

$$+\,t^2\frac{1-N\alpha}{2N}\sum_{q=1}^{N}\ddot{z}_q + o(t^2), \qquad (1.5.10)$$

$$\begin{aligned} {}^*\vec{u}_t\,[\widetilde{A}(\vec{z}_t) - A\Delta(\vec{z}_t)]\vec{e} \;&=\; -t^2\sum_{l=1}^{N-1}\dot{\zeta}_l\dot{\zeta}_{N-l}\Big[\frac{1}{1-\mu_l} + \frac{N}{2}\sum_{d=1}^{N}(\alpha_d^{(2)} - \tilde{\alpha}_d^{(2)})\omega_l^d\Big] \\ &\quad + t^2\frac{1-N\alpha+N\tilde{\alpha}}{2N}\sum_{q=1}^{N}\ddot{z}_q + o(t^2). \qquad (1.5.11)\end{aligned}$$

Proof See Appendix 1.A.3. ∎

For the sake of simplicity, the computations have been carried out in the most natural way which apparently requires $\mu_k \neq 1$ for $k < N$. In fact, the reader can convince himself that all derivations could be achieved by rendering the right member of (1.2.1) small w.r.t. the left member. This would yield the same result without any restriction on μ_k, at the expense of a somewhat longer proof.

One problem in polling models with 1-limited service strategy is that the rank of the systems of equations that we can write is $\lfloor n/2 \rfloor$, which means that $\partial F/\partial z_m(\vec{e})$ cannot be computed. Fortunately, under the following assumption, we are able to compute the most important value, that is $\partial F_m/\partial z_m(\vec{e})$.

Assumption A3 *The routing matrices are such that*

$$[I-P][I-{}^*\widetilde{P}\,] = [I-{}^*P\,][I-\widetilde{P}].$$

It is not easy to give an interpretation of **A3** nor to describe all models satisfying it. The simplest one is the classical Markov polling obtained when $P = \widetilde{P}$, for which the results of Theorem 1.5.2 thereafter are greatly simplified. Another admissible model is when the routing probabilities depend only on the absolute distance between stations, that is when $P = {}^*P$ and $\widetilde{P} = {}^*\widetilde{P}$.

Theorem 1.5.2 *Assume that* **A1**, **A2** *and* **A3** *hold. Then, for any station m, the stationary probability that the server finds station m empty is given by*

$$P(X_m = 0 \mid S = m) = \frac{1-N\alpha}{1-N\alpha+N\tilde{\alpha}}\,. \qquad (1.5.12)$$

Moreover the mean number of customers found by the server when it arrives at station m is obtained by means of the following formula

$$\mathbb{E}[X_m \mid S = m] \;=\; N\bar{\alpha} + \frac{N\tilde{\alpha}\bar{\alpha}}{1-N\alpha}\sum_{l=1}^{N-1}\frac{1}{1-\tilde{\mu}_l}$$

$$+\frac{1-N\alpha+N\tilde{\alpha}}{1-N\alpha}\frac{N}{2}\bar{\alpha}_N^{(2)}+\frac{N\alpha-N\tilde{\alpha}}{1-N\alpha}\frac{1}{2}\sum_{d=1}^{N}\bar{\alpha}_d^{(2)}$$
$$-\frac{N\tilde{\alpha}}{1-N\alpha}\sum_{l=1}^{N-1}\frac{\mu_l-\tilde{\mu}_l}{1-\tilde{\mu}_l}\frac{1}{2}\sum_{d=1}^{N}\bar{\alpha}_d^{(2)}\omega_l^d\,. \tag{1.5.13}$$

Finally, the mean number of customers present at arbitrary station m at a polling instant can be expressed as

$$\begin{aligned}\mathbb{E}X_m &= \bar{\alpha}+\frac{\tilde{\alpha}}{1-N\alpha}\sum_{l=1}^{N-1}\frac{1}{1-\tilde{\mu}_l}\\ &+\frac{1-N\alpha+N\tilde{\alpha}}{1-N\alpha}\frac{N}{2}\bar{\alpha}_N^{(2)}+\frac{N\alpha-N\tilde{\alpha}}{1-N\alpha}\frac{1}{2}\sum_{d=1}^{N}\bar{\alpha}_d^{(2)}\\ &-\frac{1-N\alpha+N\tilde{\alpha}}{1-N\alpha}\sum_{l=1}^{N-1}\frac{\mu_l-\tilde{\mu}_l}{1-\tilde{\mu}_l}\frac{1}{2}\sum_{d=1}^{N}\bar{\alpha}_d^{(2)}\omega_l^d\,.\end{aligned} \tag{1.5.14}$$

Proof See Appendix 1.A.3. ■

1.6 The mean waiting time of a customer

Theorem 1.5.2 is the key to compute of the waiting time of an arbitrary customer. In this section, we focus on the continuous-time case with compound Poisson arrivals and further assume that the arrivals at each station are independent. While Theorem 1.5.2 is valid in discrete time or with correlated arrivals, we do not give any formula in these cases, since the notation would become too involved.

The notation is as follows: customers arrive in i.i.d. batches of mean length b and second moment $b^{(2)}$ at the instants of a Poisson stream with intensity $\hat{\lambda}$ at each station. The intensity of the arrival process at each queue is $\lambda=\hat{\lambda}b$. The time between two polling instants when the first queue is not empty (resp. empty) has mean and second moment τ and $\tau^{(2)}$ (resp. $\tilde{\tau}$ and $\tilde{\tau}^{(2)}$). We must keep in mind that in general τ depends on P like $\alpha_{i;q}$ in Equation (1.3.1): if τ_d is the mean time between two consecutive polling instants at stations i and $i+d$, then we have $\tau=p_1\tau_1+\cdots+p_N\tau_N$. However, in most symmetrical polling models that have been previously analyzed, the switchover times are equal and τ does not depend on P. We will say in this case that the stations are *equidistant*, which is obviously not the case in our original taxicab problem. The same holds for $\tau^{(2)}$, $\tilde{\tau}$ and $\tilde{\tau}^{(2)}$.

Theorem 1.6.1 *Assume that* **A1**, **A2** *and* **A3** *hold. Then the mean stationary waiting time of a customer is given by the relation*

$$\mathbb{E}[W] = \frac{\tilde{\tau}}{1-N\lambda\tau}\sum_{l=1}^{N-1}\left[\frac{1}{1-\tilde{\mu}_l}\right]+\frac{N\lambda\tau^{(2)}}{2(1-N\lambda\tau)}+\frac{\tilde{\tau}^{(2)}}{2\tilde{\tau}}$$

$$+\frac{b^{(2)}-b}{2b}\left[\frac{\tau+(N-1)\tilde{\tau}}{1-N\lambda\tau}-\frac{\tilde{\tau}}{1-N\lambda\tau}\sum_{l=1}^{N-1}\frac{\mu_l-\tilde{\mu}_l}{1-\tilde{\mu}_l}\right]. \quad (1.6.1)$$

Proof See Appendix 1.A.3. ■

One surprising consequence of this formula is that when the arrival process is Poisson and the stations are *equidistant*, the mean waiting time does not depend on the routing matrix P. In other words, it does not depend on where the server goes after serving a customer. As for many phenomena arising in this model, it is difficult to give an intuitive interpretation of this property.

Beside giving the mean waiting time for a customer in our taxicab model, Equation (1.6.1) contains in fact many known formulas for waiting times corresponding to various service and polling strategies. In the next subsections, we give some of these applications.

A notation closer to the classical queuing theory notation is necessary to make the link with known results. Let w (resp. $\tilde{w}$) be the mean walking time—or switchover time—from a non-empty (resp. empty) station and let σ be the mean service time required by the customers. We denote by $w^{(2)}$, $\tilde{w}^{(2)}$ and $\sigma^{(2)}$ the associated second moments.

1.6.1 A state-independent Markovian polling model

The first possible application of our model is the classical state independent polling model with 1-limited service strategy. In this model, we have $P=\widetilde{P}$, $\tau=w+\sigma$, $\tilde{\tau}=w$ and (1.6.1) becomes

$$\begin{aligned}\mathbb{E}[W] &= \frac{w}{1-N\lambda(w+\sigma)}\sum_{l=1}^{N-1}\frac{1}{1-\mu_l}+\frac{N\lambda(w^{(2)}+2w\sigma+\sigma^{(2)})}{2(1-N\lambda(w+\sigma))}+\frac{w^{(2)}}{2w}\\ &\quad -\frac{b^{(2)}-b}{2b}\frac{Nw+\sigma}{1-N\lambda(w+\sigma)}.\end{aligned}$$

This formula is valid for all symmetric polling models with Markovian polling. It is interesting to note that, in the case of equidistant stations, there is only one term of the formula which depends on the routing. To compute this term, we remark that, if $\mathcal{P}(x)$ is the characteristic polynomial of P,

$$\sum_{l=1}^{N-1}\frac{1}{1-\mu_l} = \sum_{l=1}^{N-1}\frac{d}{dx}\log(x-\mu_l)\Big|_{x=1} = \frac{d}{dx}\log\Big[\frac{\mathcal{P}(x)}{x-1}\Big]\Big|_{x=1}.$$

The case of the cyclic polling is obtained by taking $p_1=1$ and $p_d=0$ for $d\neq 1$. Then the eigenvalues of P are $\omega_1,\ldots,\omega_N$ and $\mathcal{P}(x)=x^N-1$. So in this case, we have

$$\sum_{l=1}^{N-1} \frac{1}{1-\mu_l} = \frac{N-1}{2}.$$

Another classical polling strategy is the random polling with $p_d = 1/N$ for all d. In this case, we have $\mu_l = 0$ for $l < N$ and

$$\sum_{l=1}^{N-1} \frac{1}{1-\mu_l} = N-1.$$

The comparison between formulas for cyclic polling and random polling shows that the mean waiting time is smaller in the cyclic case. In fact this property can be generalized to all Markovian polling models.

Lemma 1.6.2 *Among all Markovian polling strategies for symmetric* 1*-limited polling models with* equidistant *queues, the cyclic polling strategy minimizes the waiting time of the customers.*

Proof This result is very easy to prove from Equation (1.6.1). We have

$$\begin{aligned}\sum_{l=1}^{N-1} \frac{1}{1-\mu_l} &= \sum_{l=1}^{N-1} \mathrm{Re}\Big(\frac{1}{1-\mu_l}\Big) \\ &= \sum_{l=1}^{N-1} \frac{1-\mathrm{Re}(\mu_l)}{2(1-\mathrm{Re}(\mu_l))+|\mu_l|^2-1},\end{aligned}$$

where $\mathrm{Re}(z)$ is the real part of z. Since $|\mu_l| \leq 1$ and $\mathrm{Re}(\mu_l) \leq 1$, we find that

$$\sum_{l=1}^{N-1} \frac{1}{1-\mu_l} \geq \frac{N-1}{2}.$$

The bound is attained if and only if for all l, $|\mu_l| = 1$. Since μ_l is the center of gravity of $\omega_l^1, \ldots, \omega_l^N$ with weights $p_1, \ldots, p_N$, this is only possible when for some $s \in \mathcal{S}$ we have $p_s = 1$ and $p_d = 0$ for $d \neq s$. When s is not a divider of N (to ensure that the routing matrix is irreducible), this is equivalent to cyclic polling. ■

This property also holds for discrete time systems and systems with correlated arrivals. The computation would be more difficult in the case where the stations are not equidistant.

1.6.2 The exhaustive and Bernoulli service strategies

An interesting application of our model is to show that some service strategies can be obtained by a proper choice of polling strategy. In this subsection, we show how we can apply our model to exhaustive and Bernoulli service strategies.

Consider the case where the stations are *equidistant*, except that the switchover time from a station to itself after a service is zero. Then, if we denote $p_N = 1 - \pi$, we have $\tau = \pi w + \sigma$, $\tau^{(2)} = \pi w^{(2)} + 2\pi w\sigma + \sigma^{(2)}$, $\tilde{\tau} = w$ and $\tilde{\tau}^{(2)} = w^{(2)}$. If we assume for the sake of simplicity that the arrival process is Poisson, (1.6.1) yields

$$\mathbb{E}[W] = \frac{w}{1 - N\lambda(\pi w + \sigma)} \sum_{l=1}^{N-1} \frac{1}{1 - \tilde{\mu}_l} + \frac{N\lambda(\pi w^{(2)} + 2\pi w\sigma + \sigma^{(2)})}{2(1 - N\lambda(\pi w + \sigma))} + \frac{w^{(2)}}{2w}.$$

The value of $\mathbb{E}[W]$ depends on P only through the value of p_N. One known model that is described by this equation is the Bernoulli strategy of Servi [52]: when the server has served a customer, it quits the queue with probability π and continues to serve it with probability $1 - \pi$. In the original model, the server polls the queues in cyclic order. Here, we have a Markovian Bernoulli polling model if we take $P = (1 - \pi)I + \pi\widetilde{P}$. It is easy to check that this choice of P satisfies **A3**. Moreover, $\mathbb{E}[W]$ is an increasing function of π that is minimal when $\pi = 0$. This case corresponds to an exhaustive service strategy with Markovian routing $\widetilde{P}$: the server polls the same queue until it is empty and then moves to another queue using the routing matrix $\widetilde{P}$. In this case, Equation (1.6.1) simply becomes

$$\mathbb{E}[W] = \frac{w}{1 - N\lambda\sigma} \sum_{l=1}^{N-1} \frac{1}{1 - \tilde{\mu}_l} + \frac{N\lambda\sigma^{(2)}}{2(1 - N\lambda\sigma)} + \frac{w^{(2)}}{2w}.$$

Using Lemma 1.6.2, we find that the above expression is minimal when $\widetilde{P}$ describes a cyclic polling scheme. Further comparisons between different polling strategies can be found in Levy, Sidi and Boxma [39].

1.A Appendix

1.A.1 Second vector field

This appendix contains some basic definitions and results from the theory of dynamical systems. These definitions have been adapted for the Markov chain $\mathcal{L}$ defined in Section 1.2. We refer the reader to [42] and [15] for a more complete treatment of the subject.

Faces. For any $\wedge \subset \{1, \ldots, N\}$, define $B^\wedge \subset \mathbb{R}_+^N$ as

$$B^\wedge \stackrel{\text{def}}{=} \{(x_1, \ldots, x_n) : x_i > 0 \Leftrightarrow i \in \wedge\}.$$

$B^\wedge$ is the *face* of $\mathbb{R}_+^N$ associated to $\wedge$. Whenever not ambiguous, the face $B^\wedge$ will be identified with $\wedge$.

Induced chains. For any $\wedge \neq \mathcal{S}$ we choose an arbitrary point $\vec{a} \in B^\wedge \cap \mathbb{Z}_+^N$ and draw a plane $C^\wedge$ of dimension $N - |\wedge|$, perpendicular to $B^\wedge$ and containing

$\vec{a}$. We define the *induced Markov chain* $\mathcal{L}^\wedge$, with state space $\mathcal{S} \times (C^\wedge \cap \mathbb{Z}_+^N)$ (by an obvious abuse in the notation, we shall write most of the time $\mathcal{S} \times C^\wedge$) and transition probabilities

$$_\wedge p_{(s,\vec{x})(t,\vec{y})} = p_{(s,\vec{x})(t,\vec{y})} + \sum_{\vec{z} \neq \vec{y}} p_{(s,\vec{x})(t,\vec{z})}, \ \forall \vec{x}, \vec{y} \in C^\wedge, \ s, t \in \mathcal{S},$$

where the summation is performed over all $\vec{z} \in \mathbb{Z}_+^N$, such that the straight line connecting $\vec{z}$ and $\vec{y}$ is perpendicular to $C^\wedge$. It is important to note that this construction does not depend on $\vec{a}$.

Assumption A4 *For any $\wedge$ the chain $\mathcal{L}^\wedge$ is irreducible and aperiodic. $\wedge$ is called* ergodic *(resp.* non-ergodic, transient*) according as $\mathcal{L}^\wedge$ is ergodic (resp. non-ergodic, transient).*

For an ergodic $\mathcal{L}^\wedge$, let $\pi^\wedge(s, \vec{x})$, $(s, \vec{x}) \in \mathcal{S} \times C^\wedge$ be its stationary transition probabilities. Moreover, let $\pi^\wedge(s)$ (resp. $\pi^\wedge(s, x_s > 0)$) be the stationary probability that the server is at station s (resp. and finds it nonempty).

First vector field. For any $(s, \vec{x}) \in \mathcal{S} \times \mathbb{Z}_+^N$, the *first vector field* is simply the mean drift of the random walk $\mathcal{L}$ at point $(s, \vec{x})$:

$$\vec{M}(s, \vec{x}) \stackrel{\text{def}}{=} \sum_{r \in \mathcal{S}, \, \vec{y} \in \mathbb{Z}_+^N} (\vec{y} - \vec{x}) P\Big[X(n+1) = \vec{y}, S(n+1) = r \,\Big|\, X(n) = \vec{x}, S(n) = s\Big].$$

Let $\vec{v}^\wedge \stackrel{\text{def}}{=} (v_1^\wedge, \ldots, v_N^\wedge)$ such that

$$\begin{aligned} v_i^\wedge &= 0, \ i \notin \wedge, \\ v_i^\wedge &= \sum_{(s,\vec{x}) \in \mathcal{S} \times C^\wedge} \pi^\wedge(s, \vec{x}) \vec{M}_i(s, \vec{x}), \ i \in \wedge. \end{aligned}$$

Intuitively, one can imagine that the random walk starts from a point which is close to $\wedge$, but sufficiently far from all other faces $B^{\wedge'}$, with $\wedge \not\subset \wedge'$. After some time (sufficiently long, but less than the minimal distance from the above mentioned $B^{\wedge'}$), the stationary regime in the induced chain will be installed. In this regime, one can ask about the mean drift along $\wedge$: it is defined exactly by $\vec{v}^\wedge$. For $\wedge = \mathcal{S}$, we call $\wedge$ *ergodic*, by definition, and put

$$\vec{v}^{\{1,\ldots,N\}} \equiv \sum_{s \in \mathcal{S}} \pi^{\{1,\ldots,N\}}(s) \vec{M}(s, \vec{x}), \ \vec{x} \in B^{\{1,\ldots,N\}}.$$

From now on, when speaking about the *components* of $\vec{v}^\wedge$, we mean the components $v_i^\wedge$ with $i \in \wedge$. To avoid technical problems related to zero drifts, we assume the following:

Assumption A5 $v_i^\wedge \neq 0$*, for each $i \in \wedge$.*

Ingoing, outgoing and neutral faces. Let us fix $\wedge, \wedge_1$, so that $\wedge \supset \wedge_1, \wedge \neq \wedge_1$, that is to say $\overline{B}^{\wedge} \supset B^{\wedge_1}$ ($\overline{B}^{\wedge}$ is the closure of $B^{\wedge}$). Let $B^{\wedge}$ be ergodic. Thus $\vec{v}^{\wedge}$ is well defined. There are three possibilities for the direction of $\vec{v}^{\wedge}$ w.r.t. $B^{\wedge_1}$. We say that $B^{\wedge}$ is an *ingoing* (resp. *outgoing*) *face* for $B^{\wedge_1}$, if all the coordinates $v_i^{\wedge}$ for $i \in \wedge \setminus \wedge_1$ are negative (resp. positive). Otherwise we say that $B^{\wedge}$ is *neutral*. As an example we give simple sufficient criteria for a face to be ergodic.

The second vector field. To any point $\vec{x} \in \mathbb{R}_+^N$, we assign a vector $v(\vec{x})$ and call this function the *second vector field*. It can be multivalued on some non-ergodic faces. We put, for ergodic faces $B^{\wedge}$,

$$\vec{v}(\vec{x}) \equiv \vec{v}^{\wedge},\ \vec{x} \in B^{\wedge}.$$

If $B^{\wedge_1}$ is non-ergodic, then at any point $\vec{x} \in B^{\wedge_1}$, $\vec{v}(\vec{x})$ takes all values $\vec{v}^{\wedge}$ for which $B^{\wedge}$ is an outgoing face with respect to $B^{\wedge_1}$. In other words, for $\vec{x}$ belonging to non-ergodic faces, with $\|\vec{x}\|$ sufficiently large,

$$\vec{x} + \vec{v}(\vec{x}) \ \in \mathbb{R}_+^N,$$

for any value $\vec{v}(\vec{x})$. If there is no such vector, we put $\vec{v}(\vec{x}) = 0$, for $\vec{x} \in B^{\wedge_1}$. Points $\vec{x} \in \mathbf{R}_+^N$, where $\vec{v}(\vec{x})$ is more than one-valued, are called branch points.

There are few interesting examples, for which only the first vector field suffices to obtain ergodicity conditions for the random walk of interest, but it is nevertheless the case for Jackson networks. In general, the second vector field must be introduced as shown in the following theorem, taken from [42] (and extended in [16] to the case of upward unbounded jumps).

Theorem 1.A.1 *Assume that there exists a nonnegative function $f : \mathbb{R}_+^N \to \mathbb{R}_+$ such that*

(*i*) *for some constant $c > 0$*

$$f(\vec{x}) - f(\vec{y}) \leq c\|\vec{x} - \vec{y}\|;$$

(*ii*) *there exists $\delta > 0$, $p > 0$ such that for all $x \in B^{\wedge}$*

$$f(\vec{x} + v(\vec{x})) - f(\vec{x}) \leq -\delta.$$

Then the Markov chain $\mathcal{L}$ is ergodic.

1.A.2 Some simple results about circulant matrices

In this appendix, we recall some well-known properties of the circulant matrices which are used throughout this chapter. These properties are given without proof, since they can easily be verified at hand. Throughout the chapter, we take the convention that every subscript less than 1 or greater than N should be shifted into the correct range.

Let $(\vec{e}_1, \ldots, \vec{e}_N)$ denote the canonical basis of $\mathbb{C}^N$ and let $\vec{e} = (1, \ldots, 1)$. Moreover, we define $\omega_k \stackrel{\text{def}}{=} \exp(2\mathrm{i}\pi k/N)$, with $\mathrm{i}^2 = -1$.

Definition *A circulant matrix M of size N is a matrix whose coefficients $m_{i,j}$ verify the relation:*

$$m_{i+k,j+k} = m_{i,j}, \text{ for all } 1 \le i, j, k \le N.$$

The following Lemma summarizes some key properties of these matrices:

Lemma 1.A.2 *Let M be a circulant matrix of the form:*

$$M = \begin{pmatrix} m_N & m_{N-1} & \cdots & m_1 \\ m_1 & m_N & \cdots & m_2 \\ \vdots & & \ddots & \vdots \\ m_{N-1} & m_{N-2} & \cdots & m_N \end{pmatrix}$$

Then for each $1 \le k \le N$, the vector $\vec{v}_k \stackrel{\text{def}}{=} \sum_{i=1}^{N} \omega_k^{-i} \vec{e}_i$ is an eigenvector of the matrix M with eigenvalue $\sum_{i=1}^{N} \omega_k^i m_i$. Moreover, $(\vec{v}_1, \ldots, \vec{v}_N)$ is an orthogonal basis of $\mathbb{C}^N$ in which $\vec{e}_1, \ldots, \vec{e}_N$ can be expressed as

$$\vec{e}_i = \frac{1}{N} \sum_{k=1}^{N} \omega_k^i \vec{v}_k.$$

One important feature of circulant matrices is that they share the same basis of eigenvectors. Note that, with the notation given before Lemma 1.A.2, we have $\vec{v}_N = \vec{e}$. Circulant matrices enjoy other properties that are not used here: for example, the product of two circulant matrices is a circulant matrix and this product commutes.

1.A.3 Proofs of the results of Sections 1.5 and 1.6

The derivations in this section are essentially analytic.

Proof of Lemma 1.5.1 This proof uses specific properties of circulant matrices to perform a fine analysis of the behavior of $A(\vec{z})$, $\widetilde{A}(\vec{z})$ and $\Delta(\vec{z})$ in the neighborhood of $\vec{z} = \vec{e}$. As pointed out later in the proof, we study these matrices only for $\vec{z} \in \mathcal{D}^N$, thus avoiding any analytical continuation. The basic relation used thereafter is a simple consequence of the definitions of $A(\vec{z})$ and $\Delta(\vec{z})$:

$$A\Delta(\vec{z})\vec{v}_k = \mu_k(\vec{z})\vec{v}_k + \sum_{q=1}^{N} (\frac{1}{z_q} - 1)\omega_k^{-q} A_{\cdot q}(\vec{z}), \tag{1.A.1}$$

where $A_{\cdot q}(\vec{z})$ stands for the q-th column of $A(\vec{z})$ and can be written as

$$\begin{aligned} A_{\cdot q}(\vec{z}) &= \sum_{i=1}^{N} p_{i-q} a_{i-q}(\vec{z}) \vec{e}_i \\ &= \sum_{i=1}^{N} p_{i-q} a_{i-q}(\vec{z}) \frac{1}{N} \sum_{l=1}^{N} \omega_l^i \vec{v}_l \\ &= \sum_{l=1}^{N} \omega_l^q \mu_l(\vec{z}) \vec{v}_l. \end{aligned}$$

So, if we define

$$\varepsilon_l(\vec{z}) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{q=1}^{N} (\frac{1}{z_q} - 1) \omega_l^q,$$

(1.A.1) can be rewritten as

$$[I - A\Delta(\vec{z})]\vec{v}_k = (1 - \mu_k(\vec{z}))\vec{v}_k - \sum_{l=1}^{N} \varepsilon_{l-k}(\vec{z}) \mu_l(\vec{z})\, \vec{v}_l \qquad (1.A.2)$$

$$[\widetilde{A}(\vec{z}) - A\Delta(\vec{z})]\vec{v}_k = (\tilde{\mu}_k(\vec{z}) - \mu_k(\vec{z}))\vec{v}_k - \sum_{l=1}^{N} \varepsilon_{l-k}(\vec{z}) \mu_l(\vec{z})\, \vec{v}_l \qquad (1.A.3)$$

Since ${}^*\vec{e}\,[I - A\Delta(\vec{e})] = {}^*\vec{e}\,[\widetilde{A}(\vec{e}) - A\Delta(\vec{e})] = \vec{0}$, we define, for any set of arbitrary complex numbers $(c_1, \ldots, c_{N-1})$, the vector $\vec{u}_t$ as follows:

$$\vec{u}_t \stackrel{\text{def}}{=} \frac{1}{N}\vec{e} + \frac{t}{N} \sum_{l=1}^{N-1} c_l \vec{v}_l.$$

With this definition, we have

$$\begin{aligned} {}^*\vec{u}_t\,[I - A\Delta(\vec{z})]\vec{e} &= \frac{{}^*\vec{e}}{N}[I - A\Delta(\vec{z})]\vec{e} + \frac{t}{N} \sum_{l=1}^{N-1} c_l {}^*\vec{v}_l\,[I - A\Delta(\vec{z})]\vec{e} \\ &= 1 - (1 + \varepsilon_0(\vec{z}))\mu_N(\vec{z}) - t \sum_{l=1}^{N-1} c_l \varepsilon_l(\vec{z}) \mu_l(\vec{z}). \qquad (1.A.4) \end{aligned}$$

When t is small, we have the relations:

$$\begin{aligned} \mu_N(\vec{z}_t) &= 1 + t\alpha \sum_{q=1}^{N} \dot{z}_q + \frac{t^2}{2} \sum_{q,r=1}^{N} \alpha_{q,r}^{(2)} \dot{z}_q \dot{z}_r \\ &\quad + \frac{t^2\alpha}{2} \sum_{q=1}^{N} \ddot{z}_q + o(t^2) \qquad (1.A.5) \\ \tilde{\mu}_N(\vec{z}_t) &= 1 + t\tilde{\alpha} \sum_{q=1}^{N} \dot{z}_q + \frac{t^2}{2} \sum_{q,r=1}^{N} \tilde{\alpha}_{q,r}^{(2)} \dot{z}_q \dot{z}_r \end{aligned}$$

$$+\frac{t^2\tilde{\alpha}}{2}\sum_{q=1}^{N}\ddot{z}_q + o(t^2) \tag{1.A.6}$$

$$\begin{aligned}\varepsilon_l(\vec{z}_t) &= -\frac{t}{N}\sum_{q=1}^{N}\dot{z}_q\omega_l^q + \frac{t^2}{N}\sum_{q=1}^{N}\dot{z}_q^2\omega_l^q \\ &\quad - \frac{t^2}{2N}\sum_{q=1}^{N}\ddot{z}_q\omega_l^q + o(t^2).\end{aligned} \tag{1.A.7}$$

The formulas given in Appendix 1.A.2 allow to relate $\dot{z}_1, \dots, \dot{z}_N$ to $\dot{\zeta}_1, \dots, \dot{\zeta}_N$:

$$\dot{\zeta}_k = \frac{1}{N}\sum_{q=1}^{N}\dot{z}_q\omega_k^q, \qquad \dot{z}_q = \sum_{l=1}^{N}\dot{\zeta}_q\omega_l^{-q}, \qquad \sum_{q=1}^{N}\dot{z}_q\dot{z}_{q+d} = N\sum_{l=1}^{N}\dot{\zeta}_l\dot{\zeta}_{N-l}\omega_l^d.$$

Using the remark after the definition of $\alpha_{q,r}^{(2)}$ and $\tilde{\alpha}_{q,r}^{(2)}$, we find that:

$$\sum_{q,r=1}^{N}\alpha_{q,r}^{(2)}\dot{z}_q\dot{z}_r = \sum_{d=1}^{N}\alpha_d^{(2)}\sum_{q=1}^{N}\dot{z}_q\dot{z}_{q+d} = N\sum_{l=1}^{N}\dot{\zeta}_l\dot{\zeta}_{N-l}\sum_{d=1}^{N}\alpha_d^{(2)}\omega_l^d.$$

Applying (1.A.5) and (1.A.7) to (1.A.4), we see that the first-order term in the expression (1.A.4) is $(\alpha - 1/N)t[\dot{z}_1 + \cdots + \dot{z}_N]$. So a necessary condition to have a formula like (1.5.10) is $\dot{z}_1 + \cdots + \dot{z}_N = 0$. Using this relation, we have

$$\begin{aligned}{}^*\vec{u}_t\,[I - A\Delta(\vec{z}_t)]\vec{e} &= t^2\sum_{l=1}^{N-1}c_l\dot{\zeta}_l\mu_l - t^2\sum_{l=1}^{N}\dot{\zeta}_l\dot{\zeta}_{N-l} \\ &\quad - t^2\frac{N}{2}\sum_{l=1}^{N}\dot{\zeta}_l\dot{\zeta}_{N-l}\sum_{d=1}^{N}\alpha_d^{(2)}\omega_l^d \\ &\quad + t^2\frac{1-N\alpha}{2N}\sum_{q=1}^{N}\ddot{z}_q + o(t^2).\end{aligned} \tag{1.A.8}$$

We have to check that it is possible to have $\vec{z}_t \in \mathcal{D}^N$ and $\dot{z}_1 + \cdots + \dot{z}_N = 0$. One easy way to satisfy these constraints is to ensure that all coordinates of $\vec{z}_t$ arrive to 1 tangentially to the unit circle as t goes to 0. This is the case when, for any q, $\dot{z}_q$ is an imaginary number and $\ddot{z}_q < 0$. Moreover, for $k < N$,

$$\begin{aligned}{}^*\vec{u}_t\,[I - A\Delta(\vec{z}_t)]\vec{v}_k &= -\varepsilon_{N-k}(\vec{z}_t)\mu_N(\vec{z}_t) + tc_k(1 - \mu_k(\vec{z}_t)) \\ &\quad - t\sum_{l=1}^{N-1}c_l\varepsilon_{l-k}(\vec{z}_t)\mu_l(\vec{z}_t) \\ &= t\dot{\zeta}_{N-k} + tc_k(1-\mu_k) + o(t).\end{aligned} \tag{1.A.9}$$

This shows that we get equations (1.5.8) and (1.5.10) from (1.A.8) and (1.A.9) if we take

$$c_k = -\frac{\dot{\zeta}_{N-k}}{1-\mu_k}.$$

With this choice of $c_1, \ldots, c_{N-1}$, we obtain equations (1.5.11) and (1.5.9) in exactly the same way. ■

Proof of Theorem 1.5.2 The idea of this proof is to apply the results of Lemma 1.5.1 to the equation

$$^*\vec{u}_t\,[I - A\Delta(\vec{z}_t)]F(\vec{z}_t) = {}^*\vec{u}_t\,[\widetilde{A}(\vec{z}_t) - A\Delta(\vec{z}_t)]\widetilde{F}(\vec{z}_t). \tag{1.A.10}$$

Define $\vec{z}_t$ as in Lemma 1.5.1. Then

$$F(\vec{z}_t) = F(\vec{e}) + t\sum_{r=1}^{N} \dot{z}_r \frac{\partial F}{\partial z_r}(\vec{e}) + \vec{o}(t).$$

Moreover, Equation (1.5.1) implies that $\partial F_i/\partial z_r(\vec{e}) = \partial F_{i-r+m}/\partial z_m(\vec{e})$ for any fixed $m \in \S$ and

$$\begin{aligned}
\frac{\partial F}{\partial z_r}(\vec{e}) &= \sum_{i=1}^{N} \frac{\partial F_i}{\partial z_r}(\vec{e})\vec{e}_i \\
&= \sum_{i=1}^{N} \Big[\frac{\partial F_{i-r+m}}{\partial z_m}(\vec{e})\frac{1}{N}\sum_{k=1}^{N} \omega_k^i \vec{v}_k\Big] \\
&= \sum_{k=1}^{N} \omega_k^{r-m}\Big[\frac{1}{N}\sum_{i=1}^{N} \frac{\partial F_{i-r+m}}{\partial z_m}(\vec{e})\omega_k^{i-r+m}\Big]\vec{v}_k \\
&= \sum_{k=1}^{N} \omega_k^{r-m}\frac{\partial \varphi_k}{\partial z_m}(\vec{e})\vec{v}_k,
\end{aligned}$$

where $(\varphi_1(\vec{z}), \ldots, \varphi_N(\vec{z}))$, defined as

$$\varphi_k(\vec{z}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} F_i(\vec{z})\omega_k^i,$$

are the coordinates of $F(\vec{z})$ in the basis $(\vec{v}_1, \ldots, \vec{v}_N)$. Finally,

$$F(\vec{z}_t) = F(\vec{e}) + tN\sum_{k=1}^{N} \dot{\zeta}_k \omega_k^{-m}\frac{\partial \varphi_k}{\partial z_m}(\vec{e})\vec{v}_k + \vec{o}(t). \tag{1.A.11}$$

With a similar definition for $\tilde{\varphi}(\vec{z})$,

$$\widetilde{F}(\vec{z}_t) = \widetilde{F}(\vec{e}) + tN\sum_{k=1}^{N} \dot{\zeta}_k \omega_k^{-m}\frac{\partial \tilde{\varphi}_k}{\partial z_m}(\vec{e})\vec{v}_k + \vec{o}(t). \tag{1.A.12}$$

We now apply Lemma 1.5.1 and Equations (1.A.11) and (1.A.12) to Equation (1.A.10) and use the fact that, by symmetry, $F(\vec{e}) = F_1(\vec{e})\vec{e}$ and $\widetilde{F}(\vec{e}) = \widetilde{F}_1(\vec{e})\vec{e}$

$$\begin{aligned}
t^2 F_1(\vec{e})\Big\{ -\sum_{l=1}^{N-1} \dot{\zeta}_l\dot{\zeta}_{N-l}\Big[\frac{1}{1-\mu_l} + \frac{N}{2}\sum_{d=1}^{N} \alpha_d^{(2)}\omega_l^d\Big] + \frac{1-N\alpha}{2N}\sum_{q=1}^{N} \ddot{z}_q\Big\} \\
= \quad t^2 \widetilde{F}_1(\vec{e})\Big\{ -\sum_{l=1}^{N-1} \dot{\zeta}_l\dot{\zeta}_{N-l}\Big[\frac{1}{1-\mu_l} + \frac{N}{2}\sum_{d=1}^{N} (\alpha_d^{(2)} - \tilde{\alpha}_d^{(2)})\omega_k^d\Big] \\
+ \frac{1-N\alpha+N\tilde{\alpha}}{2N}\sum_{q=1}^{N} \ddot{z}_q\Big\} \\
+ t^2 N \sum_{k=1}^{N-1} \dot{\zeta}_k\dot{\zeta}_{N-k}\omega_k^{-m}\frac{1-\tilde{\mu}_k}{1-\mu_k}\frac{\partial\widetilde{\varphi}_k}{\partial z_m}(\vec{e}) + o(t^2).
\end{aligned}$$

As $\vec{\ddot{z}}$ can be chosen freely, we can derive a first equality from this equation, namely

$$F_1(\vec{e})\frac{1-N\alpha}{2N} = \widetilde{F}_1(\vec{e})\frac{1-N\alpha+N\tilde{\alpha}}{2N}.$$

Since $P(X_m = 0 \mid S = m) = \widetilde{F}_1(\vec{e})/F_1(\vec{e})$, this yields Equation (1.5.12). Taking in account (1.5.6), we find

$$\begin{aligned}
\sum_{k=1}^{N-1} \dot{\zeta}_k\dot{\zeta}_{N-k}\omega_k^{-m}\frac{1-\tilde{\mu}_k}{1-\mu_k}N\frac{\partial\widetilde{\varphi}_k}{\partial z_m}(\vec{e}) \\
= \quad -\sum_{l=1}^{N-1} \dot{\zeta}_l\dot{\zeta}_{N-l}\Big[\frac{F_1(\vec{e}) - \widetilde{F}_1(\vec{e})}{1-\mu_l} + \frac{1}{2}\sum_{d=1}^{N} \bar{\alpha}_d^{(2)}\omega_l^d\Big]. \qquad (1.A.13)
\end{aligned}$$

This equation contains in fact a system of linear equations that can be built by choosing $\vec{\ddot{z}}$. The problem is that in general the rank of this system is only $\lfloor n/2 \rfloor$. However, since μ_k (resp. μ_{N-k}, $\tilde{\mu}_k$, $\tilde{\mu}_{N-k}$) is the eigenvalue of ${}^*P$ (resp. P, ${}^*\widetilde{P}$, $\widetilde{P}$) associated to the eigenvector $\vec{v}_k$, **A3** implies that, for $1 \leq k < N$,

$$\frac{1-\mu_k}{1-\tilde{\mu}_k} = \frac{1-\mu_{N-k}}{1-\tilde{\mu}_{N-k}} \in \mathbb{R}.$$

As noted in the proof of Lemma 1.5.1, $\dot{z}_1, \ldots, \dot{z}_N$ are imaginary numbers and $\dot{\zeta}_k\dot{\zeta}_{N-k}$ is a negative real number for any $k < N$. Hence we can choose in (1.A.13)

$$\dot{\zeta}_k\dot{\zeta}_{N-k} = -\min\Big[\frac{1-\mu_k}{1-\tilde{\mu}_k}, 0\Big].$$

The combination of the resulting equation with a similar equation containing only the terms where $(1-\mu_k)/(1-\tilde{\mu}_k) < 0$ yields

$$\sum_{k=1}^{N-1} \omega_k^{-m} N \frac{\partial \widetilde{\varphi}_k}{\partial z_m}(\vec{e}) = -\sum_{l=1}^{N-1} \frac{1-\mu_l}{1-\tilde{\mu}_l}\left[\frac{F_1(\vec{e})-\widetilde{F}_1(\vec{e})}{1-\mu_l} + \frac{1}{2}\sum_{d=1}^{N} \bar{\alpha}_d^{(2)} \omega_l^d\right]. \quad (1.A.14)$$

We know that, by definition, $\widetilde{F}_m(\vec{z})$ does not depend on z_m. Hence,

$$\frac{\partial \widetilde{F}_m}{\partial z_m}(\vec{e}) = \sum_{k=1}^{N} \omega_k^{-m} \frac{\partial \widetilde{\varphi}_k}{\partial z_m}(\vec{e}) = 0,$$

and, using Equation (1.A.14)

$$N \frac{\partial \tilde{\varphi}_N}{\partial z_m}(\vec{e}) = \sum_{l=1}^{N-1}\left[\frac{F_1(\vec{e})-\widetilde{F}_1(\vec{e})}{1-\tilde{\mu}_l} + \frac{1-\mu_l}{1-\tilde{\mu}_l}\frac{1}{2}\sum_{d=1}^{N} \bar{\alpha}_d^{(2)} \omega_l^d\right]. \quad (1.A.15)$$

The last step of the demonstration is to get $\partial F_m/\partial z_m(\vec{e})$ from these results. This can be done as in Lemma 1.5.1, but with $\vec{u}_t = \vec{e}/N$ and $\vec{z}_t$ chosen differently. Using Equations (1.A.2), (1.A.3) and (1.A.5)–(1.A.7) with the same notation as in Lemma 1.5.1, we have for $k < N$

$$\begin{aligned}
\frac{{}^*\vec{e}}{N}[I - A\Delta(\vec{z}_t)]\vec{e} &= 1-\mu_N(\vec{z}_t) - \varepsilon_0(\vec{z}_t)\mu_N(\vec{z}_t) \\
&= t(1-N\alpha)\dot{\zeta}_N - t^2\sum_{l=1}^{N}\dot{\zeta}_l\dot{\zeta}_{N-l} + t^2 N\alpha\dot{\zeta}_N^2 \\
&\quad - t^2\frac{N}{2}\sum_{l=1}^{N}\dot{\zeta}_l\dot{\zeta}_{N-l}\sum_{d=1}^{N}\alpha_d^{(2)}\omega_l^d + o(t^2) \\
\frac{{}^*\vec{e}}{N}[I - A\Delta(\vec{z}_t)]\vec{v}_k &= -\varepsilon_{N-k}(\vec{z}_t)\mu_N(\vec{z}_t) \;=\; t\dot{\zeta}_{N-k} + o(t) \\
\frac{{}^*\vec{e}}{N}[\widetilde{A}(\vec{z}_t) - A\Delta(\vec{z}_t)]\vec{e} &= \tilde{\mu}_N(\vec{z}_t) - \mu_N(\vec{z}_t) - \varepsilon_0(\vec{z}_t)\mu_N(\vec{z}_t) \\
&= t(1-N\alpha+N\tilde{\alpha})\dot{\zeta}_N - t^2\sum_{l=1}^{N}\dot{\zeta}_l\dot{\zeta}_{N-l} + t^2 N\alpha\dot{\zeta}_N^2 \\
&\quad + t^2\frac{N}{2}\sum_{l=1}^{N}\dot{\zeta}_l\dot{\zeta}_{N-l}\sum_{d=1}^{N}[\tilde{\alpha}_d^{(2)} - \alpha_d^{(2)}]\omega_l^d + o(t^2) \\
\frac{{}^*\vec{e}}{N}[\widetilde{A}(\vec{z}_t) - A\Delta(\vec{z}_t)]\vec{v}_k &= -\varepsilon_{N-k}(\vec{z}_t)\mu_N(\vec{z}_t) \;=\; t\dot{\zeta}_{N-k} + o(t)
\end{aligned}$$

Combining these equations with (1.A.11) and (1.A.12), we find

$$-t^2 F_1(\vec{e})\left\{\sum_{l=1}^{N}\dot{\zeta}_l\dot{\zeta}_{N-l}\left[1+\frac{N}{2}\sum_{d=1}^{N}\alpha_d^{(2)}\omega_l^d\right] - N\alpha\dot{\zeta}_N^2\right\}$$

$$
\begin{aligned}
&+ t^2 N \sum_{k=1}^{N-1} \dot{\zeta}_k \dot{\zeta}_{N-k} \omega_k^{-m} \frac{\partial \varphi_k}{\partial z_m}(\vec{e}) + t^2 N(1-N\alpha)\dot{\zeta}_N^2 \frac{\partial \varphi_N}{\partial z_m}(\vec{e}) \\
&\quad = \quad -t^2 \widetilde{F}_1(\vec{e}) \left\{ \sum_{l=1}^{N} \dot{\zeta}_l \dot{\zeta}_{N-l} \left[1 + \frac{N}{2} \sum_{d=1}^{N} [\alpha_d^{(2)} - \tilde{\alpha}_d^{(2)}] \omega_l^d \right] - N\alpha \dot{\zeta}_N^2 \right\} \\
&\qquad + t^2 N \sum_{k=1}^{N-1} \dot{\zeta}_k \dot{\zeta}_{N-k} \omega_k^{-m} \frac{\partial \tilde{\varphi}_k}{\partial z_m}(\vec{e}) \\
&\qquad + t^2 N (1 - N\alpha + N\tilde{\alpha}) \dot{\zeta}_N^2 \frac{\partial \tilde{\varphi}_N}{\partial z_m}(\vec{e}) + o(t^2). \qquad (1.A.16)
\end{aligned}
$$

This equation in turn gives for $(1-N\alpha)\dot{\zeta}_N^2 = 1$ and $\dot{\zeta}_l \dot{\zeta}_{N-l} = 1$ if $1 \leq l < N$

$$
\begin{aligned}
N \sum_{k=1}^{N} \omega_k^{-m} \frac{\partial \varphi_k}{\partial z_m}(\vec{e}) \quad = \quad & \frac{N\tilde{\alpha}}{1-N\alpha} N \frac{\partial \tilde{\varphi}_N}{\partial z_m}(\vec{e}) + N(F_1(\vec{e}) - \widetilde{F}_1(\vec{e})) \\
& + \frac{1}{2} \sum_{l=1}^{N-1} \sum_{d=1}^{N} \bar{\alpha}_d^{(2)} \omega_l^d + \frac{1}{1-N\alpha} \frac{1}{2} \sum_{d=1}^{N} \bar{\alpha}_d^{(2)}
\end{aligned}
$$

This expression, together with (1.A.15) and (1.5.7), yields (1.5.13). The derivation of (1.5.14) is done in a similar way: we use the fact that

$$
\mathbb{E} X_m \;=\; \sum_{i=1}^{N} \frac{\partial F_i}{\partial z_m}(\vec{e}) \;=\; N \frac{\partial \varphi_N}{\partial z_m}(\vec{e}).
$$

We use Equation (1.A.16) with $(1-N\alpha)\dot{\zeta}_N^2 = 1$ and $\dot{\zeta}_k \dot{\zeta}_{N-k} = 0$ for $k < N$ and find

$$
N \frac{\partial \varphi_N}{\partial z_m}(\vec{e}) = \frac{1 - N\alpha + N\tilde{\alpha}}{1-N\alpha} N \frac{\partial \tilde{\varphi}_N}{\partial z_m}(\vec{e}) + F_1(\vec{e}) - \widetilde{F}_1(\vec{e}) + \frac{1}{1-N\alpha} \frac{1}{2} \sum_{d=1}^{N} \bar{\alpha}_d^{(2)}.
$$

This gives Equation (1.5.14) and concludes the proof of the theorem. ■

Proof of Theorem 1.6.1 We see easily that the mean numbers of clients arriving between polling times are respectively $\alpha = \lambda\tau$ and $\tilde{\alpha} = \lambda\tilde{\tau}$. Moreover, using Equations (1.5.3) and (1.5.4) and the properties of generating functions we find that, for $d < N$

$$
\begin{aligned}
\alpha_d^{(2)} &= \alpha^{(2)} = \lambda^2 \tau^{(2)} \\
\alpha_N^{(2)} &= \alpha^{(2)} + \hat{\lambda}(b^{(2)} - b)\tau \\
\tilde{\alpha}_d^{(2)} &= \tilde{\alpha}^{(2)} = \lambda^2 \tilde{\tau}^{(2)} \\
\tilde{\alpha}_N^{(2)} &= \tilde{\alpha}^{(2)} + \hat{\lambda}(b^{(2)} - b)\tilde{\tau}.
\end{aligned}
$$

The computation of waiting times uses the following classical argument: a non empty queue visited by the server can be decomposed into

- the head of line customer;

- the clients who arrived after him during his waiting time;
- clients who arrived in the same batch as the first client, but are not yet served; since, by renewal arguments, the mean size of this batch is $b^{(2)}/b$, the mean number of remaining clients is $(b^{(2)}-b)/2b$.

This can be written as

$$\mathbb{E}[X_m \mid S=m, X_m>0] = 1+\lambda\mathbb{E}[W]+\frac{b^{(2)}-b}{2b}$$

and, using the fact that $\sum_{d=1}^{N}\alpha_d^{(2)}\omega_l^d = \alpha_N^{(2)}-\alpha^{(2)}$,

$$\begin{aligned}
\lambda b\mathbb{E}[W] &= \frac{\mathbb{E}[X_m \mid S=m]}{1-P(X_m=0\mid S=m)}-1-\frac{b^{(2)}-b}{2b}\\
&= \frac{\tilde{\alpha}}{1-N\alpha}\sum_{l=1}^{N-1}\frac{1}{1-\mu_l}+\frac{1}{1-N\alpha}\frac{N}{2}\alpha^{(2)}+\frac{1}{N\tilde{\alpha}}\frac{N}{2}\tilde{\alpha}^{(2)}\\
&\quad+\frac{b^{(2)}-b}{2b}\left[\frac{\alpha+(N-1)\tilde{\alpha}}{1-N\alpha}-\frac{N\tilde{\alpha}}{1-N\alpha}\frac{1}{N}\sum_{l=1}^{N-1}\frac{\mu_l-\tilde{\mu}_l}{1-\tilde{\mu}_l}\right]\\
&= \frac{\lambda\tilde{\tau}}{1-N\lambda\tau}\sum_{l=1}^{N-1}\frac{1}{1-\mu_l}+\frac{N\lambda^2\tau^{(2)}}{2(1-N\lambda\tau)}+\frac{\lambda\tilde{\tau}^{(2)}}{2\tilde{\tau}}\\
&\quad+\frac{b^{(2)}-b}{2b}\left[\frac{\lambda\tau+(N-1)\lambda\tilde{\tau}}{1-N\lambda\tau}-\frac{\lambda\tilde{\tau}}{1-N\lambda\tau}\sum_{l=1}^{N-1}\frac{\mu_l-\tilde{\mu}_l}{1-\tilde{\mu}_l}\right].
\end{aligned}$$

This gives Equation (1.6.1). ■

Chapitre 2

Lois limites pour des grands réseaux fermés

Les files d'attente considérées dans le système à polling du Chapitre 1 étaient formées par les clients d'un réseau de transport. Le modèle décrit ici est en un certain sens dual du précédent, puisque l'on étudie l'attente des véhicules dans les stations. Le contexte est le suivant. Une zone géographique contient m véhicules répartis dans n stations. Un client arrivant à l'une d'elle prend– s'il en existe– un véhicule pour se rendre à une station de son choix ; sinon il quitte immédiatement le système.

Ce modèle peut être décrit comme un réseau fermé comportant n files à un serveur, dont le débit de service est égal au taux d'arrivée des clients, et n^2 files à infinité de serveurs représentant les déplacements inter-stations. Cette représentation conduit à des réseaux de type BCMP, pour lesquels la distribution stationnaire possède la fameuse « forme produit ». Toutefois, cette forme explicite, sans précaution préalable, conduit à une explosion combinatoire, notamment à cause d'une certaine constante de normalisation. Même si ces vingt dernières années, de nombreux algorithmes avec des coûts en $O(mn^2)$ ont été proposés (voir par exemple [37]), ils n'aident pas *a priori* à comprendre l'évolution qualitative et quantitative du réseau quand m et n croissent.

La nécessité d'analyser des grands systèmes s'est imposée de façon naturelle en thermodynamique et, plus généralement, en physique statistique. On observe alors des phénomènes de transition de phase, de métastabilité, etc. Il n'est *a priori* pas surprenant que des réseaux de files d'attente (qui sont en fait des marches aléatoires générales dans $\mathbb{Z}_+^n$) puissent conduire à des phénomènes de même nature.

De telles études ont déjà été menées pour des réseaux assez simples dans [33, 34, 2, 41], mais avec des outils d'analyse complexe qui ne permettent pas de traiter simplement les exemples les plus généraux. L'approche décrite ici s'inspire du théorème de la limite centrale (TLC), ou plus exactement du théorème de Berry-Esseen, puisque l'on donne un estimé de l'erreur commise. Elle permet à la fois de supprimer bon nombre d'hypothèses « techniques » et d'exprimer les asymptotiques en termes de moments d'ordre deux et plus. En particulier, on ne suppose aucune dépendance *a priori* entre m et n, contrairement aux études précédentes qui se cantonnent au modèle où m est proportionnel à n.

L'ensemble des lois limites obtenues par le TLC coïncidant en fait avec l'ensemble des lois infiniment divisibles, il est très difficile de traiter tous les comportements possibles du réseau. Nous traitons donc deux cas couvrant une partie importante des situations rencontrées en pratique. Ils sont liés à la condition de « négligeabilité asymptotique uniforme » rencontrée dans le TLC. D'autres comportements limites plus pathologiques pourraient bien sûr être traités de la même manière. Les résultats obtenus sont schématiquement les suivants :

- le réseau est asymptotiquement équivalent (au sens de la convergence faible) à un ensemble de files d'attente indépendantes, dont la longueur totale est égale en moyenne à m ;
- on est capable de donner des développements asymptotiques pour les marginales de la distribution stationnaire et la taille moyenne des files ;
- il existe une dépendance « intéressante » du type $m = f(n)$ entre m et n pour laquelle la file la plus chargée n'est ni vide ni saturée ; f n'est linéaire que si les files ont un comportement semblable ;
- si m est « trop grand », certaines files jouant un rôle de « goulot d'étranglement » se saturent, alors que la taille des autres est bornée.

Le premier résultat peut être interprété comme un résultat d'« invariance » : l'équivalence mentionnée a lieu sous des conditions d'existence de moments d'ordre petit. Les développements obtenus dans les Théorèmes 2.3.2 et 2.4.2 sont valables dès que m et n tendent conjointement vers l'infini. La dépendance particulière mise en évidence dans les Théorèmes 2.5.1 et 2.5.3 a surtout un intérêt pour le dimensionnement du réseau, puisqu'elle permet de le charger au maximum de sa capacité. Si le nombre de clients dans le système augmente, on aura alors des phénomènes de type « transition de phase ». En fin de chapitre, on revient sur le modèle de réseau de transport initial.

Ces travaux ont été réalisés avec Guy Fayolle et ont été publiés comme rapport de recherche [13]. Il est raisonnable de conjecturer que certains réseaux non à forme produit (par exemple lorsque les files ont des capacités limitées) ont, asymptotiquement, leurs « composantes » indépendantes. Nous y travaillerons.

2.1 Introduction

In many applications (telecommunications, transportation, etc.), it is desirable to understand the behaviour and performance of stochastic networks as their size increases. From an engineering point of view, the problem can be roughly formulated as follows.

> *Consider a closed network with n nodes and exactly m customers circulating inside. Find all functions f, such that $m = f(n)$ yields an* interesting *performance of the system as n increases.*

In this chapter, we start from the famous so-called product-form networks, which play an important role in quantitative analysis of systems. Although the equilibrium state probabilities have then a simple expression (see for example Kelly [30]), non-trivial problems remain, due to an intrinsic combinatorial explosion in the formulas, especially in those involving the famous normalizing constant. To circumvent these drawbacks, the idea is to compute asymptotic expansions of the characteristic values of the network, when m and n both tend to infinity.

This approach has been considered by Knessl and Tier [32], Kogan and Birman [33, 34, 2] and Malyshev and Yakovlev [41]. However, they rely on purely analytical methods which are difficult to use in a general setting and do not really give a structural explanation of the phenomena involved.

The method proposed in this chapter has direct connections with the *central limit theorem*: for instance, instead of considering the normalizing constant as an integral in the complex plane, we express it as the distribution of a scaled sum of independent random variables. Besides giving a clear interpretation of the computations, this allows to handle directly the general case of single-chain closed networks. More interestingly, the results can be interpreted as an insensitivity result: as the number of stations n and the number of customers $m(n)$ go to infinity, the network is equivalent to n independent queues, having the same characteristics and mean total number of customers $m(n)$. It turns out that the natural scaling parameter is indeed the flow λ_n in this equivalent network. As a consequence, we can state that $m(n)$ has *a priori* no theoretical reason of being proportional to n.

It is worth to emphasize that the problem of computing quantities (such as the above mentioned normalizing constant) appears frequently in statistical physics, in particular for Gibbs distributions in Markov random fields (see for example Whittle [58]). For the sake of completeness, it is of interest to quote the studies by Mitra, McKenna and Ramakrishnan [46, 44, 45, 47], where only the number of customers goes to infinity.

The present chapter is organized as follows. The model is introduced in Section 2.2 and the main values of interest are expressed in terms of integrals of characteristic functions. In Section 2.3, asymptotics of these integrals are found in the case where the queues are in *normal* or *moderated traffic*, while Section 2.4 deals with the case where some queues are overloaded. These results are summarized in Section 2.5, and the choice of scalings is discussed. Section 2.6 is devoted to an application of these results to a service vehicle network (like the PRAXITÈLE project, now developed at INRIA); non quite intuitive results are enlightened. Section 2.7 contains some conclusive remarks and some bounds used in Sections 2.3 and 2.4 are proven in Appendix 2.A.

2.2 Description of the model

We consider a closed BCMP network $\mathcal{B}_n$ with n queues and $m(n)$ clients. The number of clients at queue k at steady state is a random variable $Q_{k,n}$. The service rate at queue k when there are q_k customers is $\mu_{k,n}(q_k)$. The routing probability from queue k to queue ℓ is $p_{k,\ell,n}$ and P_n denotes the transition matrix supposed to be ergodic, with invariant measure $\pi_n = (\pi_{1,n}, \ldots, \pi_{n,n})$, defined by:

$$\pi_n P_n = \pi_n \text{ and } \pi_{1,n} + \cdots + \pi_{n,n} = 1. \tag{2.2.1}$$

Then it is known that, for any $q_1, \ldots, q_n \geq 0$ such that $q_1 + \cdots + q_n = m(n)$,

$$\mathrm{P}_n(Q_{1,n} = q_1, \ldots, Q_{n,n} = q_n) = Z^{-1}_{m(n),n} \prod_{k=1}^{n} \frac{\pi_{k,n}^{q_k}}{\mu_{k,n}(1) \cdots \mu_{k,n}(q_k)}, \tag{2.2.2}$$

with the normalizing condition

$$Z_{m,n} = \sum_{q_1+\cdots+q_n=m} \prod_{k=1}^{n} \frac{\pi_{k,n}^{q_k}}{\mu_{k,n}(1) \cdots \mu_{k,n}(q_k)}.$$

It is worth noting that our analysis applies to any network which has a product form equilibrium distribution like (2.2.2). For example, it includes, with a proper choice of the $\mu_{k,n}$'s, all systems having transition rates of the form $p_{k,\ell,n}\alpha_{k,n}(q_k)\beta_{\ell,n}(q_\ell)$. See Serfozo [51] for further examples.

For the sake of brevity, we shall write Z_n instead of $Z_{m(n),n}$, whenever non ambiguous. To each queue k, $1 \leq k \leq n$, we associate the generating function

$$f_{k,n}(z) \stackrel{\text{def}}{=} \sum_{q=1}^{\infty} \frac{z^q}{\mu_{k,n}(1) \cdots \mu_{k,n}(q)},$$

and we define

$$F_n(z) \stackrel{\text{def}}{=} \frac{1}{z^{m(n)}} \prod_{k=1}^{n} f_{k,n}(\pi_{k,n} z).$$

To the original closed network $\mathcal{B}_n$, we let correspond a new system $\mathcal{A}_n$ which is open and consists of n parallel queues, with service rates $\mu_{k,n}(x)$ and arrival intensity $\lambda_n\pi_{k,n}$ at queue k, where the choice of λ_n will be made more precise later. The queue length $X_{k,n}$ of the k's queue of $\mathcal{A}_n$ has a distribution given by

$$P(X_{k,n}=x)=\frac{(\lambda_n\pi_{k,n})^x}{f_{k,n}(\lambda_n\pi_{k,n})\mu_{k,n}(1)\cdots\mu_{k,n}(x)},$$

and $X_{1,n},\ldots,X_{n,n}$ are independent variables. We assume that $X_{k,n}$ has some finite moments of order $r\geq 2$ and introduce the following notation:

$$\begin{aligned}
m_{k,n} &\stackrel{\text{def}}{=} \mathbb{E}X_{k,n},\\
\beta_{k,n}^{(r)} &\stackrel{\text{def}}{=} \mathbb{E}|X_{k,n}-m_{k,n}|^r,\\
\sigma_{k,n}^2 &\stackrel{\text{def}}{=} \operatorname{Var} X_{k,n} \;=\; \beta_{k,n}^{(2)},\\
S_n &\stackrel{\text{def}}{=} X_{1,n}+\cdots+X_{n,n},\\
\beta_n^{(r)} &\stackrel{\text{def}}{=} \beta_{1,n}^{(r)}+\cdots+\beta_{n,n}^{(r)},\\
\sigma_n^2 &\stackrel{\text{def}}{=} \operatorname{Var} S_n \;=\; \beta_n^{(2)}.
\end{aligned}$$

Let $\varphi_{k,n}(\theta)$ be the characteristic function of $X_{k,n}-m_{k,n}$. Then, for any real θ,

$$\varphi_{k,n}(\theta)\stackrel{\text{def}}{=}\mathbb{E}e^{\mathrm{i}(X_{k,n}-m_{k,n})\theta}=\frac{f_{k,n}(\pi_{k,n}\lambda_n e^{\mathrm{i}\theta})}{f_{k,n}(\pi_{k,n}\lambda_n)}e^{-\mathrm{i}m_{k,n}\theta}, \tag{2.2.3}$$

and

$$\varphi_n(\theta)\stackrel{\text{def}}{=}\mathbb{E}e^{\mathrm{i}(S_n-\mathbb{E}S_n)\theta}=\varphi_{1,n}(\theta)\cdots\varphi_{n,n}(\theta) \tag{2.2.4}$$

The reason why $\mathcal{A}_n$ has been introduced is that the main performance characteristics of the network $\mathcal{B}_n$ can be expressed in terms of the distribution of $X_{1,n},\ldots,X_{n,n}$, for a properly chosen value of λ_n:

Lemma 2.2.1 *For any choice of $m(n)$, there exists an unique λ_n such that*

$$\mathbb{E}S_n=\mathbb{E}[X_{1,n}+\cdots+X_{n,n}]=m(n). \tag{2.2.5}$$

With this choice of λ_n, we have

$$Z_n=\frac{F_n(\lambda_n)}{2\pi}\int_{-\pi}^{\pi}\prod_{k=1}^{n}\varphi_{k,n}(\theta)d\theta. \tag{2.2.6}$$

The mean number of clients in queue j of $\mathcal{B}_n$ is given by

$$\mathbb{E}Q_{j,n}=m_{j,n}-\frac{F_n(\lambda_n)}{2\pi Z_n}\int_{-\pi}^{\pi}\mathrm{i}\varphi_{j,n}'(\theta)\prod_{k\neq j}^{n}\varphi_{k,n}(\theta)d\theta. \tag{2.2.7}$$

For any $\ell > 0$ and $q_1, \dots, q_\ell \geq 0$, the joint distribution of the length of queues $1, \dots, \ell$ is

$$\mathrm{P}(Q_{1,n} = q_1, \dots, Q_{\ell,n} = q_\ell) = \prod_{j=1}^{\ell} \mathrm{P}(X_{j,n} = q_j) \frac{F_n(\lambda_n)}{2\pi Z_n} \int_{-\pi}^{\pi} e^{\mathrm{i} \sum_{j=1}^{\ell} (q_j - m_{j,n})\theta} \prod_{k=\ell+1}^{n} \varphi_{k,n}(\theta) d\theta. \tag{2.2.8}$$

Proof One gets immediately from (2.2.3)

$$\begin{aligned} m_{k,n} &= \pi_{k,n}\lambda_n \frac{f'_{k,n}(\pi_{k,n}\lambda_n)}{f_{k,n}(\pi_{k,n}\lambda_n)}, \\ \sigma^2_{k,n} &= \pi_{k,n}\lambda_n \frac{f'_{k,n}(\pi_{k,n}\lambda_n)}{f_{k,n}(\pi_{k,n}\lambda_n)} \\ &\quad + \pi^2_{k,n}\lambda^2_n \frac{f''_{k,n}(\pi_{k,n}\lambda_n) f_{k,n}(\pi_{k,n}\lambda_n) - f'^2_{k,n}(\pi_{k,n}\lambda_n)}{f^2_{k,n}(\pi_{k,n}\lambda_n)}, \end{aligned}$$

so that

$$\frac{\partial}{\partial \lambda_n} \mathbb{E} S_n = \frac{\sigma^2_{1,n} + \cdots + \sigma^2_{n,n}}{\lambda_n} > 0.$$

The mean number of clients in $\mathcal{A}_n$ is thus a strictly increasing function of λ_n, which equals zero when $\lambda_n = 0$ and goes to infinity with λ_n. This proves the first assertion of the lemma.

For any fixed n we consider the generating function

$$\begin{aligned} \sum_{m=0}^{\infty} z^m Z_{m,n} &= \sum_{m=0}^{\infty} \sum_{q_1+\cdots+q_n=m} \prod_{k=1}^{n} \frac{(\pi_{k,n} z)^{q_k}}{\mu_{k,n}(1) \cdots \mu_{k,n}(q_k)} \\ &= \sum_{q_1,\dots,q_n} \prod_{k=1}^{n} \frac{(\pi_{k,n} z)^{q_k}}{\mu_{k,n}(1) \cdots \mu_{k,n}(q_k)} \\ &= \prod_{k=1}^{n} f_{k,n}(\pi_{k,n} z). \end{aligned}$$

Since $m(n)$ is finite, λ_n is in the domain of convergence of this function, and Cauchy formula on the circle $\mathcal{C}$ with center 0 and radius λ_n yields

$$\begin{aligned} Z_n &= \frac{1}{2\mathrm{i}\pi} \int_{\mathcal{C}} \frac{\prod_{k=1}^{n} f_{k,n}(\pi_{k,n} z)}{z^{m(n)+1}} dz \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\prod_{k=1}^{n} f_{k,n}(\pi_{k,n}\lambda_n e^{\mathrm{i}\theta})}{\lambda_n^{m(n)}} e^{-\mathrm{i}m(n)\theta} d\theta \qquad (2.2.9) \\ &= \frac{\prod_{k=1}^{n} f_{k,n}(\pi_{k,n}\lambda_n)}{\lambda_n^{m(n)}} \frac{1}{2\pi} \int_{-\pi}^{\pi} \prod_{k=1}^{n} \frac{f_{k,n}(\pi_{k,n}\lambda_n e^{\mathrm{i}\theta})}{f_{k,n}(\pi_{k,n}\lambda_n)} e^{-\mathrm{i}m(n)\theta} d\theta, \end{aligned}$$

which, taking (2.2.3) into account, amounts to Equation (2.2.6).

It is well known that

$$\mathbb{E}Q_{j,n} = \frac{1}{Z_n}\pi_{j,n}\frac{\partial Z_n}{\partial \pi_{j,n}}.$$

From (2.2.9), we get

$$\begin{aligned}
\mathbb{E}Q_{j,n} &= \frac{1}{2\pi Z_n}\int_{-\pi}^{\pi} \pi_{j,n}\lambda_n e^{\mathrm{i}\theta} f'_{j,n}(\pi_{j,n}\lambda_n e^{\mathrm{i}\theta})\frac{\prod_{k\neq j} f_{k,n}(\pi_{k,n}\lambda_n e^{\mathrm{i}\theta})}{(\lambda_n e^{\mathrm{i}\theta})^{m(n)}}d\theta, \\
&= \frac{F_n(\lambda_n)}{2\pi Z_n}\int_{-\pi}^{\pi}[m_{j,n}\varphi_{j,n}(\theta) - \mathrm{i}\varphi'_{j,n}(\theta)]\prod_{k\neq j}\varphi_{k,n}(\theta)d\theta,
\end{aligned}$$

and (2.2.7) follows.

For any $\ell \geq 1$, the joint distributions take the form

$$\mathrm{P}(Q_{1,n} = q_1, \ldots, Q_{\ell,n} = q_\ell) = \frac{1}{Z_n}\prod_{j=1}^{\ell}\frac{\pi_{j,n}^{q_j}}{\mu_{j,n}(1)\cdots\mu_{j,n}(q_j)}Z^*,$$

where Z^* is the normalizing constant corresponding to the network of queues $\ell+1, \ldots, n$ with $m(n) - \sum_{j=1}^{\ell} q_j = m(n) - \sum q_j$ clients. Following the same lines as above, we obtain (2.2.8), since

$$\begin{aligned}
&\mathrm{P}(Q_{1,n} = q_1, \ldots, Q_{\ell,n} = q_\ell) \\
&\quad = \frac{1}{2\pi Z_n}\prod_{j=1}^{\ell}\frac{\pi_{j,n}^{q_j}}{\mu_{j,n}(1)\cdots\mu_{j,n}(q_j)}\int_{-\pi}^{\pi}\frac{\prod_{k=\ell+1}^{n} f_{k,n}(\pi_{k,n}\lambda_n e^{\mathrm{i}\theta})}{(\lambda_n e^{\mathrm{i}\theta})^{m(n)-\sum q_j+1}}d\theta.
\end{aligned}$$

■

These formulas are used in Sections 2.3 and 2.4 to derive asymptotic expansions. These expansions are given in terms of the operators O and Ω defined as follows:

$$a(\eta) = O(b(\eta)) \text{ iff } \exists K > 0,\ \forall\eta,\ |a(\eta)| \leq K|b(\eta)|,$$

$$a(\eta) = \Omega(b(\eta)) \text{ iff } a(\eta) = O(b(\eta)) \text{ and } b(\eta) = O(a(\eta)),$$

where η is unspecified argument. Unless otherwise stated, all these bounds are uniform with respect to n and all queue indexes.

2.3 Asymptotic expansion for normal traffic

In this section, we compute estimates of several performance measures of $\mathcal{B}_n$ when the queues are not saturated (in a sense that will be made precise in Lemma 2.3.1). The method that we use to compute our estimates is related to the Central Limit Theorem or, more precisely, to Berry-Esseen type expansions. The results of next lemma can be interpreted as a proof that $(S_n - m(n))/\sigma_n$ converge in distribution to a normal law as $n \to \infty$. This is not surprising in view of the inversion relation

$$\mathrm{P}(S_n = m(n)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \varphi_n(\theta) d\theta.$$

Further information on this and related subjects can be found, for example, in Feller [17]. Note that it is possible by truncation methods to prove similar results without assuming the existence of moments.

Define $\chi^2_{k,n}$ from $X_{k,n}$ as in Lemma 2.A.1 of the appendix, and let

$$\chi_n^2 \stackrel{\text{def}}{=} \chi_{1,n}^2 + \cdots + \chi_{n,n}^2 \le \sigma_n^2.$$

Lemma 2.3.1 *Define, for any $0 < r \le 1$,*

$$\delta_n^r \stackrel{\text{def}}{=} \frac{1}{2} \frac{\sigma_n^2}{\beta_n^{(2+r)}}.$$

Let $\chi_n \delta_n \to \infty$ as $n \to \infty$. Then we have the following estimates.

(*i*) *For any integer $a \ge 0$,*

$$\begin{aligned} \int_{-\pi}^{\pi} \theta^a \varphi_n(\theta) d\theta &= \frac{1}{\sigma_n^{1+a}} \int_{-\infty}^{\infty} u^a e^{-\frac{u^2}{2}} du + O\Big(\frac{1}{\sigma_n^{1+a}} \frac{\beta_n^{(2+r)}}{\sigma_n^{2+r}}\Big) \\ &\quad + O\Big(\frac{1}{\chi_n^2 \delta_n^{1-a}} \exp\Big(-\frac{\chi_n^2 \delta_n^2}{5}\Big)\Big). \end{aligned} \qquad (2.3.1)$$

(*ii*) *When a is an even number, Equation (2.3.1) holds for $1 < r \le 2$, provided that $X_{1,n}, \ldots, X_{n,n}$ admit moments of order $2+r$ and choosing*

$$\delta_n = \frac{1}{2} \frac{\sigma_n^2}{\beta_n^{(3)}}.$$

(*iii*) *For any real number $a \ge 0$,*

$$\int_{-\pi}^{\pi} |\theta^a \varphi_n(\theta)| d\theta = O\Big(\frac{1}{\sigma_n^{1+a}}\Big) + O\Big(\frac{1}{\chi_n^2 \delta_n^{1-a}} \exp\Big(-\frac{\chi_n^2 \delta_n^2}{5}\Big)\Big). \qquad (2.3.2)$$

Proof We only prove (*i*) and (*ii*), since (*iii*) follows trivially from the computations. Define,

$$\begin{aligned} I_n &\stackrel{\text{def}}{=} \int_{-\pi}^{\pi} \theta^a \varphi_n(\theta) d\theta - \int_{-\infty}^{\infty} \theta^a e^{-\frac{\sigma_n^2 \theta^2}{2}} d\theta \\ &= \int_{-\delta_n}^{\delta_n} \theta^a \Big(\varphi_n(\theta) - e^{-\frac{\sigma_n^2 \theta^2}{2}}\Big) d\theta \\ &\quad - \int_{|\theta| \ge \delta_n} \theta^a e^{-\frac{\sigma_n^2 \theta^2}{2}} d\theta + \int_{\delta_n \le |\theta| \le \pi} \theta^a \varphi_n(\theta) d\theta, \end{aligned}$$

noting that $\delta_n \le 1/2$. It is known that

$$\int_{|\theta|\ge\delta_n} \theta^a e^{-\frac{\sigma_n^2\theta^2}{2}} d\theta \sim \frac{2}{\sigma_n^2\delta_n^{1-a}} e^{-\frac{\sigma_n^2\delta_n^2}{2}}.$$

Applying Lemma 2.A.1 to φ_n, we get

$$|\int_{\delta_n\le|\theta|\le\pi} \theta^a \varphi_n(\theta) d\theta| \le \int_{|\theta|\ge\delta_n} |\theta|^a e^{-\frac{\chi_n^2\theta^2}{5}} d\theta = O\Big(\frac{1}{\chi_n^2\delta_n^{1-a}} e^{-\frac{\chi_n^2\delta_n^2}{5}}\Big).$$

Finally,

$$\begin{aligned} |I_n| \quad &\le \quad \int_{-\delta_n}^{\delta_n} |\theta|^a \Big|\varphi_n(\theta) - e^{-\frac{\sigma_n^2\theta^2}{2}}\Big| d\theta \\ &\quad + O\Big(\frac{1}{\sigma_n^2\delta_n^{1-a}} e^{-\frac{\sigma_n^2\delta_n^2}{2}}\Big) + O\Big(\frac{1}{\chi_n^2\delta_n^{1-a}} e^{-\frac{\chi_n^2\delta_n^2}{5}}\Big). \end{aligned} \tag{2.3.3}$$

We proceed now to estimate the above integral, so that implicitly $|\theta| \le \delta_n$. The derivation relies on the following simple inequality, valid for all complex numbers $x_1, \dots, x_n$ and $y_1, \dots, y_n$:

$$|x_1 \cdots x_n - y_1 \cdots y_n| \le \sum_{k=1}^{n} |x_1 \cdots x_{k-1}||x_k - y_k||y_{k+1} \cdots y_n|, \tag{2.3.4}$$

which will be used with $x_k = \varphi_{k,n}(\theta)$ and $y_k = \exp(-\sigma_{k,n}^2\theta^2/2)$.

The characteristic function $\varphi_{k,n}$ of the random variable $X_{k,n}$ satisfies (see for example Loève [40])

$$\Big|\varphi_{k,n}(\theta) - 1 + \sigma_{k,n}^2\frac{\theta^2}{2}\Big| \le \beta_{k,n}^{(2+r)}\frac{|\theta|^{2+r}}{2}. \tag{2.3.5}$$

This implies that

$$\begin{aligned} \Big|\varphi_{k,n}(\theta) - e^{-\frac{\sigma_{k,n}^2\theta^2}{2}}\Big| \quad &\le \quad \Big|\varphi_{k,n}(\theta) - 1 + \sigma_{k,n}^2\frac{\theta^2}{2}\Big| + \Big|e^{-\frac{\sigma_{k,n}^2\theta^2}{2}} - 1 + \sigma_{k,n}^2\frac{\theta^2}{2}\Big| \\ &\le \quad \beta_{k,n}^{(2+r)}\frac{|\theta|^{2+r}}{2} + \sigma_{k,n}^{2+r}\frac{|\theta|^{2+r}}{2} \;\le\; \beta_{k,n}^{(2+r)}|\theta|^{2+r}. \end{aligned} \tag{2.3.6}$$

We have to find an upper bound for $|\varphi_{k,n}|$, assuming first $\sigma_{k,n}\delta_n \le 1$, so that

$$\begin{aligned} |\varphi_{k,n}(\theta)| \quad &\le \quad 1 - \sigma_{k,n}^2\frac{\theta^2}{2} + \beta_{k,n}^{(2+r)}\frac{|\theta|^{2+r}}{2} \\ &\le \quad \exp(-\sigma_{k,n}^2 + \beta_{k,n}^{(2+r)}\delta_n^r)\frac{\theta^2}{2} \;\stackrel{\text{def}}{=}\; \psi_{k,n}(\theta). \end{aligned} \tag{2.3.7}$$

Assume now that $\sigma_{k,n}\delta_n \ge 1$. Then,

$$-\sigma_{k,n}^2 + \beta_{k,n}^{(2+r)}\delta_n^r \ge -\sigma_{k,n}^2 + \sigma_{k,n}^{2+r}\delta_n^r \ge 0,$$

and $|\varphi_{k,n}(\theta)| \le 1 \le \psi_{k,n}(\theta)$.

Since, from the statement of the lemma, $\delta_n\sigma_n \to \infty$, we can choose n to ensure $\delta_n \ge 2/\sigma_n$. Thus,

$$\psi_{k,n}(\theta) \ge \exp\Big(-\sigma_{k,n}^2\frac{\theta^2}{2}\Big) \ge \exp\Big(-\sigma_n^2\frac{\theta^2}{8}\Big).$$

Using (2.3.4), (2.3.6) and (2.3.7), we find

$$\begin{aligned}\Big|\varphi_n(\theta) - e^{-\frac{\sigma_n^2\theta^2}{2}}\Big| &\le \beta_n^{(2+r)}|\theta|^{2+r}\exp\Big(-\sigma_n^2 + \beta_n^{(2+r)}\delta_n^r + \frac{\sigma_n^2}{4})\Big)\frac{\theta^2}{2}\\ &\le \beta_n^{(2+r)}|\theta|^{2+r}\exp\Big(-\sigma_n^2\frac{\theta^2}{8}\Big), \qquad (2.3.8)\end{aligned}$$

and (2.3.1) follows, since the integral in (2.3.3) is bounded by

$$\begin{aligned}\int_{-\delta_n}^{\delta_n}|\theta|^a\Big|\varphi_n(\theta) - e^{-\frac{\sigma_n^2\theta^2}{2}}\Big|d\theta &\le \beta_n^{(2+r)}\int_{-\infty}^{\infty}|\theta|^{2+r+a}\exp\Big(-\sigma_n^2\frac{\theta^2}{8}\Big)d\theta\\ &= O\Big(\frac{1}{\sigma_n^{1+a}}\frac{\beta_n^{(2+r)}}{\sigma_n^{2+r}}\Big).\end{aligned}$$

The proof of assertion (*ii*) of the lemma is similar, although the computations be more involved. In this case, since a is even , I_n can be rewritten as

$$I_n \stackrel{\text{def}}{=} \int_{-\pi}^{\pi}\theta^a\varphi_n(\theta)d\theta - \int_{-\infty}^{\infty}\theta^a\Big(1 - \mathrm{i}\bar\beta_n^{(3)}\frac{\theta^3}{6}\Big)e^{-\frac{\sigma_n^2\theta^2}{2}}d\theta,$$

where $\bar\beta_{k,n}^{(3)} \stackrel{\text{def}}{=} \mathbb{E}[X_{k,n} - m_{k,n}]^3$ and $\bar\beta_n^{(3)} \stackrel{\text{def}}{=} \mathbb{E}[X_n - m(n)]^3 = \bar\beta_{1,n}^{(3)} + \cdots + \bar\beta_{n,n}^{(3)}$. To find a bound for $|I_n|$, we have to estimate

$$\begin{aligned}&\Big|\varphi_n(\theta) - \Big(1 - \mathrm{i}\bar\beta_n^{(3)}\frac{\theta^3}{6}\Big)e^{-\frac{\sigma_n^2\theta^2}{2}}\Big| \qquad (2.3.9)\\ &\qquad\le \Big|\varphi_n(\theta) - e^{-\frac{\sigma_n^2\theta^2}{2} - \mathrm{i}\frac{\bar\beta_n^{(3)}\theta^3}{6}}\Big| + \Big|e^{-\mathrm{i}\frac{\bar\beta_n^{(3)}\theta^3}{6}} - 1 + \mathrm{i}\bar\beta_n^{(3)}\frac{\theta^3}{6}\Big|e^{-\frac{\sigma_n^2\theta^2}{2}}.\end{aligned}$$

The second part of the r.h.s. of (2.3.9) is trivially less than

$$\Big|\bar\beta_n^{(3)}\frac{\theta^3}{6}\Big|^{1+\frac{r}{3}}e^{-\frac{\sigma_n^2\theta^2}{2}} \le \beta_n^{(3+r)}\frac{|\theta|^{3+r}}{6}e^{-\frac{\sigma_n^2\theta^2}{2}},$$

and the first one is evaluated as above with (2.3.4) and

$$\psi_{k,n}(\theta) = \exp(-\sigma_{k,n}^2 + \beta_{k,n}^{(3)}\delta_n)\frac{\theta^2}{2}.$$

■

We are now in a position to state the main result of this section, which involves a Lyapounov's condition, popular in the Central Limit Problem.

Theorem 2.3.2 *Let $s \le r$ be two real numbers such that $0 < r \le 2$ and $0 < s \le 1$. Assume that $\sigma_n = O(\chi_n)$, that $\beta_n^{(2+r)}$ exists and $\beta_n^{(2+r)}/\sigma_n^{2+r} \to 0$ as $n \to \infty$. Then the following asymptotic expansions hold.*

(*i*)

$$Z_n = \frac{F_n(\lambda_n)}{\sigma_n\sqrt{2\pi}}\Big[1 + O\Big(\frac{\beta_n^{(2+r)}}{\sigma_n^{2+r}}\Big)\Big]. \qquad (2.3.10)$$

(*ii*) *For any $j \le n$,*

$$\mathbb{E}Q_{j,n} = \mathbb{E}X_{j,n} + O\Big(\frac{\beta_n^{(2+r)}}{\sigma_n^{2+r}}\Big) + O\Big(\frac{\beta_n^{(2+s)}\sigma_{j,n}^2 + \beta_{j,n}^{(2+s)}\sigma_n^2}{\sigma_n^{3+s}}\Big). \qquad (2.3.11)$$

(*iii*) *For any $\ell \le n$, if $[\sum_{j=1}^{\ell} m_{j,n}]/\sigma_n \to 0$,*

$$\mathrm{P}(Q_{1,n} = q_1, \ldots, Q_{\ell,n} = q_\ell) = \prod_{k=1}^{\ell} \mathrm{P}(X_{k,n} = q_k)\Big[1 + O(\varepsilon_n)\Big], \qquad (2.3.12)$$

$$\varepsilon_n = \frac{\beta_n^{(2+r)}}{\sigma_n^{2+r}} + \frac{\beta_n^{(2+s)}\sum_{j=1}^{\ell} m_{j,n}}{\sigma_n^{3+s}} + \frac{[\sum_{j=1}^{\ell} m_{j,n}]^2}{\sigma_n^2}.$$

Proof Equation (2.3.10) is a simple application of Lemma 2.3.1 to (2.2.6), since

$$\begin{aligned} Z_n &= \frac{F_n(\lambda_n)}{2\pi}\int_{-\pi}^{\pi}\prod_{k=1}^{n}\varphi_{k,n}(\theta)d\theta \\ &= \frac{F_n(\lambda_n)}{2\pi}\Big[\frac{\sqrt{2\pi}}{\sigma_n} + O\Big(\frac{\beta_n^{(2+r)}}{\sigma_n^{3+r}}\Big) + O\Big(\frac{1}{\chi_n^2\delta_n}\exp\Big(-\frac{\chi_n^2\delta_n^2}{5}\Big)\Big)\Big]. \end{aligned}$$

Like in (2.3.5), we can write

$$\varphi'_{j,n}(\theta) = -\sigma_{j,n}^2\theta + O(\beta_{j,n}^{(2+s)})\theta^{1+s},$$

and (2.2.7) reads:

$$\begin{aligned} \mathbb{E}Q_{j,n} &= m_{j,n} + \frac{F_n(\lambda_n)}{2\pi Z_n}\Big[\mathrm{i}\sigma_{j,n}^2\int_{-\pi}^{\pi}\theta\prod_{k\ne j}^{n}\varphi_{k,n}(\theta)d\theta \\ &\quad + O(\beta_{j,n}^{(2+s)})\int_{-\pi}^{\pi}\theta^{1+s}\prod_{k\ne j}^{n}\varphi_{k,n}(\theta)d\theta\Big] \\ &= m_{j,n} + \frac{F_n(\lambda_n)}{2\pi Z_n}\Big[O\Big(\frac{\beta_n^{(2+s)}\sigma_{j,n}^2}{\sigma_n^{4+s}}\Big) + O\Big(\frac{\beta_{j,n}^{(2+s)}}{\sigma_n^{2+s}}\Big)\Big], \end{aligned}$$

and (2.3.11) follows. Equation (2.3.12) is obtained in the same way from (2.2.8) and

$$e^{\mathrm{i}\sum_{j=1}^{\ell}(q_j-m_{j,n})\theta}=1-\mathrm{i}\Big[\sum_{j=1}^{\ell}q_j-m_{j,n}\Big]\theta+O\Big(\Big[\sum_{j=1}^{\ell}q_j-m_{j,n}\Big]^2\Big)\theta^2.$$

■

2.4 Asymptotic expansion with heavily loaded queues

In this section, we analyze the network $\mathcal{B}_n$ when some queues become heavily loaded. This, in particular, implies that the Lyapounov condition $\beta_n^{(2+r)}/\sigma_n^{2+r}\to 0$ is no more valid. In fact, after proper scaling, $S_n-m(n)$ will be shown to converge in distribution to a random variable having a gamma distribution, under the broad assumption that the first singularities of the relevant generating functions are algebraic.

Lemma 2.4.1 *Assume that we can write for some fixed $\xi\geq 1$*

$$\varphi_n(\theta)=\omega_n^{\xi}(\theta)\widehat{\varphi}_n(\theta),$$

where $\widehat{\varphi}_n$ is a characteristic function and

$$\omega_n(\theta)\stackrel{\text{def}}{=}\frac{(1-\rho_n)e^{-\mathrm{i}\alpha_n\theta}}{1-\rho_n e^{\mathrm{i}\theta}},$$

with $0\leq\rho_n<1$ and $\alpha_n\stackrel{\text{def}}{=}\rho_n/(1-\rho_n)$. We fix r, $0<r\leq 1$. Let $\hat{\sigma}_n$, $\hat{\beta}_n^{(2+r)}$, $\hat{\chi}_n$ and $\hat{\delta}_n$ be quantities having the same meaning as in Lemma 2.3.1, but related to $\widehat{\varphi}_n(\theta)$.

Let $\hat{\sigma}_n/\alpha_n\to 0$ and $\hat{\delta}_n\hat{\chi}_n\to\infty$ as $n\to\infty$. Then the following estimates hold:

$$\begin{aligned}\alpha_n\int_{-\pi}^{\pi}\varphi_n(\theta)d\theta &= \int_{-\infty}^{\infty}\frac{e^{-\mathrm{i}\xi u}}{(1-\mathrm{i}u)^{\xi}}\exp\Big(-\frac{\hat{\sigma}_n^2}{\alpha_n^2}\frac{u^2}{2}\Big)du \qquad (2.4.1)\\ &\quad+O\Big(\frac{1}{\alpha_n}+\frac{\hat{\beta}_n^{(2+r)}}{\hat{\sigma}_n^{2+r}}\Big(\frac{\hat{\sigma}_n}{\alpha_n}\Big)^{2+r}+\frac{\hat{\beta}_n^{(2+r)}}{\hat{\sigma}_n^{2+r}}\Big(\frac{\hat{\sigma}_n}{\alpha_n}\Big)^{\xi-1}\Big)\\ &\quad+O\Big(\frac{e^{-\frac{\hat{\chi}_n^2\hat{\delta}_n^2}{5}}}{\hat{\chi}_n^2\hat{\delta}_n^{\xi+1}\alpha_n^{\xi-1}}\Big),\\ \alpha_n\int_{-\pi}^{\pi}|\theta\varphi_n(\theta)|d\theta &= O\Big(\frac{1}{\alpha_n}+\frac{1}{\hat{\sigma}_n}\Big(\frac{\hat{\sigma}_n}{\alpha_n}\Big)^{\xi-1}\Big)+O\Big(\frac{e^{-\frac{\hat{\chi}_n^2\hat{\delta}_n^2}{5}}}{\hat{\chi}_n^2\hat{\delta}_n^{\xi}\alpha_n^{\xi-1}}\Big). \qquad (2.4.2)\end{aligned}$$

Proof The proof of this lemma is similar to the proof of Lemma 2.3.1 and is only sketched here. Define

$$\omega(u)\stackrel{\text{def}}{=}\frac{e^{-\mathrm{i}u}}{1-\mathrm{i}u},$$

and

$$\begin{aligned}
I_n &\stackrel{\text{def}}{=} \alpha_n \int_{-\pi}^{\pi} \omega_n^{\xi}(\theta)\widehat{\varphi}_n(\theta)d\theta - \int_{-\infty}^{\infty} \omega^{\xi}(u)\exp\Big(-\frac{\hat{\sigma}_n^2}{\alpha_n^2}\frac{u^2}{2}\Big)du \\
&= \int_{-\pi\alpha_n}^{\pi\alpha_n} \omega_n^{\xi}(u/\alpha_n)\Big[\widehat{\varphi}_n(u/\alpha_n) - \exp\Big(-\frac{\hat{\sigma}_n^2}{\alpha_n^2}\frac{u^2}{2}\Big)\Big]du \\
&\quad + \int_{-\pi\alpha_n}^{\pi\alpha_n} \Big[\omega_n^{\xi}(u/\alpha_n) - \omega^{\xi}(u)\Big]\exp\Big(-\frac{\hat{\sigma}_n^2}{\alpha_n^2}\frac{u^2}{2}\Big)du \\
&\quad - \int_{|u|\geq\pi\alpha_n} \omega^{\xi}(u)\exp\Big(-\frac{\hat{\sigma}_n^2}{\alpha_n^2}\frac{u^2}{2}\Big)du. \qquad (2.4.3)
\end{aligned}$$

The evaluation of these integrals depends on the following straightforward estimations, valid for $|u| < \pi\alpha_n$:

$$\begin{aligned}
|\omega_n^{\xi}(u/\alpha_n)| &= O\Big(\frac{1}{(1+u^2)^{\xi/2}}\Big) \\
|\omega_n^{\xi}(u/\alpha_n) - \omega^{\xi}(u)| &= O\Big(\frac{1}{\alpha_n}\frac{u^2}{(1+u^2)^{\xi}}\Big),
\end{aligned}$$

and from (2.3.8), which reads for $|u| < \alpha_n\hat{\delta}_n$:

$$\Big|\widehat{\varphi}_n(u/\alpha_n) - \exp\Big(-\frac{\hat{\sigma}_n^2}{\alpha_n^2}\frac{u^2}{2}\Big)\Big| = O\Big(\frac{\hat{\beta}_n^{(2+r)}}{\alpha_n^{2+r}}\Big)u^{2+r}\exp\Big(-\frac{\hat{\sigma}_n^2}{\alpha_n^2}\frac{u^2}{8}\Big).$$

Moreover, we use the following approximation, valid for $a, b > 0$ and for small z:

$$J(a,b,z) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} \frac{|u|^a}{(1+u^2)^b}e^{-z^2u^2}du = O(1) + O(z^{2b-a-1}).$$

These relations, together with (2.4.3), yield:

$$\begin{aligned}
I_n &= O\Big(\frac{1}{\alpha_n}\frac{\hat{\beta}_n^{(2+r)}}{\alpha_n^{2+r}}\Big)J\Big(2+r, \frac{\xi}{2}, \frac{\hat{\sigma}_n}{\alpha_n}\Big) + O\Big(\frac{1}{\alpha_n^2}\Big)J\Big(2, \xi, \frac{\hat{\sigma}_n}{\alpha_n}\Big) \\
&\quad + O\Big(\Big(\frac{\hat{\sigma}_n}{\alpha_n}\Big)^{\xi-1}\Big)\int_{v\geq\pi\hat{\sigma}_n} v^{-\xi}e^{-\frac{v^2}{2}}dv \\
&\quad + O\Big(\Big(\frac{\hat{\chi}_n}{\alpha_n}\Big)^{\xi-1}\Big)\int_{v\geq\delta_n\hat{\chi}_n} v^{-\xi}e^{-\frac{v^2}{2}}dv,
\end{aligned}$$

and (2.4.1) follows. Relation (2.4.2) is obtained from the estimation

$$\begin{aligned}
\int_{-\delta_n\alpha_n}^{\delta_n\alpha_n} \Big|u\omega_n^{\xi}(u/\alpha_n)\varphi_n(u/\alpha_n)\Big|du &\leq \int_{-\infty}^{\infty} \frac{|u|}{(1+u^2)^{\xi/2}}\exp\Big(-\frac{\hat{\sigma}_n^2}{\alpha_n^2}\frac{u^2}{4}\Big)du \\
&= O(1) + O\Big(\Big(\frac{\hat{\sigma}_n}{\alpha_n}\Big)^{\xi-2}\Big).
\end{aligned}$$

■

The estimates of Lemma 2.4.1 allow to establish the main result of this section, with the same notation.

Theorem 2.4.2 *Assume that the queues can be split into two categories:*

(*i*) *a set $\mathcal{S}_n$ of "saturated" queues, for which*

$$\varphi_{k,n}(\theta) = \omega_n^{\xi_{k,n}}(\theta)\psi_{k,n}(\theta),$$

with $\psi_{k,n}(\theta) = O(1)$ and $\psi'_{k,n}(\theta) = O(\theta)$, uniformly in k and n;

(*ii*) *"normal" queues, with total characteristic function $\widehat{\varphi}_n(\theta)$.*

Define $\xi_n = \sum_{k\in\mathcal{S}_n} \xi_{k,n}$ and let r be a real number, $0 < r \le 1$.

Assume $\hat{\sigma}_n/\alpha_n \to 0$, $\hat{\beta}_n^{(2+r)}/\hat{\sigma}_n^{2+r} \to 0$ and $\hat{\sigma}_n = O(\hat{\chi}_n)$ as $n \to \infty$. Then the following expansions hold.

(*i*)

$$Z_n = \frac{F_n(\lambda_n)}{\alpha_n}\frac{\xi_n^{\xi_n-1}e^{-\xi_n}}{\Gamma(\xi_n)}[1 + O(\varepsilon_{1,n})], \tag{2.4.4}$$

with

$$\varepsilon_{1,n} = \Big(\frac{\hat{\sigma}_n}{\alpha_n}\Big)^2 + \frac{1}{\alpha_n} + \frac{\hat{\beta}_n^{(2+r)}}{\hat{\sigma}_n^{2+r}}\Big(\frac{\hat{\sigma}_n}{\alpha_n}\Big)^{\xi_n-1};$$

(*ii*) *for any $j \le n$,*

$$\mathbb{E}Q_{j,n} = \mathbb{E}X_{j,n}[1 + O(\varepsilon_{1,n} + \varepsilon_{2,n})], \tag{2.4.5}$$

with

$$\varepsilon_{2,n} = \begin{cases} \dfrac{\sigma_{j,n}^2}{\alpha_n} + \dfrac{\sigma_{j,n}^2}{\hat{\sigma}_n}\Big(\dfrac{\hat{\sigma}_n}{\alpha_n}\Big)^{\xi_n-1} & \text{if } j \notin \mathcal{S}_n,\\[2ex] \dfrac{1}{\alpha_n} + \dfrac{1}{\hat{\sigma}_n}\Big(\dfrac{\hat{\sigma}_n}{\alpha_n}\Big)^{\xi_n-1} & \text{otherwise;} \end{cases}$$

(*iii*) *for any $\ell \le n$, if $\mathcal{S}_n \cap [1, j] = \emptyset$,*

$$\mathrm{P}(Q_{1,n} = q_1, \ldots, Q_{\ell,n} = q_\ell) = \prod_{k=1}^{n} \mathrm{P}(X_{k,n} = q_k)\Big[1 + O(\varepsilon_{1,n} + \varepsilon_{3,n})\Big], \tag{2.4.6}$$

with

$$\varepsilon_{3,n} = \sum_{j=1}^{\ell} m_{j,n}\Big(\frac{1}{\alpha_n} + \frac{1}{\hat{\sigma}_n}\Big(\frac{\hat{\sigma}_n}{\alpha_n}\Big)^{\xi_n-1}\Big).$$

Proof The proof is essentially identical to the proof of Theorem 2.3.2. The first term of (2.4.1) is evaluated using Parseval's identity and classical complex analysis tools, giving

$$\int_{-\infty}^{\infty} \omega_n^{\xi}(u)\exp\Big(-\frac{\hat{\sigma}_n^2}{\alpha_n^2}\frac{u^2}{2}\Big)du = \frac{\xi_n^{\xi_n-1}e^{-\xi_n}}{\Gamma(\xi_n)}\Big[1 + O\Big(\frac{\hat{\sigma}_n^2}{\alpha_n^2}\Big)\Big].$$

∎

2.5 The choice of a good scaling

One thing that has been left out in the previous sections is the determination of "interesting" values for $m(n)$ or, equivalently, for λ_n. The analysis in [32, 33, 34, 2, 41] was done assuming that the number of customers in the system should be proportional to the number of stations. Indeed, we shall see that this is not always the suitable scaling. In order to exhibit meaningful behaviours, it will be convenient to make additional assumptions on the model. However, it should be pointed out that the results of Sections 2.3 and 2.4 are valid in a more general setting, which would allow to carry out special cases via the same methods.

The queues are partitioned as follows:

$$
\begin{aligned}
\mathcal{F}_n &\stackrel{\text{def}}{=} \{0 \le k \le n \ :\ \liminf_{q\to\infty} \mu_{k,n}(q) < +\infty\},\\
\mathcal{I}_n &\stackrel{\text{def}}{=} \{0 \le k \le n \ :\ \lim_{q\to\infty} \mu_{k,n}(q) = +\infty\},
\end{aligned}
$$

and we define

$$
\begin{aligned}
\mu_{k,n} &\stackrel{\text{def}}{=} \begin{cases} \liminf_{q\to\infty} \mu_{k,n}(q) & \text{if } k \in \mathcal{F}_n \\ \mu_{k,n}(1) & \text{if } k \in \mathcal{I}_n \end{cases},\\
\rho_{k,n} &\stackrel{\text{def}}{=} \frac{\lambda_n \pi_{k,n}}{\mu_{k,n}},\\
\rho_n &\stackrel{\text{def}}{=} \sup_{k\in\mathcal{F}_n} \rho_{k,n},\\
\frac{1}{\bar{\mu}_n(1)} &\stackrel{\text{def}}{=} \sum_{k=1}^{n} \frac{\pi_{k,n}}{\mu_{k,n}(1)},
\end{aligned}
$$

The analysis is done under the following assumption on the service disciplines.

Assumption A6 *There exist sequences $R(q)$ and $T(q)$ such that*

$$\lim_{q\to\infty} R(q) = 1,$$

$$\lim_{q\to\infty} T(q) = +\infty,$$

and, for any $q > 0$,

$$
\begin{aligned}
\frac{\mu_{k,n}(q)}{\mu_{k,n}} &\ge R(q) \quad \textit{for } k \in \mathcal{F}_n,\\
\frac{\mu_{k,n}(q)}{\mu_{k,n}} &\ge T(q) \quad \textit{for } k \in \mathcal{I}_n.
\end{aligned}
$$

This assumption holds in particular for any mixing of infinite-server and multiple-server queues, the number of servers remaining bounded. Moreover, for the sake of simplicity, we will assume

Assumption A7 $\bar{\mu}_n(1) = \Omega(1)$.

This ensures that the mean service times are bounded away from zero and from infinity. When this assumption is not satisfied, the results of this section still hold, with λ_n replaced by $\lambda_n/\bar{\mu}_n(1)$ in the error terms.

The first theorem deals with scalings, where no queue of the network $\mathcal{B}_n$ becomes saturated.

Theorem 2.5.1 *Assume that* **A6** *and* **A7** *hold and that there exists constants A and B such that*

$$\lim_{n\to\infty} \lambda_n = +\infty, \tag{2.5.1}$$

$$\rho_{k,n} \le B < 1, \ \text{for all } k \in \mathcal{F}_n, \tag{2.5.2}$$

$$\rho_{k,n} \le A, \ \text{for all } k \in \mathcal{I}_n. \tag{2.5.3}$$

Then the following expansions hold, for all $j \le n$.

$$Z_n = \frac{F_n(\lambda_n)}{\sigma_n\sqrt{2\pi}}\Big[1 + O\Big(\frac{1}{\lambda_n}\Big)\Big], \tag{2.5.4}$$

$$\mathbb{E}Q_{j,n} = \mathbb{E}X_{j,n}\Big[1 + O\Big(\frac{1}{\lambda_n}\Big)\Big], \tag{2.5.5}$$

$$\begin{aligned} &\mathrm{P}(Q_{1,n} = q_1, \ldots, Q_{\ell,n} = q_j) \\ &\quad = \mathrm{P}(X_{1,n} = q_1, \ldots, X_{\ell,n} = q_j)\Big[1 + O\Big(\frac{1}{\lambda_n}\Big)\Big]. \end{aligned} \tag{2.5.6}$$

Moreover, $m(n) = \Omega(\lambda_n)$.

Proof This theorem is a simple rewriting of Theorem 2.3.2, using **A6** and (2.5.1)–(2.5.3). For a queue $k \in \mathcal{F}_n$, and for all $a \in \mathbb{N}$, we have

$$\sum_{q=0}^{\infty} q^a \frac{(\lambda_n \pi_{k,n})^q}{\mu_{k,n}(1)\cdots\mu_{k,n}(q)} \le \sum_{q=0}^{\infty} \frac{q^a B^q}{R(1)\cdots R(q)} < +\infty.$$

In particular, $f_{k,n}(\lambda_n\pi_{k,n}) = \Omega(1)$ and

$$m_{k,n} = \frac{\lambda_n\pi_{k,n}}{\mu_{k,n}(1) f_{k,n}(\lambda_n\pi_{k,n})} \sum_{q=1}^{\infty} q \frac{(\lambda_n\pi_{k,n})^{q-1}}{\mu_{k,n}(2)\cdots\mu_{k,n}(q)} = \Omega\Big(\frac{\lambda_n\pi_{k,n}}{\mu_{k,n}(1)}\Big).$$

Similarly, for any $r \in \mathbb{N}$,

$$\beta_{k,n}^{(r)} = \Omega(\frac{\lambda_n\pi_{k,n}}{\mu_{k,n}(1)}),$$

and the results of Theorem 2.3.2 apply. The remaining computations are done in a similar fashion for all types of queues. ∎

Theorem 2.5.1 can be easily generalized to the situation where some queues of $\mathcal{I}_n$ become saturated, in which case (2.5.3) fails to be satisfied.

Corollary 2.5.2 *Let* **A6**, **A7**, *(2.5.1) and (2.5.2) be satisfied and assume that the queues for which (2.5.3) does not hold are infinite-server queues. Then the conclusions of Theorem 2.5.1 remain valid, except (2.5.6), which holds only for queues satisfying (2.5.2) or (2.5.3).*

Proof This proof relies on the fact that the characteristic function of the number of clients X in an infinite server queue with parameter ρ can be written as

$$\mathbb{E}e^{\mathrm{i}\theta X} = \exp\Big(\rho(e^{\mathrm{i}\theta} - 1)\Big) = \left[\exp\Big(\frac{\rho}{\lfloor\rho\rfloor}(e^{\mathrm{i}\theta} - 1)\Big)\right]^{\lfloor\rho\rfloor},$$

which means that a saturated infinite server queue can be replaced by several non-saturated infinite-server queues without changing the distribution of S_n. With this substitution, the proof Theorem 2.5.1 applies without change. ■

To handle the case of saturated queues in $\mathcal{F}_n$, we have to introduce the subset of $\mathcal{F}_n$

$$\mathcal{S}_n \stackrel{\text{def}}{=} \{k \in \mathcal{F}_n \ : \ \rho_{k,n} = \rho_n\},$$

and to assume:

Assumption A8 *The cardinal of $\mathcal{S}_n$ is uniformly bounded and, for all $k \in \mathcal{S}_n$, $f_{k,n}(z)$ has its first singularity at $z = \mu_{k,n}$, which is algebraic of order $\xi_{k,n}$. Moreover, for some $\xi > 0$,*

$$\xi_n \stackrel{\text{def}}{=} \sum_{k\in\mathcal{S}_n} \xi_{k,n} \in [1, \xi].$$

While this assumption covers a wide range of known queues, other kind of singularities could be handled via the same method.

The next theorem illustrates what happens when $\rho_n \to 1$.

Theorem 2.5.3 *Assume that* **A6**, **A7** *and* **A8** *hold and that there exists constants A and B such that*

$$\lim_{n\to\infty} \rho_n = 1, \tag{2.5.7}$$

$$\rho_{k,n} \le B < 1, \ \textit{for all } k \in \mathcal{F}_n \setminus \mathcal{S}_n, \tag{2.5.8}$$

$$\rho_{k,n} \le A, \ \textit{for all } k \in \mathcal{I}_n. \tag{2.5.9}$$

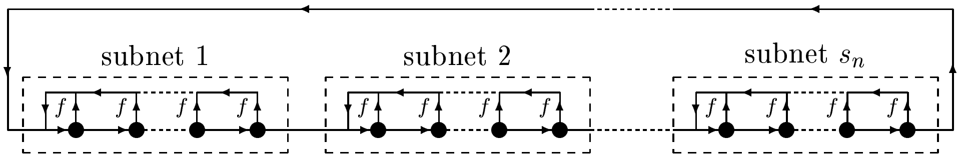


Figure 2.1: a compound network of tandem queues

(i) *If* $\lim_{n\to\infty}(1-\rho_n)^2\lambda_n = +\infty$, *then the following expansions hold for any* $j \le n$:

$$
\begin{aligned}
Z_n &= \frac{F_n(\lambda_n)}{\sigma_n\sqrt{2\pi}}\left[1+O\left(\frac{1}{\lambda_n}+\frac{1}{\lambda_n^2(1-\rho_n)^4}\right)\right],\\
\mathbb{E}Q_{j,n} &= \mathbb{E}X_{j,n}\left[1+O\left(\frac{1}{\lambda_n}+\frac{1}{\lambda_n^2(1-\rho_n)^4}\right)\right], \ \textit{if} \ \ j\notin\mathcal{S}_n,\\
\mathbb{E}Q_{j,n} &= \mathbb{E}X_{j,n}\left[1+O\left(\frac{1}{\lambda_n(1-\rho_n)^2}\right)\right], \ \textit{if} \ \ j\in\mathcal{S}_n,\\
\mathrm{P}(Q_{1,n} &= q_1,\ldots,Q_{\ell,n}=q_j)\\
&= \prod_{k=1}^{n}\mathrm{P}(X_{k,n}=q_k)\left[1+O\left(\frac{1}{\lambda_n}+\frac{1}{\lambda_n^2(1-\rho_n)^4}\right)\right],\\
&\qquad\qquad\qquad\qquad \textit{if}\ [1,j]\cap\mathcal{S}_n=\emptyset.
\end{aligned}
$$

(ii) *if* $\lim_{n\to\infty}(1-\rho_n)^2\lambda_n = 0$, *then the expansions of Theorem 2.4.2 hold with* $r=1$ *and*

$$
\hat\beta_n^{(3)} = \Omega(\lambda_n), \quad \hat\sigma_n^2 = \Omega(\lambda_n),
$$
$$
\alpha_n = \Omega\left(\frac{1}{1-\rho_n}\right), \qquad m(n) = \Omega\left(\frac{1}{1-\rho_n}\right) + \Omega(\lambda_n).
$$

The proof of this theorem is straightforward and is therefore omitted. There is also an obvious equivalent of Corollary 2.5.2, (when (2.5.9) does not hold) which will not be explicitly stated.

As pointed out at the beginning of this section, there are cases of interest with $\lambda_n = o(n)$ and $m(n) = o(n)$. This will be illustrated in the next example.

Example Consider a closed network consisting of s_n subnetworks of $M/M/1$ queues having each a unique entry point, in which a fixed number m of tasks circulate. The queues are subject to failures, taking place with some probability f. When a failure occurs, the task returns to the entry point of its current subnetwork. Tasks visit the various subnetworks according to some probability matrix.

This model exhibits tight bottlenecks, when the number and the size of the subnetworks grow. This fact, for the sake of simplicity, will be illustrated on a very simple topology, presented in Figure 2.1: all subnetworks are associated in tandem, and each of them consists itself of ℓ_n queues in tandem, with unit processing rates.

Here, the invariant measure of the routing matrix has the form

$$(\pi_1, \ldots, \pi_{\ell_n}; \pi_1, \ldots; \ldots, \pi_{\ell_n}),$$

where π_k is the invariant probability associated to the k-th queue of any subnetwork. A straightforward computation, using symmetry properties, yields

$$\pi_k = \frac{1}{s_n} \frac{f(1-f)^k}{1-(1-f)^{\ell_n}}.$$

Therefore, the s_n entry points of the subnetworks will act as bottlenecks for the system, since the size of the queues remain uniformly bounded if, and only if, $\lambda_n \pi_1 < B < 1$. It suffices to take

$$\lambda_n \leq \frac{s_n B}{f(1-f)}.$$

Thus, from Theorem 2.5.1, it follows that

$$m(n) = \Omega(\lambda_n) = \Omega(s_n).$$

Thus, since $n = \ell_n s_n$, the function $m(n)$ has an arbitrary asymptotic behaviour.

2.6 Application to a service vehicle network

Consider a fleet of vehicles serving an area consisting of n stations forming a fully connected graph. These vehicles are used to transport goods or passengers. Vehicles wait at stations until they receive a request, in which case they go to an other station. The routing among stations is done according to some routing matrix P_n. When a request arrives to an empty station, it is immediately lost. The request arrivals form a Poisson stream at each queue.

We model this system as follows: for all $0 \leq k \leq n$, station k is represented as a single-server queue with service rate $\mu_{k,n}$ which is equal to the arrival rate at station k, since arrivals are lost when the station is empty. When a vehicle leaves station k, it chooses its destination according to the Markovian routing matrix $P_n = (p_{k\ell,n})$. The duration of the journey between two stations k and ℓ is represented by an infinite server queue placed on the edge between them. The service rate of this queue when there are q vehicles traveling between k and ℓ is $\mu_{k\ell,n} q$. Note that, contrary to the convention used throughout this chapter, the total number of *queues* is $n^2 + n$.

Let $(\pi_{1,n},\ldots,\pi_{n,n})$ be the invariant measure of P_n, defined as in (2.2.1). Then, with obvious notation, for all $k,\ell\in[1,n]$, for all $\theta\in[-\pi,\pi]$,

$$\begin{aligned}
\rho_{k,n} &\stackrel{\text{def}}{=} \frac{\lambda_n \pi_{k,n}}{2\mu_{k,n}},\\
m_{k,n} &\stackrel{\text{def}}{=} \frac{\rho_{k,n}}{1-\rho_{k,n}},\\
\varphi_{k,n}(\theta) &\stackrel{\text{def}}{=} \frac{(1-\rho_{k,n})e^{-\mathrm{i}m_{k,n}\theta}}{1-\rho_{k,n}e^{\mathrm{i}\theta}},\\
\rho_{k\ell,n} &\stackrel{\text{def}}{=} \frac{\lambda_n \pi_{k,n} p_{k\ell,n}}{2\mu_{k\ell,n}},\\
m_{k\ell} &\stackrel{\text{def}}{=} \rho_{k\ell,n},\\
\varphi_{k\ell,n}(\theta) &\stackrel{\text{def}}{=} \exp\Big(\rho_{k\ell,n}(e^{\mathrm{i}\theta}-1-\mathrm{i}\theta)\Big),\\
\varphi_n(\theta) &\stackrel{\text{def}}{=} \exp\left[(e^{\mathrm{i}\theta}-1-\mathrm{i}\theta)\sum_{k,\ell=1}^{n}\rho_{k\ell,n}\right]\prod_{k=1}^{n}\frac{(1-\rho_{k,n})e^{-\mathrm{i}m_{k,n}\theta}}{1-\rho_{k,n}e^{\mathrm{i}\theta}},
\end{aligned}$$

Some questions of interest come to mind:

(i) what is the maximal efficiency that can be expected from this system?

(ii) how many vehicles should be provided?

Among all performance measures which can be derived from Theorems 2.5.1 and 2.5.3, we define the *loss probability* as

$$\mathcal{P}_{loss}(n) \stackrel{\text{def}}{=} \frac{\sum_{k=1}^{n}\mu_{k,n}\,\mathrm{P}(Q_{k,n}=0)}{\sum_{k=1}^{n}\mu_{k,n}}$$

Moreover, we introduce the *mean station time* and the *mean travel time* of a vehicle, respectively defined as:

$$\begin{aligned}
\bar{\tau}_n^{(s)} &\stackrel{\text{def}}{=} \sum_{k=1}^{n}\frac{\pi_{k,n}}{2\mu_{k,n}},\\
\bar{\tau}_n^{(t)} &\stackrel{\text{def}}{=} \sum_{k,\ell=1}^{n}\frac{\pi_{k,n}p_{k\ell,n}}{2\mu_{k\ell,n}}.
\end{aligned}$$

and we assume that these quantities are uniformly bounded

$$\bar{\tau}_n^{(s)}=\Omega(1)\quad \bar{\tau}_n^{(t)}=\Omega(1),$$

so that **A7** holds. Note that when these assumptions are not fulfilled, the system can be considered as flawed. Define $\mathcal{S}_n$ as in Section 2.5 and assume that its cardinal is some fixed integer $K\ge 1$. This implies that assumptions **A6** and **A8** hold. Theorems 2.5.1 and 2.5.3 can then be rewritten as:

Theorem 2.6.1 *Assume that (2.5.8) holds. Recall from Section 2.5 that*

$$\rho_n = \lambda_n \sup_{1\le k\le n} \frac{\pi_{k,n}}{2\mu_{k,n}} < 1.$$

(*i*) *if* $\lim_{n\to\infty}(1-\rho_n)^2\lambda_n = +\infty$ *then*

$$\begin{aligned}
\mathcal{P}_{loss}(n) &= \frac{\sum_{k\notin\mathcal{S}_n}\mu_{k,n}(1-\rho_{k,n})}{\sum_{k=1}^n \mu_{k,n}}\left[1+O\left(\frac{1}{\lambda_n}+\frac{1}{\lambda_n^2(1-\rho_n^4)}\right)\right],\\
m(n) &= \Omega(\lambda_n);
\end{aligned}$$

(*ii*) *if* $\lim_{n\to\infty}(1-\rho_n)^2\lambda_n = 0$ *then*

$$\begin{aligned}
\mathcal{P}_{loss}(n) &= \frac{\sum_{k\notin\mathcal{S}_n}\mu_{k,n}(1-\rho_{k,n})}{\sum_{k=1}^n \mu_{k,n}}\left[1+O\left(\frac{1}{\sqrt{\lambda_n}}\right)\right], \ \textit{if } K=1,\\
\mathcal{P}_{loss}(n) &= \frac{\sum_{k\notin\mathcal{S}_n}\mu_{k,n}(1-\rho_{k,n})}{\sum_{k=1}^n \mu_{k,n}}\left[1+O(1-\rho_n)\right], \ \textit{if } K\ge 2,\\
m(n) &= \Omega\left(\frac{1}{1-\rho_n}\right)+\Omega(\lambda_n).
\end{aligned}$$

Define the minimum loss probability

$$\mathcal{P}^*_{loss}(n) = \inf_{m(n)\in\mathbb{N}} \mathcal{P}_{loss}(n).$$

Since, as $n\to\infty$,

$$\mathcal{P}^*_{loss}(n) \sim 1 - \frac{\inf_{1\le k\le n}\frac{\mu_{k,n}}{\pi_{k,n}}}{\sum_{k=1}^n \mu_{k,n}},$$

it is useless to increase the number of vehicles in the system beyond a quantity of order

$$m(n) = \Omega\left(\inf_{1\le k\le n}\frac{\mu_{k,n}}{\pi_{k,n}}\right).$$

Consequently, having a number of vehicle proportional to the number of stations can be a poor choice, especially when some stations are more loaded than the others. These stations act as *bottlenecks* of the system, which can be removed by changing the routing probabilities.

2.7 General remarks

(*i*) In Section 2.3, the reader may be tempted to raise questions about possible caracterizations of the original network by means of other families of limit laws. For example, it is easy to choose a proper normalization a_n ensuring that

$$\frac{S_n}{a_n} = \frac{X_{1,n} + \cdots + X_{n,n}]}{a_n} \xrightarrow{w} \mathcal{P}(\lambda),$$

under standard assumptions, where $\mathcal{P}(\lambda)$ denotes a Poisson law of parameter λ. The key point is then to "lift" these results onto the original network, otherwise they would be of little interest. In fact, this might well be the case for the following reason: most of the estimates in Section 2.3 were in fact obtained by computing deviations of S_n around $\mathbb{E}S_n = m(n)$. In the case just quoted (*i.e.* Poisson limit law), there is no natural reason for working "around the mean".

(*ii*) In the case of saturated queues, additional assumptions had to be introduced. The main reason is that, in this case, the standard "uniform asymptotic negligeability" conditions do not hold, which means in particular that we are no more in the classical framework of the Central Limit Theorem.

2.A Appendix: a bound on periodic characteristic functions

One of the problems arising in the computation of convergence rates in the Central Limit Theorem is to find upper bounds on the modulus of a characteristic function $\varphi(\theta)$ for θ away from 0. One typical property used is usually stated as follows:

there exist $\theta_0 > 0$ and $a < 1$ such that, for all $|\theta| > \theta_0$, $|\varphi(\theta)| < a$.

It is pointed out in Feller [17] that this condition is usually easy to fulfill in practice, as long as X does not have a lattice distribution. Unfortunately, we are in the lattice case and we must cope with problem of periodicity of φ.

We show therefore in next lemma how a bound on $|\varphi(\theta)|$ can be derived for $|\theta| \leq \pi$.

Lemma 2.A.1 *Let X be an integer-valued random variable with distribution $P(X = k) = p_k$, $k \in \mathbb{N}$. Define*

$$\chi^2 \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} \frac{p_{2k}p_{2k+1}}{p_{2k} + p_{2k+1}},$$

where the summands are taken to be zero when $p_{2k} = p_{2k+1} = 0$. Then, for any $\theta \in [0, \pi]$, the characteristic function φ of X satisfies:

$$|\varphi(\theta)| \leq \exp\Big(-\frac{\chi^2}{5}\theta^2\Big). \tag{2.A.1}$$

χ^2 is always less than $\min(\operatorname{Var} X, 1/4)$*. If, in addition, there exists a real number ρ such that $p_{k+1} \le \rho p_k$, then*

$$\frac{1-p_0}{(1+\rho)^2} \le \chi^2 \le \frac{\rho(p_0+\rho)}{(1+\rho)^2}. \tag{2.A.2}$$

Proof We have

$$|\varphi(\theta)| = \Big|\sum_{k=0}^{\infty} p_k e^{\mathrm{i}k\theta}\Big| \le \sum_{k=0}^{\infty} \Big|p_{2k} + p_{2k+1} e^{\mathrm{i}\theta}\Big|.$$

Moreover,

$$\begin{aligned} \Big|p_{2k} + p_{2k+1} e^{\mathrm{i}\theta}\Big| &= \sqrt{(p_{2k} + p_{2k+1}\cos\theta)^2 + p_{2k+1}^2 \sin^2\theta} \\ &= \sqrt{(p_{2k} + p_{2k+1})^2 - 2p_{2k}p_{2k+1}(1-\cos\theta)} \\ &\le p_{2k} + p_{2k+1} - \frac{p_{2k}p_{2k+1}}{p_{2k}+p_{2k+1}}(1-\cos\theta). \end{aligned}$$

Hence, for $\theta \in [0,\pi]$,

$$\begin{aligned} |\varphi(\theta)| &\le 1 - (1-\cos\theta) \sum_{k=0}^{\infty} \frac{p_{2k}p_{2k+1}}{p_{2k}+p_{2k+1}} \\ &\le 1 - \frac{2}{\pi^2}\theta^2\chi^2 \\ &\le \exp(-\frac{2\chi^2}{\pi^2}\theta^2), \end{aligned}$$

which gives (2.A.1). That $\chi^2 \le \operatorname{Var} X$ can be seen by a Taylor expansion of φ in the neighborhood of $\theta = 0$, while the relation $\chi^2 \le 1/4$ follows from the trivial inequality

$$\frac{p_{2k}p_{2k+1}}{p_{2k}+p_{2k+1}} \le \frac{p_{2k}+p_{2k+1}}{4}.$$

To prove (2.A.2), remark that, when $p_{k+1} \le \rho p_k$,

$$\chi^2 \ge \frac{1}{1+\rho} \sum_{k=0}^{\infty} p_{2k+1}. \tag{2.A.3}$$

Since

$$\rho \sum_{k=0}^{\infty} p_{2k+1} \ge \sum_{k=0}^{\infty} p_{2k+2} = 1 - p_0 - \sum_{k=0}^{\infty} p_{2k+1},$$

inequality (2.A.3) becomes

$$\chi^2 \ge \frac{1}{1+\rho}\frac{1-p_0}{1+\rho},$$

and the first half of (2.A.2) is proved. The second half is proved in the same way from the inequality

$$\chi^2 \leq \frac{\rho}{1+\rho} \sum_{k=0}^{\infty} p_{2k}.$$

∎

χ has the desirable property to be zero when X is an integer variable with a span strictly greater than 1, in which case the period of φ is less than 2π. Another desirable property would be that $\chi \to \infty$ when the moments of X are unbounded; since $\chi \leq 1/4$, this is obviously not possible here. That this "feature" is somehow unavoidable can be seen on the following example:

$$\begin{aligned} \varphi(\theta) &\stackrel{\text{def}}{=} \frac{2+e^{\mathrm{i}\theta}}{4} + \frac{1}{4}\sum_{k=2}^{\infty} \frac{e^{\mathrm{i}k\theta}}{k(k-1)} \\ &= \frac{1+e^{\mathrm{i}\theta}}{2} + (1-e^{\mathrm{i}\theta})\ln(1-e^{\mathrm{i}\theta}). \end{aligned}$$

The random variable having φ as characteristic function admits no finite moment of order greater or equal to 1, but no bound on $|\varphi|$ is substantially better than (2.A.1).

Chapitre 3

Analyse de perturbations pour une file stationnaire

Une partie de l'intérêt des modèles présentés dans les deux premiers chapitres est qu'ils conduisent à des formules explicites permettant de décrire en partie le comportement du système sous-jacent. De telles formules, même si elles sont parfois complexes, donnent une mesure directe ou par approximation numérique de l'influence des différents paramètres du modèle. Cependant, on est souvent amené à considérer des modèles dont la complexité est telle que seules des méthodes de simulation permettent de les étudier.

La simulation, si elle a le mérite d'être beaucoup plus simple conceptuellement que l'analyse mathématique, pose un certain nombre de problèmes :

– l'utilisation des résultats nécessite des intervalles de confiance, qui ne sont pas faciles à obtenir quand le modèle n'a pas de propriétés d'indépendance ou de régénération ;

– les grands systèmes sont souvent difficiles à simuler, que ce soit à cause de la durée des calculs ou de la précision des résultats ;

– l'étude de l'influence des paramètres du modèle conduit à effectuer un grand nombre de simulations et devient rapidement impraticable lorsque trop de paramètres entrent en jeu.

L'« analyse de sensibilité » regroupe un certain nombre de techniques qui tentent de résoudre ce dernier problème : on se donne une partie compacte Θ de $\mathbb{R}^d$, $d \geq 1$, et on définit pour tout $\theta \in \Theta$ un espace de probabilité $(\Omega, \mathcal{F}, \mathrm{P}_\theta)$. Si L_θ est une variable aléatoire sur cet espace, on cherche une autre variable aléatoire D_θ telle que

$$\mathbb{E}_\theta D_\theta = \nabla_\theta \Big[\mathbb{E}_\theta L_\theta\Big],$$

où ∇_θ est l'opérateur gradient par rapport au vecteur θ. La variable D_θ doit bien sûr pouvoir être obtenue simplement au cours de la simulation et avoir une variance la plus petite possible. Elle permettra alors de mesurer localement la dépendance de la mesure de performance $\mathbb{E}_\theta L_\theta$ par rapport à θ. Une autre utilisation importante de ce type d'estimateur est la recherche de valeurs « optimales » de θ à l'aide d'algorithmes de gradient stochastique de type Robbins-Monro, comme le font par exemple Tang et Chen [57].

Il existe plusieurs méthodes d'analyse de sensibilité, dont les références sont citées en section 3.1. La méthode utilisée dans ce chapitre est l'« analyse de perturbations infinitésimales » –en anglais IPA pour « Infinitesimal Perturbation Analysis ». On se propose de construire un espace de probabilité indépendant de θ et on cherche des hypothèses sous lesquelles prévaut l'égalité

$$\nabla_\theta\Big[\mathbb{E}L_\theta\Big] = \mathbb{E}\Big[\nabla_\theta L_\theta\Big].$$

Les conditions nécessaires et suffisantes sous lesquelles cette relation est vraie sont en théorie connues, mais elles impliquent des propriétés d'uniforme intégrabilité inutilisables en pratique. On cherche donc des conditions suffisantes, s'exprimant simplement en fonction du modèle, sous lesquelles le théorème de convergence dominée s'applique. De telles conditions sont données pour certains modèles par Glasserman [23] ou Ho et Cao [29].

On construit ici un estimateur de type IPA pour une fonctionnelle de la charge d'une file $G/G/1/\infty$ en régime stationnaire (Théorèmes 3.3.1 et 3.4.1). L'approche traditionnellement utilisée dans le cas de tels systèmes en régime stationnaire est de construire un estimateur « à horizon fini » pour la file prise entre les temps 0 et T, puis de montrer qu'il converge lorsque $T \to \infty$. Pour une file $G/G/1/\infty$ dont les temps de service dépendent d'un paramètre réel θ, on donne des estimateurs stationnaires de la dérivée par rapport à θ de l'espérance $\mathbb{E}f(W_\theta(0))$, où $W_\theta(t)$ est la charge de la file et f une fonction à variation bornée non nécessairement continue. A cet effet, on utilise une idée de Konstantopoulos et Zazanis [35] fondée sur la formule d'inversion de Palm, mais on évite le passage à la limite en effectuant de manière globale l'analyse de type « level-crossing » grâce au théorème de Fubini. Cette méthode de preuve permet de traiter le cas où la distribution de W_θ charge des points de discontinuité de f ; nous modifions la formule donnée dans [35] en conséquence. Le cas où le paramètre est la vitesse du serveur ou le facteur d'échelle des temps d'interarrivée est également traité avec la même approche (Théorèmes 3.5.1 et 3.5.2).

Ces résultats sont issus d'un travail commun avec Pierre Brémaud et ont été publiés dans le journal *Journal of Discrete Event Dynamic Systems* [9] (voir aussi [8]).

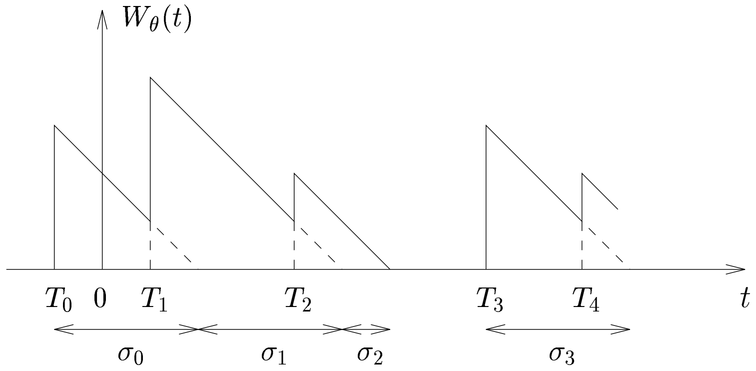


Figure 3.1: workload of a $G/G/1$ queue.

3.1 Introduction

Consider a stationary $G/G/1/\infty$ queue in which customers arrive according to a stationary process $\{T_n\}_{n\in\mathbb{Z}}$. The customer n asks for a service time $\sigma_n(\theta)$, where θ is a real parameter in the compact interval Θ and $\{\sigma_n(\theta)\}_{n\in\mathbb{Z}}$ is an i.i.d sequence. Let $\{\tau_n\}_{n\in\mathbb{Z}}$ denote the inter-arrival times process satisfying $\tau_n = T_{n+1} - T_n$. Assume that the queue is stationary and let $W_\theta(t)$ be the remaining work in the system at time t—see Figure 3.1— given by Lindley's equation

$$W_\theta(t) = \Big(W_\theta(T_n-) + \sigma_n(\theta) - (t - T_n)\Big)^+, \; t \in [T_n, T_{n+1}), \qquad (3.1.1)$$

with the notation $x^+ \stackrel{\text{def}}{=} \max(x, 0)$. Given a real function f, consider the functional $J(\theta)$ defined as

$$J(\theta) \stackrel{\text{def}}{=} \mathbb{E}f(W_\theta(0)).$$

We want to estimate, if it exists, the derivative of J with respect to θ. To this end, we use Infinitesimal Perturbation Analysis (IPA), a method first introduced by Ho and Cao [28] and further developed by Cao [11], Suri and Zazanis [55] and recently Konstantopoulos and Zazanis [35]. Glasserman [23] and Ho and Cao [29] summarize and review most previous results on IPA. Alternative methods have been used to estimate derivatives, namely Smooth Perturbation Analysis (SPA, see Suri and Zazanis [59], Gong and Ho [27], Glasserman and Gong [24], Fu and Hu [21]), Likelihood Ratio Method (LRM, see e.g. Reiman and Weiss [49] or Glynn [25]) and Rare Perturbation Analysis (RPA, see Brémaud and Vázquez-Abad [10] and Brémaud [6]).

In this chapter, we aim to prove that, under appropriate conditions

$$\frac{\partial}{\partial\theta}\mathbb{E}f(W_\theta(0)) \;=\; \lim_{h\to 0}\frac{1}{h}\Big[\mathbb{E}f(W_{\theta+h}(0)) - \mathbb{E}f(W_\theta(0))\Big]$$

$$
\begin{aligned}
&= \mathbb{E} \lim_{h\to 0} \frac{1}{h}\Big[f(W_{\theta+h}(0)) - f(W_\theta(0))\Big] \\
&= \mathbb{E}\frac{\partial}{\partial\theta} f(W_\theta(0)) \qquad (3.1.2)
\end{aligned}
$$

and we give a formula replacing (3.1.2) when f is not differentiable but is of bounded variation. This formula was obtained by Konstantopoulos and Zazanis [35] under stronger assumptions on the service times distributions. However, due to the difficulty of passing to the limit in their approximation procedure, their formula does not give any insight on the equality of the left-hand and right-hand derivatives; this information is crucial for practical use of the derivative estimator. This shortcoming has been fixed after the parution of Brémaud and Lasgouttes [8], in a short communication [36]. Our method of proof avoids the passage to the limit and therefore allows for better control of the computations. Moreover, it can be extended in many ways to handle different situations.

The chapter is organized as follows: in Section 3.2, we give a construction of the $G/G/1$ queue and we derive some basic properties. The main result of the chapter is given in Section 3.3 and the same method is applied to second-order derivatives in Section 3.4; Section 3.5 shows how our method can be extended to other parameters, respectively the speed of the server and the rate of arrival in the system. Section 3.6 discusses the implementation of the estimates and a short review of Palm probabilities can be found in the appendix.

3.2 Construction of the $G/G/1$ queue

In a formula like (3.1.2), the probability space does not depend on θ. To obtain this independence, we use the inversion representation (see Suri [54]) to generate service times: let $\{\xi_n\}_{n\in\mathbb{Z}}$ be a sequence of random variables uniformly distributed on $[0,1]$. Let $F(\cdot,\theta)$ be the common distribution function of service times; we can define its inverse function

$$G(\xi,\theta) = \sup(x \geq 0 \ : \ F(x,\theta) \leq \xi).$$

Then $\sigma_n(\theta) \stackrel{\text{def}}{=} G(\xi_n,\theta)$ is distributed according to $F(\cdot,\theta)$. This means that, if we choose as basic stationary random sequences $\{\tau_n\}_{n\in\mathbb{Z}}$ and $\{\xi_n\}_{n\in\mathbb{Z}}$, we define the queue on a probability space independent from θ. We note λ the intensity of the input process and P^0 the associated Palm probability—see Appendix for notations and details. In order to apply IPA, the following assumption on service times is needed:

Assumption A9 *The distribution of service times verifies the following conditions:*

(*i*) $\theta \mapsto G(\xi,\theta)$ *is differentiable and Lipschitz, that is*

$$|G(\xi,\theta_1) - G(\xi,\theta_2)| \leq K^\sigma(\xi)|\theta_1 - \theta_2|,\ \forall\,\theta_1,\theta_2 \in \Theta;$$

(*ii*) $\lambda \mathbb{E}^0 \sigma_0^* < 1$, *with the notation*

$$\sigma_n^* \stackrel{\text{def}}{=} \sup_{\theta\in\Theta} \sigma_n(\theta) = \sup_{\theta\in\Theta} G(\xi_n,\theta).$$

Condition **A9**-(i) ensures that we have enough smoothness with respect to θ in the distribution of the service times. However, in a number of cases, ξ_n will not be directly known, in particular when observing a real experiment; this difficulty can be overcome with the following classical proposition (Suri [54]; for this formulation see Glasserman [23]):

Proposition *Assume that*

(*i*) $F(\cdot,\theta)$ *has a density* $\partial_x F(\cdot,\theta)$ *which is strictly positive on an open interval* I_θ *and zero elsewhere;*

(*ii*) F *is continuously differentiable on* $I_\theta \times \Theta$.

Then, for all $n \in \mathbb{Z}$, $\sigma_n(\theta)$ *is differentiable and*

$$\sigma_n'(\theta) = -\frac{\partial_\theta F(\sigma_n(\theta),\theta)}{\partial_x F(\sigma_n(\theta),\theta)}.$$

In the above formula, the prime denotes the derivative with respect to θ. A case of particular interest is when θ is a scale parameter of the service times, that is when $\sigma_n(\theta) = \theta\eta_n$ for some random variable η. Then we have directly

$$\sigma_n'(\theta) = \eta_n = \frac{\sigma_n(\theta)}{\theta}.$$

In particular, we do not need to know the real distribution of service times unless we actually want to simulate them. Note that **A9**-(i) is similar to assumption (i) of Section 1 in Konstantopoulos and Zazanis [35]; it is the classical assumption on smoothed distributions needed for IPA.

Using Assumption **A9**-(ii), we derive a bound on the size of the busy periods of the system for all possible values of θ. We shall note that we don't know *a priori* whether σ_n^* is P^0-a.s. finite or not; however, this condition is weaker than assumptions (ii) and (iii) of [35]. With that in mind, let $\{R_k^*\}_{k\in\mathbb{Z}}$ be the regeneration times at which the arriving customers of the $*$-system find the queue empty—the $*$-system is the queue with service times $\{\sigma_n^*\}_{n\in\mathbb{Z}}$, whereas the θ-system uses service times $\{\sigma_n(\theta)\}_{n\in\mathbb{Z}}$. We can build

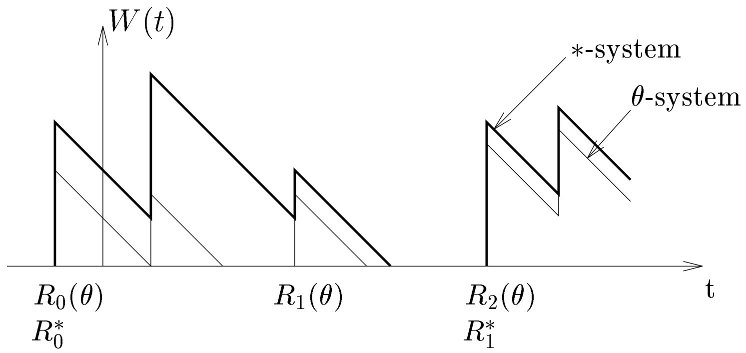


Figure 3.2: the domination property.

the θ-system from the busy period process $\{R_k^*\}_{k\in\mathbb{Z}}$ but with the service times given by $\{\sigma_n(\theta)\}_{n\in\mathbb{Z}}$, so that the following domination property holds for the respective stationary workload of the queues—see Figure 3.2:

$$W_\theta(t) \le W^*(t),\ \forall\,\theta \in \Theta,\ \forall\ t \in \mathbb{R}. \tag{3.2.1}$$

With the above construction, we get

$$\{R_k^*\}_{k\in\mathbb{Z}} \subseteq \{R_k(\theta)\}_{k\in\mathbb{Z}},$$

where $\{R_k(\theta)\}_{k\in\mathbb{Z}}$ (or simply $\{R_k\}_{k\in\mathbb{Z}}$) denotes the beginning of busy period process for the θ-system. Moreover, we have the boundary property

$$R_-^*(t) \le R_-(\theta)(t) \le t < R_+(\theta)(t) \le R_+^*(t). \tag{3.2.2}$$

3.3 Gradient estimation for non-decreasing functions

In this section, we show that IPA applies with any non-decreasing *càdlàg* function f. But since f is not required to be continuous, we cannot apply (3.1.2) as such. First of all, we need to introduce an assumption similar to assumptions A1, A2 and A3′ of Konstantopoulos and Zazanis [35]:

Assumption A10 *The following inequalities hold:*

(i) $\mathbb{E}^0[K^\sigma(\xi_0)]^4 < \infty$;

(ii) $\mathbb{E}^0[A([R_0^*, R_1^*))]^4 < \infty$;

(iii) $\mathbb{E}^0[f(W^*(0))]^2 < \infty$.

Theorem 3.3.1 *Let μ_f be the measure on $\mathbb{R}$ associated with f. Assume* **A9** *and* **A10** *hold. Then J admits a right derivative with respect to θ given by*

$$\begin{aligned} J_r'(\theta) \;=\; & \lambda \mathbb{E}^0 W_\theta'(0)\Big[f(W_\theta(0)) - f(W_\theta(T_1-)) \\ & - \mathbb{1}_{\{W_\theta'(0)<0\}}\big[\mu_f(\{W_\theta(0)\}) - \mu_f(\{W_\theta(T_1-)\})\big]\Big]. \end{aligned} \tag{3.3.1}$$

and its left derivative is

$$\begin{aligned} J_l'(\theta) &= \lambda \mathbb{E}^0 W_\theta'(0)\Big[f(W_\theta(0)) - f(W_\theta(T_1-)) \\ &\qquad - \mathbb{1}_{\{W_\theta'(0)>0\}}\big[\mu_f(\{W_\theta(0)\}) - \mu_f(\{W_\theta(T_1-)\})\big]\Big] \end{aligned} \tag{3.3.2}$$

Example 3.3.2 With $f(w) = \mathbb{1}_{\{w \geq x\}}$, Theorem 3.3.1 yields

$$\begin{aligned} \frac{\partial_r}{\partial\theta} \mathrm{P}(W_\theta(0) > x) &= \lambda \mathbb{E}^0 W_\theta'(0)\Big[\mathbb{1}_{(W_\theta(T_1-),W_\theta(0)]}(x) \\ &\qquad - \mathbb{1}_{\{W_\theta'(0)<0\}}\big[\mathbb{1}_{\{W_\theta(0)=x\}} - \mathbb{1}_{\{W_\theta(T_1-)=x\}}\big]\Big] \\ \frac{\partial_l}{\partial\theta} \mathrm{P}(W_\theta(0) > x) &= \lambda \mathbb{E}^0 W_\theta'(0)\Big[\mathbb{1}_{(W_\theta(T_1-),W_\theta(0)]}(x) \\ &\qquad - \mathbb{1}_{\{W_\theta'(0)>0\}}\big[\mathbb{1}_{\{W_\theta(0)=x\}} - \mathbb{1}_{\{W_\theta(T_1-)=x\}}\big]\Big]. \end{aligned}$$

Theorem 3.3.1 shows that $J(\theta)$ admits right and left derivatives even when f is not continuous. But in a number of cases, we can get the equality of these two derivatives:

Corollary 3.3.3 *Assume* **A9** *and* **A10** *hold. If f is continuous or if $W_\theta(0)$ and $W_\theta(T_1-)$ admit densities with respect to* P^0 *then $J(\theta)$ is differentiable and*

$$J'(\theta) = \lambda \mathbb{E}^0 W_\theta'(0)[f(W_\theta(0)) - f(W_\theta(T_1-))]. \tag{3.3.3}$$

Proof If f is continuous, then $w \mapsto \mu_f(\{w\}) \equiv 0$. If $W_\theta(0)$ admits a P^0-density, say $\gamma^0(w)$, we can use the fact that $\mu_f(\{\cdot\}) = 0$ almost everywhere for the Lebesgue measure:

$$\begin{aligned} |\mathbb{E}^0 \mathbb{1}_{\{W_\theta'(0)<0\}} \mu_f(\{W_\theta(0)\})| &\leq \mathbb{E}^0 \mu_f(\{W_\theta(0)\}) \\ &= \int_0^\infty \mu_f(\{w\}) \gamma^0(w)\, dw = 0. \end{aligned}$$

In either case, the result is proved. ■

Remark In the case where f admits a derivative f', we can use the inversion formula (3.A.2) of Appendix and write (3.3.3) as

$$J'(\theta) = \lambda \mathbb{E}^0 [\int_0^{T_1} W_\theta'(t) f'(W_\theta(t)) dt] = \mathbb{E}[W_\theta'(0) f'(W_\theta(0))],$$

thus obtaining the expected IPA estimate (3.1.2). In this computation, we used the fact that $W_\theta'(t)$ is constant between arrivals during busy periods, and zero during idle periods. A comparison between the two estimates is made in Section 3.6.

Before starting the proof of the theorem, let us mention that our derivation is different from Konstantopoulos and Zazanis [35] in two respects: first we do not require an approximation procedure and we treat directly a non decreasing function f. This is made possible by the simple crucial observation that

$$f(y) - f(x) = \int_{(x,y]} \mu_f(dz) \quad \text{for all } x \leq y,$$

which allows us to have a better view of the residual terms in the level crossing analysis that follows. The result can be applied to any function of bounded variation if assumption **A10** is verified by both the increasing and decreasing parts of the function. Secondly, we do not need switch back and forth between the Palm probabilities with respect to the arrival process and with respect to the regeneration points as in [35]. However, we retain the fundamental idea of [35] by starting with its expression in terms of the Palm probability P^0.

Proof of Theorem 3.3.1 Assume that $f(0) = 0$, so that f is non-negative. The Palm inversion formula (3.A.2) gives

$$\begin{aligned}
\mathbb{E} f(W_\theta(0)) &= \lambda \mathbb{E}^0 \int_0^{T_1} f(W_\theta(t))\, dt \\
&= \lambda \mathbb{E}^0 \int_0^{T_1} \int_{\mathbb{R}_+} \mathbb{1}_{\{W_\theta(t) > x\}}\, \mu_f(dx)\, dt \\
&= \lambda \mathbb{E}^0 \int_{\mathbb{R}_+} \int_0^{T_1} \mathbb{1}_{\{W_\theta(t) > x\}}\, dt\, \mu_f(dx)
\end{aligned}$$

and therefore

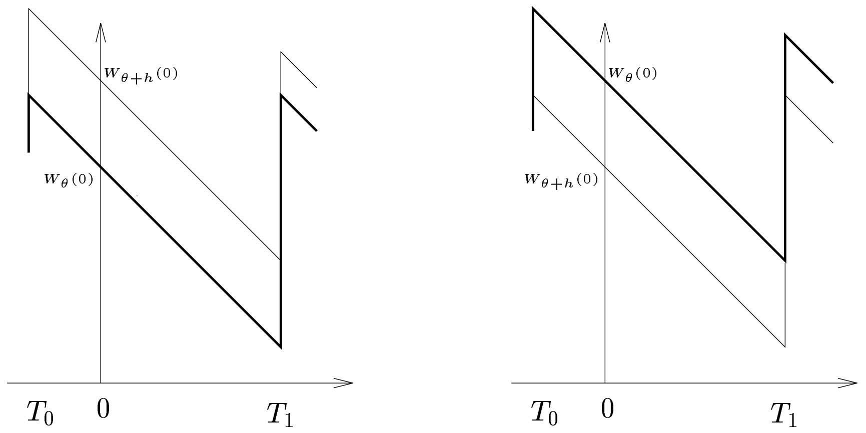
$$\begin{aligned}
&\frac{1}{h} \mathbb{E}[f(W_{\theta+h}(0)) - f(W_\theta(0))] \\
&\quad = \frac{\lambda}{h} \mathbb{E}^0 \int_{\mathbb{R}_+} \int_0^{T_1} [\mathbb{1}_{\{W_{\theta+h}(t) > x\}} - \mathbb{1}_{\{W_\theta(t) > x\}}]\, dt\, \mu_f(dx).
\end{aligned}$$

In order to simplify the notations, let:

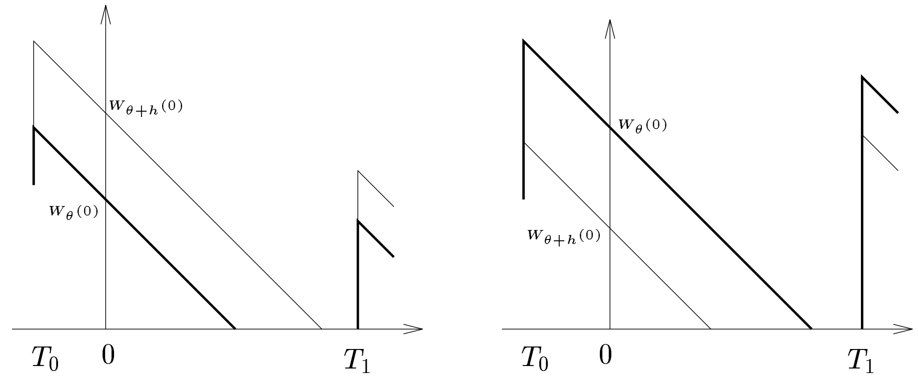
$$\begin{aligned}
\varphi(x,t) &\stackrel{\text{def}}{=} \mathbb{1}_{\{W_{\theta+h}(t) > x\}} - \mathbb{1}_{\{W_\theta(t) > x\}} \\
\Phi(\theta,h) &\stackrel{\text{def}}{=} \int_{\mathbb{R}_+} \int_0^{T_1} \varphi(x,t)\, dt\, \mu_f(dx).
\end{aligned}$$

The first step of our proof is to compute $\lim_{h\to 0} \Phi(\theta,h)/h$. We will have to integrate a function taking its values in $\{-1, 0, 1\}$ with respect to $dt\, \mu_f(dx)$. Define also for any $t \in [0, T_1)$:

$$\Delta W_{\theta,h} \stackrel{\text{def}}{=} W_{\theta+h}(t) - W_\theta(t).$$



Case 1: $W'(0) > 0$ and $W(T_1-) > 0$ Case 2: $W'(0) < 0$ and $W(T_1-) > 0$



Case 1′: $W'(0) > 0$ and $W(T_1-) = 0$ Case 2′: $W'(0) < 0$ and $W(T_1-) = 0$

Figure 3.3: Four different cases for the computation of Φ.

Assume first that $h > 0$. As shown in Figure 3.3, we must consider different cases depending on the relative position of $W_\theta(0)$ and $W_{\theta+h}(0)$. We have to add cases 3 and $3'$, where $W'_\theta(0) = 0$, preventing us to guess their relative positions. In fact, all the terms of the formula can be found in the first two cases and we will leave the other ones to the reader's attention.

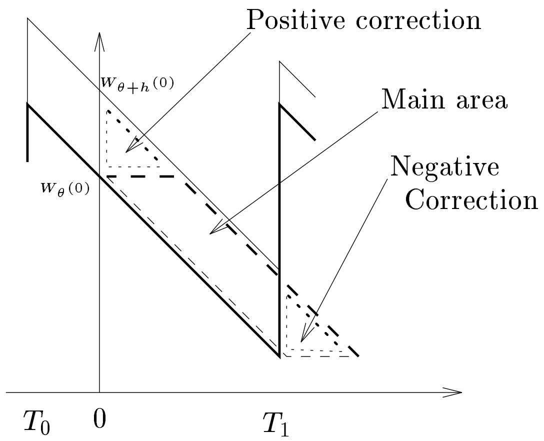


Figure 3.4: computation of Φ in case 1.

Case 1: for h small enough, $W_{\theta+h}(0) > W_\theta(0)$ and $\varphi = 1$. The way to compute $\Phi(\theta, h)$ can be best understood with the help of Figure 3.4. Φ is equal to the area with a dashed border plus the dotted triangle on the left, minus the right one. Here the borders included in the areas are in bold; since all functions are *càdlàg*, these borders are the top and right ones.

$$\begin{aligned}
\frac{1}{h}\Phi(\theta,h) &= \Big[f(W_\theta(0)) - f(W_\theta(T_1-))\Big]\frac{\Delta W_{\theta,h}}{h} \\
&\quad + \frac{1}{h}\int_0^{\Delta W_{\theta,h}} \mu_f((W_\theta(0), W_\theta(0)+y])\,dy \\
&\quad - \frac{1}{h}\int_0^{\Delta W_{\theta,h}} \mu_f((W_\theta(T_1-), W_\theta(T_1-)+y])\,dy. \quad (3.3.4)
\end{aligned}$$

The first term converges to $[f(W_\theta(0)) - f(W_\theta(T_1-))]W'_\theta(0)$. Moreover,

$$\mu_f((W_\theta(0), W_\theta(0)+y]) = \mu_f((W_\theta(0), W_\theta(0)+y)) + \mu_f(\{W_\theta(0)+y\})$$

and since $\mu_f(\{W_\theta(0)+y\}) = 0$ dy-a.e., the second term of r.h.s. of equation (3.3.4) reads

$$\frac{1}{h}\int_0^{\Delta W_{\theta,h}} \mu_f((W_\theta(0), W_\theta(0)+y))\,dy,$$

which is less or equal than

$$(W'_\theta(0) + o(1))\,\mu_f((W_\theta(0), W_{\theta+h}(0))).$$

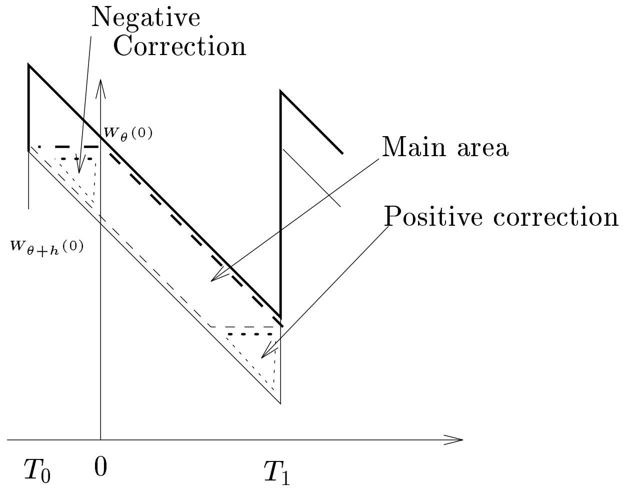


Figure 3.5: computation of Φ in case 2.

Since $W_\theta(0)$ is continuous in the neighborhood of θ, this goes to zero with h. The third term converges to 0 for the same reasons. So we have in this case:

$$\lim_{h\to 0+} \frac{1}{h}\Phi(\theta,h) = [f(W_\theta(0)) - f(W_\theta(T_1-))]W'_\theta(0).$$

Case 2: here $W_{\theta+h}(0) < W_\theta(0)$ and $\varphi = -1$. Due to the order of $W_{\theta+h}(0)$ and $W_\theta(0)$, we find a formula different from equation (3.3.4)—see Figure 3.5:

$$\begin{aligned}
\frac{1}{h}\Phi(\theta,h) &= -\Big\{\Big[f(W_\theta(0)) - f(W_\theta(T_1-))\Big]\frac{-\Delta W_{\theta,h}}{h} \\
&\quad - \frac{1}{h}\int_0^{-\Delta W_{\theta,h}} \mu_f((W_\theta(0)-y, W_\theta(0)])\,dy \\
&\quad + \frac{1}{h}\int_0^{-\Delta W_{\theta,h}} \mu_f((W_\theta(T_1-)-y, W_\theta(T_1-)])\,dy\Big\}. \qquad (3.3.5)
\end{aligned}$$

The first term is the same as in case 1, but the second is equal to

$$\frac{1}{h}\int_0^{-\Delta W_{\theta,h}} \mu_f((W_\theta(0)-y, W_\theta(0)))\,dy \; - \; \mu_f(\{W_\theta(0)\})\frac{\Delta W_{\theta,h}}{h}$$

and its limit is $\mu_f(\{W_\theta(0)\})W'_\theta(0)$. The last term of $\Phi(\theta,h)/h$ is computed in a similar way. Finally:

$$\begin{aligned}
\lim_{h\to 0+}\frac{1}{h}\Phi(\theta,h) &= \Big[f(W_\theta(0)) - f(W_\theta(T_1-)) \\
&\qquad - \mu_f(\{W_\theta(0)\}) + \mu_f(\{W_\theta(T_1-)\})\Big]W'_\theta(0).
\end{aligned}$$

We can summarize the above cases in the following formula:

$$\begin{aligned}\lim_{h\to 0+}\frac{1}{h}\Phi(\theta,h) &= W'_\theta(0)\Big[f(W_\theta(0)) - f(W_\theta(T_1-))\\ &\quad - 1\!\!1_{\{W'_\theta(0)<0\}}[\mu_f(\{W_\theta(0)\}) - \mu_f(\{W_\theta(T_1-)\})]\Big].\end{aligned}$$

The next step is to find a bound for $\Phi(\theta,h)/h$ which has a finite mean with respect to P^0. The formulas for each case give:

$$\begin{aligned}\Big|\frac{1}{h}\Phi(\theta,h)\Big| &\le \Big(f(W_\theta(0)) - f(W_\theta(T_1-))\Big)\Big|\frac{\Delta W_{\theta,h}}{h}\Big|\\ &\quad + |f(W_{\theta+h}(0)) - f(W_\theta(0))|\cdot\Big|\frac{\Delta W_{\theta,h}}{h}\Big|\\ &\quad + |f(W_{\theta+h}(T_1-)) - f(W_\theta(T_1-))|\cdot\Big|\frac{\Delta W_{\theta,h}}{h}\Big|\\ &\le 3f(W^*(0))K^W_\theta(0).\end{aligned}$$

The last inequality takes advantage of the fact that f is non-decreasing and of the domination property (3.2.1). $K^W_\theta(t)$ is a Lipschitz coefficient for $W(t)$ w.r.t. θ. Finally,

$$\Big|\frac{1}{h}\Phi(\theta,h)\Big| \le 3f(W^*(0))K^W_\theta(0).$$

The latter expression is independent from h. Moreover, it has a finite mean under P^0: from Cauchy-Schwartz inequality,

$$\mathbb{E}^0\Big[f(W^*(0))K^W_\theta(0)\Big] \le \sqrt{\mathbb{E}^0[f(W^*(0))]^2}\sqrt{\mathbb{E}^0[K^W_\theta(0)]^2}.$$

The first mean is finite from assumption **A10**-(iii). To prove that the second one is also finite, we must first give an expression of $K^W_\theta(0)$:

$$\begin{aligned}\Big|\frac{W_{\theta+h}(0) - W_\theta(0)}{h}\Big| &\le \Big|\frac{W_{\theta+h}(T_{-1}) - W_\theta(T_{-1})}{h}\Big| + \Big|\frac{\sigma_0(\theta+h) - \sigma_0(\theta)}{h}\Big|\\ &\le \sum_{n\in\mathbb{Z}}\Big|\frac{\sigma_n(\theta+h) - \sigma_n(\theta)}{h}\Big| 1\!\!1_{[R^*_-(T_0),0)}(T_n)\\ &\le \sum_{n\in\mathbb{Z}} K^\sigma(\xi_n) 1\!\!1_{[R^*_-(0),R^*_+(0))}(T_n)\\ &\stackrel{\text{def}}{=} K^W_\theta(0).\end{aligned}$$

The first inequality comes from equation (3.1.1) and inequality $|a^+ - b^+| \le |a-b|$; then we use the boundary property (3.2.2) and last the Lipschitz property **A9**-(i). To prove that $\mathbb{E}^0[K^W_\theta(0)]^2$ is finite, we can use the inequality $(x_1+\cdots+x_n)^p \le n^{p-1}(x_1^p+\cdots+x_n^p)$ and

$$\mathbb{E}^0\Big[\sum_{n\in\mathbb{Z}} K^\sigma(\xi_n) 1\!\!1_{[R^*_-(0),R^*_+(0))}(T_n)\Big]^2$$

$$
\begin{aligned}
&\leq \mathbb{E}^0\Big[\sum_{n\in\mathbb{Z}} A([R_0^*, R_1^*))[K^\sigma(\xi_n)]^2 \mathbb{1}_{[R_-^*(0), R_+^*(0))}(T_n)\Big] \\
&\leq \mathbb{E}^0[A([R_0^*, R_1^*))]^2[K^\sigma(\xi_0)]^2 \\
&\leq \sqrt{\mathbb{E}^0[A([R_0^*, R_1^*))]^4}\sqrt{\mathbb{E}^0[K^\sigma(\xi_0)]^4},
\end{aligned}
$$

which is finite from **A10**-(i) and **A10**-(ii) . Here, the second inequality uses Lemma 3.A.1.

Summing up our results, we can apply the Dominated Convergence Theorem:

$$
\begin{aligned}
J_r'(\theta) &\stackrel{\text{def}}{=} \lim_{h\to 0+} \mathbb{E}[f(W_{\theta+h}(0)) - f(W_\theta(0))] \\
&= \lim_{h\to 0+} \lambda \mathbb{E}^0 \frac{1}{h}\Phi(\theta, h) \\
&= \lambda \mathbb{E}^0 \lim_{h\to 0+} \frac{1}{h}\Phi(\theta, h)
\end{aligned}
$$

This gives equation (3.3.1). The case of $h < 0$ is handled in the same way and gives equation (3.3.2)—loosely speaking, the above cases used the sign of $W_{\theta+h}(0) - W_\theta(0)$; this sign is inverted if $h < 0$. This concludes the proof of the theorem. ■

Remark Assumption **A10** ensures that $\mathbb{E}^0[f(W^*(0))K_\theta^W(0)] < \infty$. If we know that f is bounded, for example, the only assumptions we need are

(i) $\mathbb{E}^0[K^\sigma(\xi_0)]^2 < \infty$;

(ii) $\mathbb{E}^0[A([R_0^*, R_1^*))]^2 < \infty$.

This reduced set of assumptions can for instance be used in Example 3.3.2.

It is important to point out that Corollary 3.3.3 cannot always be applied. We show such a case in next example :

Example 3.3.4 Consider a $D/D/1$ queue, that is with deterministic interarrival time τ and service time $\theta < \tau$. In order to have a stationary queue, T_0 must be uniformly spread in $[-\tau, 0]$. As we can see in Figure 3.6, we have

$$
W(T_n) = \theta,\ W(T_n-) = 0,\ W_\theta'(0) = 1\ \ \mathrm{P}^0\text{-a.s.}
$$

For $x > 0$, take $f(w) = \mathbb{1}_{\{w\geq x\}}$ as in Example 3.3.2. Then if $\theta \leq x$, $J(\theta) = 0$, else

$$
J(\theta) = \int_{-\tau}^{0} \mathbb{1}_{\{\theta+t\geq x\}} \frac{dt}{\tau} = \frac{\theta - x}{\tau}.
$$

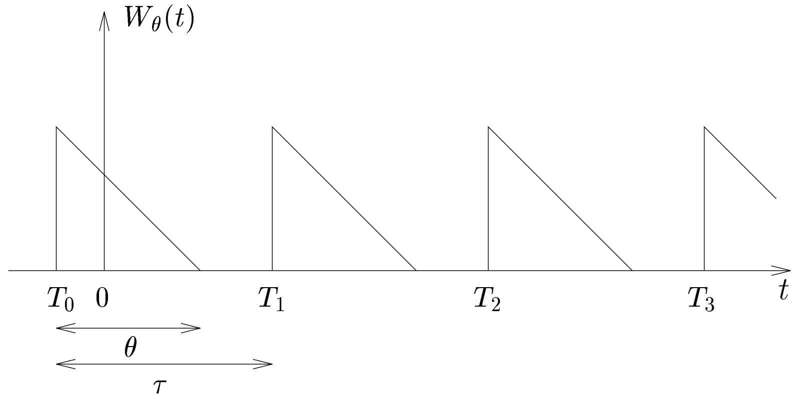


Figure 3.6: workload of the $D/D/1$ queue.

Finally $J(\theta) \equiv \mathrm{P}(W_\theta(0) \geq x) = \left(\frac{\theta - x}{\tau}\right)^+$, which is not differentiable at point $\theta = x$. Besides,

$$\begin{aligned} J_r'(\theta) &= \lambda \mathbb{E}^0[\mathbb{1}_{\{\theta \geq x\}} - \mathbb{1}_{\{0 \geq x\}}] \\ &= \frac{1}{\tau}\mathbb{1}_{\{\theta \geq x\}} \\ J_l'(\theta) &= \lambda \mathbb{E}^0[\mathbb{1}_{\{\theta \geq x\}} - \mathbb{1}_{\{0 \geq x\}} + \mathbb{1}_{\{0 = x\}} - \mathbb{1}_{\{\theta = x\}}] \\ &= \frac{1}{\tau}\mathbb{1}_{\{\theta > x\}} \end{aligned}$$

3.4 Second order derivative

The method used in Section 3.3 can be used for higher-order derivatives. We need assumptions on the properties of our system and some new moment conditions:

Assumption A11 *G and f verify the following:*

(*i*) *$\theta \mapsto G(\xi, \theta)$ is twice differentiable and there exists a function $\xi \mapsto K^{\sigma'}(\xi)$ such that*

$$|G(\xi, \theta + 2h) - 2G(\xi, \theta + h) + G(\xi, \theta)| \leq h^2 K^{\sigma'}(\xi);$$

(*ii*) *$w \mapsto f(w)$ is non-decreasing and differentiable.*

Assumption A12 *The following inequalities hold:*

(*i*) $\mathbb{E}^0[K^\sigma(\xi_0)]^8 < \infty$;

(*ii*) $\mathbb{E}^0[K^{\sigma'}(\xi_0)]^4 < \infty$;

(*iii*) $\mathbb{E}^0[A([R_0^*, R_1^*))]^8 < \infty$;

(*iv*) $\mathbb{E}^0[f(W^*(0))]^2 < \infty$;

(*v*) $\mathbb{E}^0[\sup_\theta f'(W_\theta(0))]^2 < \infty$;

(*vi*) $\mathbb{E}^0[\sup_\theta f'(W_\theta(T_1-))]^2 < \infty$.

The main result of this section is:

Theorem 3.4.1 *Assume* **A9**, **A11** *and* **A12** *hold; then J admits a right second derivative with respect to θ given by*

$$\begin{aligned} J_r''(\theta) &= \lambda \mathbb{E}^0 \Big[W_\theta''(0) \Big[f(W_\theta(0)) - f(W_\theta(T_1-)) \Big] \\ &\quad + [W'(0)]^2 \Big[f'(W_\theta(0)) - f'(W_\theta(T_1-)) \\ &\quad - \mathbb{1}_{\{W_\theta'(0)<0\}} [\mu_{f'}(\{W_\theta(0)\}) - \mu_{f'}(\{W_\theta(T_1-)\})] \Big] \Big], \end{aligned} \tag{3.4.1}$$

and its left second derivative is

$$\begin{aligned} J_l''(\theta) &= \lambda \mathbb{E}^0 \Big[W_\theta''(0) \Big[f(W_\theta(0)) - f(W_\theta(T_1-)) \Big] \\ &\quad + [W'(0)]^2 \Big[f'(W_\theta(0)) - f'(W_\theta(T_1-)) \\ &\quad - \mathbb{1}_{\{W_\theta'(0)>0\}} [\mu_{f'}(\{W_\theta(0)\}) - \mu_{f'}(\{W_\theta(T_1-)\})] \Big] \Big]. \end{aligned} \tag{3.4.2}$$

Corollary 3.4.2 *Assume* **A9**, **A11** *and* **A12** *hold; if f' is continuous or if $W_\theta(0)$ and $W_\theta(T_1-)$ admit densities w.r.t.* P^0 *then $J(\theta)$ is differentiable twice and*

$$\begin{aligned} J''(\theta) &= \lambda \mathbb{E}^0 \Big[W_\theta''(0) \Big[f(W_\theta(0)) - f(W_\theta(T_1-)) \Big] \\ &\quad + [W'(0)]^2 \Big[f'(W_\theta(0)) - f'(W_\theta(T_1-)) \Big] \Big]. \end{aligned}$$

Proof of Theorem 3.4.1 As this proof is very similar to that of Theorem 3.3.1, we will omit the parts of it which are not new. We want to compute the limit as $h \to 0$ of

$$\frac{1}{h^2} \mathbb{E}\Big[f(W_{\theta+2h}(0)) - 2f(W_{\theta+h}(0)) + f(W_\theta(0)) \Big] = \frac{\lambda}{h^2} \mathbb{E}^0 \Phi_2(\theta, h),$$

with

$$\begin{aligned} \Phi_2(\theta, h) &\stackrel{\mathrm{def}}{=} \int_{\mathbb{R}_+} \int_0^{T_1} \Big[[\mathbb{1}_{\{W_{\theta+2h}(t)>x\}} - \mathbb{1}_{\{W_{\theta+h}(t)>x\}}] \\ &\quad - [\mathbb{1}_{\{W_{\theta+h}(t)>x\}} - \mathbb{1}_{\{W_\theta(t)>x\}}] \Big] dt \, \mu_f(dx). \end{aligned}$$

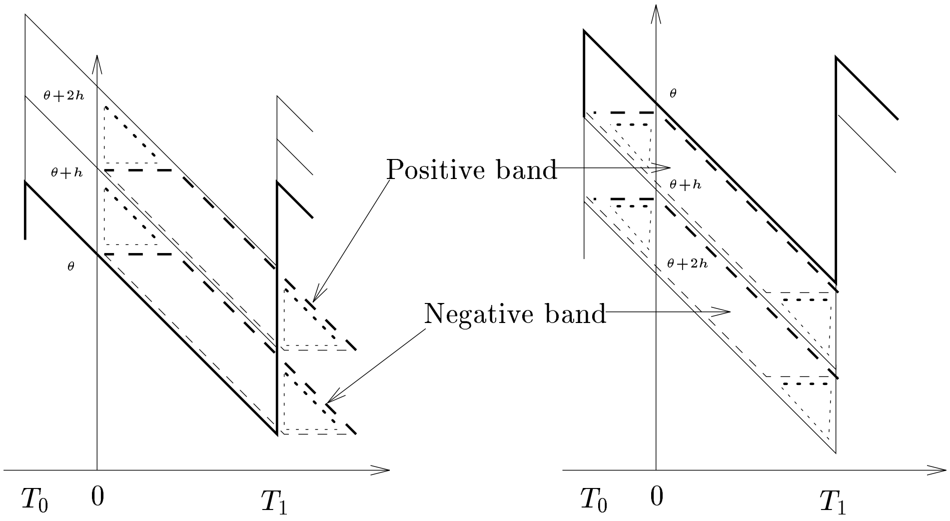


Figure 3.7: computation of Φ_2 in cases 1 and 2

We will once more distinguish two important cases among all possible ones, depending on the sign of $W'_\theta(0)$. Suppose first that $h > 0$.

Case 1: $W'_\theta(0) > 0$; for h small enough, $W_\theta(0) < W_{\theta+h}(0) < W_{\theta+2h}(0)$—see Figure 3.7. We have here to subtract the areas of two bands which are of the same sort as in Theorem 3.3.1:

$$\begin{aligned}\Phi_2(\theta,h) &= \Delta W_{\theta+h,h}\Big[f(W_{\theta+h}(0)) - f(W_{\theta+h}(T_1-))\Big] \\ &\quad - \Delta W_{\theta,h}\Big[f(W_\theta(0)) - f(W_\theta(T_1-))\Big] \\ &\quad + A_{\theta,h}(0) - A_{\theta,h}(T_1-)\end{aligned}$$

where

$$\begin{aligned}A_{\theta,h}(t) &= \int_0^{\Delta W_{\theta+h,h}} \mu_f((W_{\theta+h}(t), W_{\theta+h}(t)+y])\,dy \\ &\quad - \int_0^{\Delta W_{\theta,h}} \mu_f((W_\theta(t), W_\theta(t)+y])\,dy.\end{aligned}$$

The main term is equal to

$$\begin{aligned}&\Delta W_{\theta+h,h}\Big[f(W_{\theta+h}(0)) - f(W_{\theta+h}(T_1-)) - f(W_\theta(0)) + f(W_\theta(T_1-))\Big] \\ &+ \Delta^2 W_{\theta,h}\Big[f(W_\theta(0)) - f(W_\theta(T_1-))\Big].\end{aligned}$$

Moreover,

$$A_{\theta,h}(0) = \int_0^{\Delta W_{\theta,h}} \Big[f(W_{\theta+h}(0)+y) - f(W_{\theta+h}(0))$$

$$
\begin{aligned}
&\quad - f(W_\theta(0)+y) + f(W_\theta(0))\Big] dy + o(h^2) \\
&= \int_0^{\Delta W_{\theta,h}} h W_\theta'(0) \mu_{f'}((W_\theta(0), W_\theta(0)+y]) dy + o(h^2)
\end{aligned}
$$

As in Theorem 3.3.1-case 1, $\lim_{h\to 0} A_{\theta,h}(0)/h^2 = 0$; the limit is the same for $A_{\theta,h}(T_1-)$. Consequently,

$$
\begin{aligned}
\lim_{h\to 0+} \frac{1}{h^2}\Phi_2(\theta,h) &= [W_\theta'(0)]^2 \Big[f'(W_\theta(0)) - f'(W_\theta(T_1-))\Big] \\
&\quad + W_\theta''(0)\Big[f(W_\theta(0)) - f(W_\theta(T_1-))\Big].
\end{aligned}
$$

Case 2: $W_\theta'(0) < 0$; for h small enough, $W_\theta(0) > W_{\theta+h}(0) > W_{\theta+2h}(0)$ and

$$
\begin{aligned}
\Phi_2(\theta,h) &= -1 \cdot \Big\{ -\Delta W_{\theta+h,h}\Big[f(W_{\theta+h}(0)) - f(W_{\theta+h}(T_1-))\Big] \\
&\quad - \Delta W_{\theta,h}\Big[f(W_\theta(0)) - f(W_\theta(T_1-))\Big] \\
&\quad - B_{\theta,h}(0) + B_{\theta,h}(T_1-)\Big\}
\end{aligned}
$$

where

$$
\begin{aligned}
B_{\theta,h}(t) &\stackrel{\text{def}}{=} \int_0^{-\Delta W_{\theta+h,h}} \mu_f((W_{\theta+h}(t)-y, W_{\theta+h}(t)])\, dy \\
&\quad - \int_0^{-\Delta W_{\theta,h}} \mu_f((W_\theta(t)-y, W_\theta(t)])\, dy.
\end{aligned}
$$

While the main part has the same limit as in case 1, we have

$$
\begin{aligned}
B_{\theta,h}(0) &= \int_0^{-\Delta W_{\theta,h}} h W_\theta'(0)\mu_{f'}((W_\theta(0)-y, W_\theta(0)])dy + o(h^2) \\
&= -h^2 [W_\theta'(0)]^2 \mu_{f'}(\{W_\theta(0)\}) + o(h^2).
\end{aligned}
$$

Finally, in case 2,

$$
\begin{aligned}
\lim_{h\to 0+} \frac{1}{h^2}\Phi_2(\theta,h) &= [W_\theta'(0)]^2 \Big[f'(W_\theta(0)) - f'(W_\theta(T_1-))\Big] \\
&\quad + W_\theta''(0)\Big[f(W_\theta(0)) - f(W_\theta(T_1-))\Big] \\
&\quad - [W_\theta'(0)]^2 \Big[\mu_{f'}(\{W_\theta(0)\}) - \mu_{f'}(\{W_\theta(T_1-)\})\Big].
\end{aligned}
$$

Besides,

$$
\begin{aligned}
\left|\frac{1}{h^2}\Phi_2(\theta,h)\right| &\le 3\Big[\sup_\theta f'(W_\theta(0)) + \sup_\theta f'(W_\theta(T_1-))\Big]\Big[K_\theta^W(0)\Big]^2 \\
&\quad + f(W^*(0)) K_\theta^{W'}(0),
\end{aligned}
$$

where $K_\theta^W(0)$ is the same as in Theorem 3.3.1 and

$$K_\theta^{W'}(0) \stackrel{\text{def}}{=} \sum_{n\in\mathbb{Z}} K^{\sigma'}(\xi_n) 1\!\!1_{[R_-^*(0),R_+^*(0))}(T_n).$$

As in theorem 3.3.1, we use Lemma 3.A.1, Cauchy-Schwartz inequality and assumption **A12** to prove that $|\Phi_2(\theta,h)/h^2|$ has a finite mean under P^0. Using the Dominated Convergence Theorem, we find expressions (3.4.1) and (3.4.2) for the second derivatives of J. ∎

3.5 Other parameters of the queue

Let us consider a setting slightly different from the original one: we still deal with a $G/G/1$ queue, but now working at speed ν. Lindley's equation for the workload of the queue reads:

$$W_\nu(t) = \Big(W_\nu(T_n-) + \sigma_n - \nu(t-T_n)\Big)^+,\ t\in[T_n,T_{n+1}).$$

We address the same problem as in Section 3.3 in this new setting. Our method can apply in this case in the same way as for variable service times; we will try to keep the notations as close as possible to those of Section 3.2 to point out the similitudes, replacing θ with ν when necessary.

Remark As an anonymous reviewer pointed out, if we define $\bar{W}_\nu(t) = \nu W_\nu(t)$ we have the relation:

$$\bar{W}_\nu(t) = \Big(\bar{W}_\nu(T_n-) + \frac{\sigma_n}{\nu} - (t-T_n)\Big)^+,\ t\in[T_n,T_{n+1}).$$

This means that the queue with workload $\bar{W}_\nu$ fits in the framework of Sections 3.2 and 3.3. Nevertheless, what we want to estimate is $(\partial/\partial\nu)\mathbb{E}f(\nu\bar{W}_\nu(0))$, which does not follow directly from Theorem 3.3.1. The extra computations needed would cancel the gain of using Theorem 3.3.1. Note also that this result will be useful in the second part of this section to deal with parameters of the arrival process.

Assume that $\nu \geq \nu^* > 0$; then as in Section 3.2, we can construct all the queues for different values of ν so that for all $\nu \geq \nu^*$ and $t\in\mathbb{R}$, we have the relation (see Figure 3.8)

$$W_\nu(t) \leq W_{\nu^*}(t) \tag{3.5.1}$$

$$R_-^*(t) \leq R_-(\nu)(t) \leq t < R_+(\nu)(t) \leq R_+^*(t), \tag{3.5.2}$$

The assumption on moments we need is much like **A10**:

Assumption A13 *The following moments are finite:*

(*i*) $\mathbb{E}^0[\tau_0]^4 < \infty$;

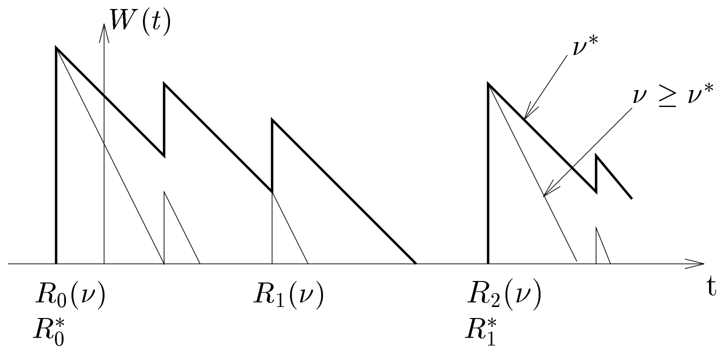


Figure 3.8: the domination property for the speed

(*ii*) $\mathbb{E}^0[A([R_0^*, R_1^*))]^4 < \infty$;

(*iii*) $\mathbb{E}^0[f(W_{\nu^*}(0))]^2 < \infty$.

The first real difference with the results of Section 3.3 is that the expressions for the derivative use a primitive of f, whereas only f appeared in Theorem 3.3.1.

Theorem 3.5.1 *Let F be a primitive of f; if* **A13** *holds, then J has a right-hand derivative equal to:*

$$\begin{aligned} J'_r(\nu) &= \frac{\lambda}{\nu^2}\mathbb{E}^0\Big\{\nu W'_\nu(0)\Big[f(W_\nu(0)) - f(W_\nu(T_1-))\Big] \\ &\quad + F(W_\nu(0)) - F(W_\nu(T_1-)) \\ &\quad - [W_\nu(0) - W_\nu(T_1-)]f(W_\nu(T_1-))\Big\} \end{aligned} \tag{3.5.3}$$

and its left-hand derivative is

$$\begin{aligned} J'_l(\nu) &= \frac{\lambda}{\nu^2}\mathbb{E}^0\Big\{\nu W'_\nu(0)\Big[f(W_\nu(0)) - f(W_\nu(T_1-))\Big] \\ &\quad + F(W_\nu(0)) - F(W_\nu(T_1-)) \\ &\quad - [W_\nu(0) - W_\nu(T_1-)]f(W_\nu(T_1-)) \\ &\quad + \nu W'_\nu(0)[\mu_f(\{W_\nu(0)\}) - \mu_f(\{W_\nu(T_1-)\})] \\ &\quad - [W_\nu(0) - W_\nu(T_1-)]\mu_f(\{W_\nu(T_1-)\})\Big\}. \end{aligned}$$

If f is continuous or if both $W_\nu(0)$ and $W_\nu(T_1-)$ admit densities w.r.t. P^0*, then J is differentiable and its derivative is equal to J'_r.*

Remark The expressions in Theorem 3.5.1 seem really complicated when compared to those obtained in Theorem 3.3.1; in fact, in the case where f is

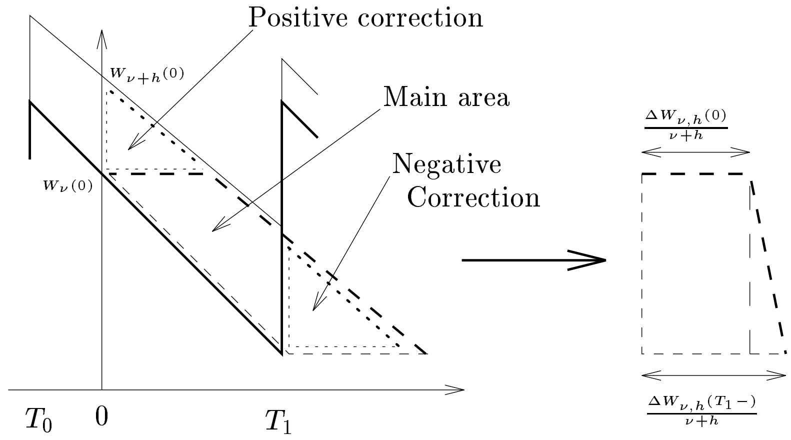


Figure 3.9: computation of Φ for $h > 0$.

differentiable, the inversion formula applied to (3.5.3) gives the classic IPA formula

$$J'(\nu) = \mathbb{E}W'_\nu(0) f(W_\nu(0)).$$

The complexity of (3.5.3) comes from the fact that $W'_\nu(t)$ is not constant on $[T_0, T_1)$.

Proof of Theorem 3.5.1 We once more proceed as in the proof of Theorem 3.3.1—more details can be found in [8]. Define

$$\Phi(\nu, h) \stackrel{\text{def}}{=} \int_{\mathbb{R}_+} \int_0^{T_1} [\mathbb{1}_{\{W_{\nu+h}(t)>x\}} - \mathbb{1}_{\{W_\nu(t)>x\}}]\, dt\, \mu_f(dx)$$

and remark that

$$\frac{1}{h}\mathbb{E}[f(W_{\nu+h}(0)) - f(W_\nu(0))] = \frac{\lambda}{h}\mathbb{E}^0\Phi(\nu, h).$$

We will consider only right-hand derivatives; left-hand derivatives are obtained with the same method. Figure 3.9 shows how Φ can be computed: the main area is equal to the area of the trapezium on the right. As W_ν is linear in ν, we have

$$\Delta W_{\nu,h}(T_1-) - \Delta W_{\nu,h}(0) = hT'_1,$$

where

$$T'_1 \stackrel{\text{def}}{=} \min\Big[\frac{W_\nu(0)}{\nu}, T_1\Big] = \frac{W_\nu(0) - W_\nu(T_1-)}{\nu}.$$

The area of the trapezium of Figure 3.9 is equal to

$$\mathcal{A} \;=\; \frac{\Delta W_{\nu,h}(0)}{\nu + h}\Big[f(W_\nu(0)) - f(W_\nu(T_1-))\Big]$$

$$
\begin{aligned}
&+ \int_0^{\frac{hT_1'}{\nu+h}} \mu_f\Big(\Big(W_\nu(T_1-), W_\nu(0) - \frac{\nu(\nu+h)}{h} y\Big]\Big)\, dy \\
= \;& \frac{h}{\nu+h}\Big\{ \frac{\Delta W_{\nu,h}(0)}{h}\Big[f(W_\nu(0)) - f(W_\nu(T_1-))\Big] \\
&+ \frac{1}{\nu}\Big[F(W_\nu(0)) - F(W_\nu(T_1-))\Big] \\
&- \frac{W_\nu(0) - W_\nu(T_1-)}{\nu} f(W_\nu(T_1-))\Big\}.
\end{aligned}
$$

The additional terms read:

$$
\begin{aligned}
&\int_0^{\frac{\Delta W_{\nu,h}(0)}{\nu+h}} \mu_f((W_\nu(0), W_\nu(0) + (\nu+h)y])\, dy \\
-&\int_0^{\frac{\Delta W_{\nu,h}(T_1-)}{\nu+h}} \mu_f((W_\nu(T_1-), W_\nu(T_1-) + (\nu+h)y])\, dy.
\end{aligned}
$$

As we have shown in the proof of Theorem 3.3.1, this kind of expression is an $o(h)$ and

$$
\begin{aligned}
\lim_{h\to 0+} \Phi(\nu, h) \;=\; & \frac{W_\nu'(0)}{\nu}\Big[f(W_\nu(0)) - f(W_\nu(T_1-))\Big] \\
&+ \frac{1}{\nu^2}\Big[F(W_\nu(0)) - F(W_\nu(T_1-))\Big] \\
&- \frac{W_\nu(0) - W_\nu(T_1-)}{\nu} f(W_\nu(T_1-)).
\end{aligned}
$$

Moreover, as in Theorem 3.3.1, we have

$$
\Big|\frac{1}{h}\Phi(\nu, h)\Big| \le \frac{1}{\nu}[3K_\nu^W(0) + 2\tau_0] f(W_{\nu^*}(0)),
$$

where $K_\nu^W(0)$ is a Lipschitz coefficient for $W_\nu(0)$ w.r.t. ν, which can be expressed as in Theorem 3.3.1 as

$$
K_\nu^W(0) \stackrel{\text{def}}{=} \sum_{n\in\mathbb{Z}} \tau_n \mathbb{1}_{[R_-^*(0), R_+^*(0))}(T_n).
$$

One can easily check that assumption **A13** suffices to prove that $|\Phi/h|$ is bounded by an integrable variable. Consequently, we can apply the Dominated Convergence Theorem and find the expected result. ■

The method used so far does not apply to the case where the parameter of interest is a parameter of the inter-arrival times; in this case, the Palm measure associated to the arrival process depends on the parameter and the method fails. We show how a change of time scale can be used in some cases. We consider a $G/G/1$ queue with inter-arrival times $\{\tau_n(\alpha)\}_{n\in\mathbb{Z}}$, $\alpha \ge \alpha^* > 0$ and we will restrict our attention to the following case:

Assumption A14 *α is a scale parameter for $\tau_n(\alpha)$, that is $\tau_n(\alpha) = \alpha\eta_n$.*

Lindley's equation takes the form

$$W_\alpha(t) = \Big(W_\alpha(T_n(\alpha)-) + \sigma_n - (t - T_n(\alpha))\Big)^+, \; t \in [T_n(\alpha), T_{n+1}(\alpha)) \quad (3.5.4)$$

Now define a $G/G/1$ queue *with speed* α which inter-arrival times, service times and arrival process are given by:

$$\begin{aligned} \tilde{\tau}_n &\stackrel{\text{def}}{=} \frac{\tau_n(\alpha)}{\alpha} = \eta_n \\ \tilde{\sigma}_n &\stackrel{\text{def}}{=} \sigma_n \\ \widetilde{T}_n &\stackrel{\text{def}}{=} \frac{T_n(\alpha)}{\alpha}. \end{aligned}$$

These processes are stationary with respect to the measurable flow $\tilde{\theta}_t \stackrel{\text{def}}{=} \theta_{\alpha t}$ and the queue they define is stable whenever the original one is; this queue will be referred to as the "auxiliary system". Throughout this section, we will use the same notations as for the main system, but with a tilde. Lindley's equation for the auxiliary system reads:

$$\widetilde{W}_\alpha(t) = \Big(\widetilde{W}_\alpha(\widetilde{T}_n-) + \tilde{\sigma}_n - \alpha(t - \widetilde{T}_n)\Big)^+, \; t \in [\widetilde{T}_n, \widetilde{T}_{n+1}). \quad (3.5.5)$$

Comparing equations (3.5.4) and (3.5.5) and noting that the process $W_\alpha(\alpha t)$ is stationary with respect to the flow $\tilde{\theta}_t$, uniqueness in Loynes' Stability Theorem—see Baccelli and Brémaud [1]— yields

$$W_\alpha(t) = \widetilde{W}_\alpha(t/\alpha).$$

The effect of the change of time scale can be seen on Figure 3.10. Moreover,

$$\begin{aligned} \tilde{\lambda} &= \mathbb{E}\widetilde{A}((0,1]) \\ &= \mathbb{E}A((0,\alpha]) = \alpha\lambda(\alpha) \\ \widetilde{W}'_\alpha(0) &= W'_\alpha(0). \end{aligned}$$

In the computation of $\tilde{\lambda}$, we use the fact that the auxiliary system is defined on the same probability space than the main one. It has its own Palm measure associated to $\{\widetilde{T}_n\}_{n\in\mathbb{Z}}$, say $\widetilde{\mathrm{P}}^0$. The way to switch between probability measures P^0_α and $\widetilde{\mathrm{P}}^0$ will be shown in the proof of Theorem 3.5.2. Before proceeding, we need a set of **A13**-like conditions:

Assumption A15 *The following conditions hold:*

(*i*) $\mathbb{E}^0_\alpha[\tau_0]^4 < \infty$;

(*ii*) $\mathbb{E}^0_{\alpha^*}[A([R^*_0, R^*_1))]^4 < \infty$;

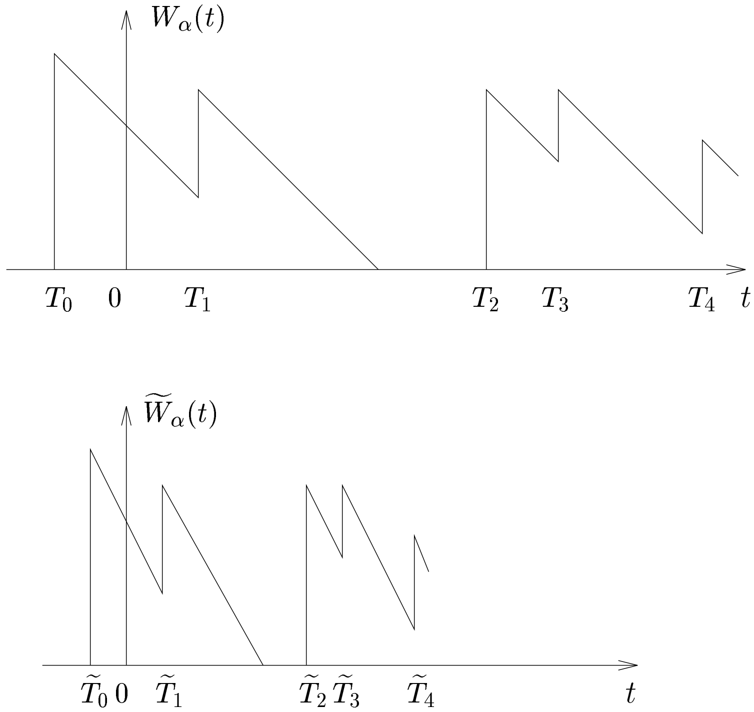


Figure 3.10: Change of time scale for $\alpha = 2$.

(*iii*) $\mathbb{E}^0_{\alpha^*}[f(W_{\alpha^*}(0))]^2 < \infty$.

Using this model, we find the following result:

Theorem 3.5.2 *Assume* **A14** *and* **A15** *hold; then*

$$\begin{aligned} J_r'(\alpha) &= \frac{\lambda}{\alpha}\mathbb{E}^0_\alpha\Big\{\alpha W_\alpha'(0)\Big[f(W_\alpha(0)) - f(W_\alpha(T_1-))\Big] \\ &\quad + \Big[F(W_\alpha(0)) - F(W_\alpha(T_1-))\Big] \\ &\quad - [W_\alpha(0) - W_\alpha(T_1-)]f(W_\alpha(T_1-))\Big\} \end{aligned} \tag{3.5.6}$$

$$\begin{aligned} J_l'(\alpha) &= \frac{\lambda}{\alpha}\mathbb{E}^0_\alpha\Big\{\alpha W_\alpha'(0)\Big[f(W_\alpha(0)) - f(W_\alpha(T_1-))\Big] \\ &\quad + \Big[F(W_\alpha(0)) - F(W_\alpha(T_1-))\Big] \\ &\quad - [W_\alpha(0) - W_\alpha(T_1-)]f(W_\alpha(T_1-)) \\ &\quad - \alpha W_\alpha'(0)\Big[\mu_f(\{W_\alpha(0)\}) - \mu_f(\{W_\alpha(T_1-)\})\Big] \\ &\quad + [W_\alpha(0) - W_\alpha(T_1-)]\mu_f(\{W_\alpha(T_1-)\})\Big\}. \end{aligned} \tag{3.5.7}$$

If f is continuous or if $W_\alpha(0)$ and $W_\alpha(T_1(\alpha)-)$ admit densities with *respect to* P^0_α *then* J *is differentiable w.r.t.* α *and its derivative is equal to* J'_r.

Proof We have

$$J(\alpha) = \mathbb{E} f(W_\alpha(0)) = \mathbb{E} f(\widetilde{W}_\alpha(0))$$

where $\widetilde{W}_\alpha(0)$ is the workload of the auxiliary queue with speed α. We aim to apply Theorem 3.5.1 to this queue and then adapt the result to the main queue. The three conditions of **A13** correspond to the three ones of **A15**: for condition **A13**-(i), note that

$$\begin{aligned}
\widetilde{\mathbb{E}}^0[\tilde{\tau}_0]^4 &= \frac{1}{\tilde{\lambda}} \mathbb{E} \sum_{n\in\mathbb{Z}} [\tilde{\tau}_n]^4 \mathbb{1}_{\{\widetilde{T}_n \in (0,1]\}} \\
&= \frac{1}{\alpha\lambda(\alpha)} \mathbb{E} \sum_{n\in\mathbb{Z}} \Big[\frac{\tau_n(\alpha)}{\alpha}\Big]^4 \mathbb{1}_{\{T_n(\alpha)\in(0,\alpha]\}} \\
&= \frac{1}{\alpha^4} \mathbb{E}^0_\alpha [\tau_0]^4 < \infty
\end{aligned}$$

and for **A13**-(ii),

$$\begin{aligned}
\widetilde{\mathbb{E}}^0[\widetilde{A}([\widetilde{R}^*_0, \widetilde{R}^*_1))]^4 &= \frac{1}{\tilde{\lambda}} \mathbb{E} \sum_{n\in\mathbb{Z}} [\widetilde{A}([\widetilde{R}^*_-(\widetilde{T}_n), \widetilde{R}^*_+(\widetilde{T}_n)))]^4 \mathbb{1}_{\{\widetilde{T}_n \in (0,1]\}} \\
&= \frac{1}{\alpha\lambda(\alpha)} \mathbb{E} \sum_{n\in\mathbb{Z}} [A([R^*_-(T_n), R^*_+(T_n)))]^4 \mathbb{1}_{\{T_n(\alpha^*)\in(0,\alpha^*]\}} \\
&= \mathbb{E}^0_{\alpha^*}[A([R^*_0, R^*_1))]^4 < \infty.
\end{aligned}$$

Finally, for **A13**-(iii),

$$\widetilde{\mathbb{E}}^0[f(\widetilde{W}_{\alpha^*}(0))]^2 = \mathbb{E}^0_{\alpha^*}[f(W_{\alpha^*}(0))]^2 < \infty.$$

So we apply Theorem 3.5.1 and find for the right-hand derivative:

$$\begin{aligned}
J'_r(\alpha) &= \frac{\tilde{\lambda}}{\alpha^2} \widetilde{\mathbb{E}}^0 \Big\{ \alpha \widetilde{W}'_\alpha(0) \Big[f(\widetilde{W}_\alpha(0)) - f(\widetilde{W}_\alpha(\widetilde{T}_1 -)) \Big] \\
&\quad + \Big[F(\widetilde{W}_\alpha(0)) - F(\widetilde{W}_\alpha(\widetilde{T}_1 -)) \Big] \\
&\quad - [\widetilde{W}_\alpha(0) - \widetilde{W}_\alpha(\widetilde{T}_1 -)] f(\widetilde{W}_\alpha(\widetilde{T}_1 -)) \Big\}.
\end{aligned}$$

This gives Equation (3.5.6); the left-hand derivative is derived similarly. ■

3.6 Implementation of the method

The formulas given in preceding sections will be interesting only if they provide estimates which are (i) easy to compute and (ii) strongly consistent, which means that they converge a.s. to their expected values. In this section we show how the estimate given by Corollary 3.3.3 can be used in simulation when the system is ergodic. In this case, ergodic theorem (3.A.6) applied to equation (3.3.3) reads:

$$\begin{aligned} J'(\theta) &= \lim_{n\to\infty} \frac{\lambda}{n} \sum_{k=0}^{n-1} W'_\theta(T_k)[f(W_\theta(T_k)) - f(W_\theta(T_{k+1}-))] \\ &= \lim_{n\to\infty} \frac{\lambda}{n} \sum_{k=0}^{n-1} W'_\theta(T_k)[f(W_\theta(T_k)) - f(W_\theta(T_k-))]. \qquad (3.6.1) \end{aligned}$$

The different ingredients of this formula are easy to evaluate once the simulation of the queue is set up: $W_\theta(T_k)$ and $W_\theta(T_k-)$ are known when customer k joins the queue; to get $W'_\theta(T_k)$, we use equation (3.1.1), keeping in mind that both $W(t)$ and $W'(t)$ are *càdlàg* processes and find:

$$W'_\theta(T_k) = \begin{cases} \sigma'_k(\theta) & \text{if customer } k \text{ finds the system empty} \\ W'_\theta(T_{k-1}) + \sigma'_k(\theta) & \text{else.} \end{cases}$$

As shown in Section 3.2, σ'_k can most of the time be expressed as a function of σ_k and θ, say $\sigma'_k = D(\sigma_k, \theta)$. So if we define $w_k = W_\theta(T_k-)$ and $d_k = W'_\theta(T_k)$, we have

$$\begin{aligned} w_k &= (w_{k-1} + \sigma_{k-1} - \tau_{k-1})^+ \\ d_k &= d_{k-1}\mathbb{1}\{w_k > 0\} + D(\sigma_k, \theta), \end{aligned}$$

and equation (3.6.1) shows that

$$\phi_n \stackrel{\text{def}}{=} \frac{\lambda}{n} \sum_{k=0}^{n-1} d_k[f(w_k + \sigma_k) - f(w_k)]$$

is a strongly consistent estimate of $J'(\theta)$. Since our Palm estimate does not require differentiability for f, one will want to check whether it is as accurate as the classic IPA estimate: if the system is ergodic, the ergodic theorem (3.A.5) of Appendix applied to equation (3.1.2) gives

$$\begin{aligned} J'(\theta) &= \lim_{t\to\infty} \frac{1}{t} \int_0^t W'_\theta(s) f'(W_\theta(s))\, ds \\ &= \lim_{n\to\infty} \frac{1}{T_n} \sum_{k=0}^{n-1} \int_{T_k}^{T_{k+1}} W'_\theta(s) f'(W_\theta(s))\, ds \\ &= \lim_{n\to\infty} \frac{1}{T_n} \sum_{k=0}^{n-1} W'_\theta(T_k)[f(W_\theta(T_k)) - f(W_\theta(T_{k+1}-))], \quad (3.6.2) \end{aligned}$$

where all the limits are valid P^0-a.s. or P-a.s. indifferently. In the third equality, we used the fact that $W'_\theta(t)$ is zero during idle periods. Comparing equations (3.6.1) and (3.6.2), we see that the estimates based on the same amount of data give very close expressions; in fact, they are even equal when λ needs to be estimated. For comparisons between time-average and customer-average estimates, see for instance Glynn and Whitt [26]. The implementation of an estimate of the second derivative of J would be done exactly in the same way, except that the formulas involved are slightly more complicated.

3.A Appendix: a short introduction to Palm theory

In this appendix we give without proof some basic results on Palm theory; interested readers can refer to Baccelli and Brémaud [1] for a more complete presentation of the subject. The stationary framework is the following: given a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, let θ_t, $t \in \mathbb{R}$ be a measurable flow $(\Omega, \mathcal{F}) \to (\Omega, \mathcal{F})$, i.e.:

- $(t, \omega) \mapsto \theta_t\omega$ is measurable w.r.t. $\mathcal{B}(\mathbb{R}) \otimes \mathcal{F}$,
- θ_t is bijective for all $t \in \mathbb{R}$,
- $\theta_t \circ \theta_s = \theta_{t+s}$ for all $t, s \in \mathbb{R}$. In particular, $\theta_0 =$*identity* and $\theta_t^{-1} = \theta_{-t}$.

Note that there is nothing common between the flow θ_t and the parameter θ of the queue; these are the traditional notations in sensitivity analysis and Palm theory. We assume that $\mathrm{P} \circ \theta_t = \mathrm{P}$. Let $\{T_n\}_{n\in\mathbb{Z}}$ and $\{U_n\}_{n\in\mathbb{Z}}$ be two simple point processes and let A and D be the associated counting measures, that is, for all borelian set $C \subset \mathbb{R}$,

$$A(C) = \sum_{n\in\mathbb{Z}} \mathbb{1}_C(T_n), \; D(C) = \sum_{n\in\mathbb{Z}} \mathbb{1}_C(U_n)$$

and assume that for each $n \in \mathbb{Z}$, $U_n - T_n \overset{\text{def}}{=} W_n > 0$. We take the convention $T_0 \leq 0 < T_1$ and note:

$$\begin{aligned} T_-(t) &= \sup(T_n : T_n \leq t), \\ T_+(t) &= \inf(T_n : T_n > t). \end{aligned}$$

A and D are viewed as arrival and departure processes and we note $X(t)$ a queueing process associated with them. Let $B(t)$ be a non decreasing *càdlàg*—i.e. right continuous with left limits—real-valued process and $Z(t)$ a non-negative real-valued stochastic process. We assume that these processes are compatible with the flow θ_t, that is

$$\begin{aligned} A(\omega, C + t) &= A(\theta_t\omega, C) \\ Z(\omega, t) &= Z(\theta_t\omega, 0). \end{aligned}$$

Similar equalities hold for D and X; if we define $\lambda_A = \mathbb{E}[A((0,1])]$, then there exists a probability measure called the Palm probability of the stationary process $(A, \theta_t, \mathrm{P})$ verifying the Swiss Army Formula (Brémaud [7]):

$$\lambda_A \mathbb{E}_A^0[\int_{(0,W_0]} Z(s)dB(s)] = \frac{1}{t}\mathbb{E}[\int_{(0,t]} X(s-)Z(s)dB(s)] \tag{3.A.1}$$

The Swiss Army Formula is not the definition of the Palm measure, but we will see that it contains this definition and the classic formulas of Palm theory. We shall add that under P_A^0, $T_0 = 0$ a.s.. We derive now some useful formulas from (3.A.1). The first one is the inversion formula: Take $U_n = T_{n+1}$ and $B(s) = s$; then $X(t) = 1$ and

$$\mathbb{E}[Z(0)] = \lambda_A \mathbb{E}_A^0[\int_0^{T_1} Z(s)ds]. \tag{3.A.2}$$

The second formula is Neveu's exchange formula (Neveu [48]): we take D as for the inversion formula and remark that if $B \equiv A$, (3.A.1) reads

$$\lambda_A \mathbb{E}_A^0[Z(0)] = \frac{1}{t}\mathbb{E}[\int_{(0,t]} Z(s)dA(s)].$$

This is Mecke's definition of Palm probability. Now, if B is a point process, we use the above equation and (3.A.1) to derive the exchange formula:

$$\lambda_A \mathbb{E}_A^0[\int_0^{T_1} Z(s)dB(s)] = \lambda_B \mathbb{E}_B^0[Z(0)]. \tag{3.A.3}$$

We will now prove a simple lemma which replaces Wald's identity for stationary systems:

Lemma 3.A.1 *Let $\{R_n\}_{n\in\mathbb{Z}}$ be a stationary stochastic point process with associated measure A'. The following holds:*

$$\mathbb{E}_A^0\Big[\int_{[R_0,R_1)} Z(s)dA(s)\Big] = \mathbb{E}_A^0[A([R_0,R_1))\, Z(0)]. \tag{3.A.4}$$

Proof If we note Y the random variable inside the expectation of the l.h.s. of equation (3.A.4), then

$$Y \circ \theta_{T_i} = \int_{[R_-(T_i),R_+(T_i))} Z(s)dA(s) = \sum_{n\in\mathbb{Z}} Z(T_n)\mathbb{1}_{[R_-(T_i),R_+(T_i))}(T_n).$$

Since $R_\pm(T_i) = R_\pm(0)$ if $T_i \in [R_-(0), R_+(0))$,

$$\begin{aligned}\sum_{T_i\in[R_-(0),R_+(0))} Y \circ \theta_{T_i} &= \sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} Z(T_n)\mathbb{1}_{[R_-(0),R_+(0))}(T_n)\mathbb{1}_{[R_-(0),R_+(0))}(T_i)\\ &= A([R_0,R_1))\sum_{n\in\mathbb{Z}} Z(T_n)\,\mathbb{1}_{[R_0,R_1)}(T_n).\end{aligned}$$

If we now apply Neveu's exchange formula (3.A.3) between P_A^0 and $P_{A'}^0$, we obtain:

$$\begin{aligned}
\lambda_A \mathbb{E}_A^0 Y &= \lambda_{A'} \mathbb{E}_{A'}^0 \sum_{T_i \in [R_-(0), R_+(0))} Y \circ \theta_{T_i} \\
&= \lambda_{A'} \mathbb{E}_{A'}^0 \Big[\sum_{n \in \mathbb{Z}} A([R_0, R_1))\, Z(T_n)\, \mathbb{1}_{[R_0,R_1)}(T_n) \Big] \\
&= \lambda_A \mathbb{E}_A^0 [A([R_0, R_1))\, Z(0)],
\end{aligned}$$

which is exactly equality (3.A.4). ■

Formula (3.A.4) can be seen as an extension of Wald's identity that can be used for stationary sequences instead of i.i.d. variables and applies to any stationary process. It is not as convenient as Wald's identity is, but is valid in a wider setting.

Remark Lemma 3.A.1 is also a corollary of the extended $H = \lambda G$ formula (6.2) of Brémaud [7].

Palm probabilities can also be given an interpretation which relates them to simulation. When (P, θ_t) is ergodic, the ergodic theorems for P and P^0 read:

$$\begin{aligned}
\mathbb{E}[Y] &= \lim_{t \to \infty} \frac{1}{t} \int_0^t Y \circ \theta_s ds \\
\mathbb{E}_A^0[Y] &= \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} Y \circ \theta_{T_k}
\end{aligned}$$

which imply that:

$$\mathbb{E}[Z(0)] = \lim_{t \to \infty} \frac{1}{t} \int_0^t Z(s) ds \tag{3.A.5}$$

$$\mathbb{E}_A^0[Z(0)] = \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} Z(T_k). \tag{3.A.6}$$

These equalities are valid P-a.s. and P_A^0-a.s. This shows that $\mathbb{E}[Z(0)]$ is the *time-average* of the process $Z(t)$, whether $\mathbb{E}_A^0[Z(0)]$ is its *customer-average*.

Bibliographie

[1] Baccelli, F., and Brémaud, P. *Palm Probabilities and Stationary Queues*, vol. 41 of *Lecture Notes in Statistics*. Springer-Verlag, 1987.

[2] Birman, A., and Kogan, Y. Asymptotic evaluation of closed queueing networks with many stations. *Communications in Statistics—Stochastic Models 8*, 3 (1992), 543–563.

[3] Borovkov, A. A., and Schassberger, R. Ergodicity of a polling network. *Stochastic Processes and their Applications 50*, 2 (1994), 253–262.

[4] Boxma, O. J., and Weststrate, J. Waiting time in polling systems with Markovian server routing. In *Messung, Modellierung und Bewertung von Rechensysteme* (Berlin, 1989), Proc. Conference Braunschweig, Springer-Verlag, pp. 89–104.

[5] Bradlow, H. S., and Byrd, H. F. Mean waiting time evaluation of packet switches for centrally controlled PBX's. *Performance Evaluation 7*, 4 (1987), 309–327.

[6] Brémaud, P. Maximal coupling and rare perturbation analysis. *Queueing Systems, Theory and Applications 11*, 4 (1992), 307–333.

[7] Brémaud, P. A Swiss Army formula of Palm calculus. *Journal of Applied Probability 30* (1993), 40–51.

[8] Brémaud, P., and Lasgouttes, J.-M. Stationary IPA estimates for non-smooth functions of the GI/G/1/∞ workload. Rapport de Recherche 1677, INRIA, Rocquencourt BP 105 – 78153 Le Chesnay – France, May 1992.

[9] Brémaud, P., and Lasgouttes, J.-M. Stationary IPA estimates for non-smooth G/G/1/∞ functionals via Palm inversion and level-crossing analysis. *Discrete Event Dynamic Systems: Theory and Applications 3*, 4 (1993), 347–374.

[10] Brémaud, P., and Vázquez-Abad, F. J. On the pathwise computation of derivatives with respect to the rate of a point process: the phantom RPA method. *Queueing Systems, Theory and Applications 10*, 3 (1992), 249–270.

[11] Cao, X.-R. Convergence of parameter sensitivity estimates in a stochastic experiment. *IEEE Transactions on Automatic Control 30*, 9 (1985), 834–843.

[12] Fayolle, G., and Lasgouttes, J.-M. A state-dependent polling model with Markovian routing. Rapport de Recherche 2279, INRIA, Rocquencourt BP 105 – 78153 Le Chesnay – France, June 1994.

[13] Fayolle, G., and Lasgouttes, J.-M. Limit laws for large product-form networks: connections with the Central Limit Theorem. Rapport de Recherche 2513, INRIA, Mar. 1995.

[14] Fayolle, G., and Lasgouttes, J.-M. A state-dependent polling model with Markovian routing. In *The IMA Volumes in Mathematics and its Applications*, F. P. Kelly and R. R. Williams, Eds., vol. 71 of *Stochastic Networks*. Springer-Verlag, 1995, pp. 283–312.

[15] Fayolle, G., Malyshev, V. A., and Menshikov, M. V. *Topics in the Constructive Theory of Countable Markov Chains*. Cambridge University Press, 1995.

[16] Fayolle, G., and Zamyatin, A. A. Controlled random walks in Z_+^N and their applications to queueing networks. *Preprint* (1993). To appear.

[17] Feller, W. *An Introduction to Probability Theory and its Applications, Vol. II*, 2 ed. John Wiley & Sons, 1971.

[18] Ferguson, M. J. Mean waiting time for a token ring with station dependent overheads. In *Local Area & Multiple Access Networks*, R. L. Pickholtz, Ed. Computer Science Press, Rockville, Maryland, 1986, ch. 3, pp. 43–67.

[19] Fricker, C., and Jaïbi, M. R. Monotonicity and stability of periodic polling models. *Queueing Systems, Theory and Applications 15* (1994), 211–238.

[20] Fricker, C., and Jaïbi, M. R. Stability of a polling model with a Markovian scheme. Rapport de Recherche 2278, INRIA, Rocquencourt BP 105 – 78153 Le Chesnay – France, May 1994.

[21] Fu, M. C., and Hu, J.-Q. Extensions and generalizations of smoothed perturbation analysis in a generalized semi-Markov process framework. *IEEE Transactions on Automatic Control 37*, 10 (1992), 1483–1500.

[22] Georgiadis, L., and Szpankowski, W. Stability of token passing rings. *Queueing Systems, Theory and Applications*, 11 (1992), 7–33.

[23] Glasserman, P. *Gradient Estimation via Perturbation Analysis.* Kluwer Academic Publishers, Norwell, Massachusetts, 1991.

[24] Glasserman, P., and Gong, W.-B. Smoothed perturbation analysis for a class of discrete event systems. *IEEE Transactions on Automatic Control 35*, 11 (1991), 1218–1230.

[25] Glynn, P. W. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM 33*, 10 (1990), 76–84.

[26] Glynn, P. W., and Whitt, W. Indirect estimation via $L = \lambda W$. *Operations Research 37*, 1 (1989), 82–103.

[27] Gong, W.-B., and Ho, Y.-C. Smoothed (conditional) perturbation analysis of discrete event dynamical systems. *IEEE Transactions on Automatic Control 32*, 10 (1987), 858–866.

[28] Ho, Y.-C., and Cao, X.-R. Optimization and perturbation analysis of queueing networks. *Journal of Optimization Theory and Applications 40* (1983), 559–582.

[29] Ho, Y.-C., and Cao, X.-R. *Perturbation Analysis of Discrete Event Dynamic Systems.* Kluwer Academic Publishers, 1991.

[30] Kelly, F. P. *Reversibility and Stochastic Networks.* Wiley, 1979.

[31] Kleinrock, L., and Levy, H. The analysis of random polling systems. *Operations Research 36*, 5 (1988), 716–732.

[32] Knessl, C., and Tier, C. Asymptotic expansions for large closed queueing networks. *Journal of the ACM 37*, 1 (1990), 144–174.

[33] Kogan, Y. Another approach to asymptotic expansions for large closed queueing networks. *Operations Research Letters 11* (1992), 317–321.

[34] Kogan, Y., and Birman, A. Asymptotic analysis of closed queueing networks with bottlenecks. In *Proceedings of the International Conference on Performance of Distributed Systems and Integrated Communication Networks* (Kyoto, 1991), T. Hasegawa, H. Takagi, and Y. Takahashi, Eds., pp. 237–252.

[35] Konstantopoulos, P., and Zazanis, M. Sensitivity analysis for stationary and ergodic queues. *Advances in Applied Probability 24*, 3 (1992), 738–750.

[36] Konstantopoulos, P., and Zazanis, M. Sensitivity analysis for stationary and ergodic queues: Additional results. *Advances in Applied Probability 26* (1994), 556–560.

[37] Lavenberg, S. S. *Computer Performance Modelling Handbook.* Academic Press, 1983.

[38] Levy, H., and Sidi, M. Polling systems: Applications, modeling and optimization. *IEEE Transactions on Communications 38*, 10 (1990), 1750–1760.

[39] Levy, H., Sidi, M., and Boxma, O. J. Dominance relations in polling systems. *Queueing Systems, Theory and Applications 6* (1990), 155–172.

[40] Loève, M. *Probability Theory*, fourth ed. D. Van Nostrand Company, 1977.

[41] Malyshev, V., and Yakovlev, A. Condensation in large closed Jackson networks. *Annals of Applied Probability 6*, 1 (1996), 92–115.

[42] Malyshev, V. A., and Menshikov, M. V. Ergodicity, continuity and analyticity of countable Markov chains. *Trans. Moscow. Math. Soc. 39* (1979), 3–48.

[43] Mazumdar, R. R., Badrinath, V., Guillemin, F., and Rosenberg, C. On pathwise rate conservation for a class of semi-martigales. *Stochastic Processes and their Applications 47* (1993), 119–130.

[44] McKenna, J., and Mitra, D. Integral representations and asymptotic expansions for closed Markovian queueing networks: Normal usage. *Bell System Technical Journal 61* (1982), 661–683.

[45] McKenna, J., and Mitra, D. Asymptotic expansions and integral representations of moments of queue lengths in closed Markovian networks. *Journal of the ACM 31* (1984), 346–360.

[46] McKenna, J., Mitra, D., and Ramakrishnan, K. G. A class of closed Markovian queueing networks: Integral representations, asymptotic expansions and generalizations. Tech. rep., AT&T Bell Laboratories, 1990.

[47] Mitra, D., and McKenna, J. Asymptotic expansions for closed Markovian queueing networks with state dependent service rates. *Journal of the ACM 33* (1986), 568–592.

[48] Neveu, J. Sur les mesures de Palm de deux processus ponctuels stationnaires. *Zeitschrift für Wahrscheinlichkeitstheorie und verw. Geb. 34* (1976), 199–203.

[49] Reiman, M. I., and Weiss, A. Sensitivity analysis for simulations via likelihood ratios. *Operations Research 37*, 5 (1989), 830–844.

[50] Schassberger, R. Stability of polling networks with state-dependent server routing. *Preprint* (1993).

[51] Serfozo, R. F. Markovian network processes: Congestion-dependent routing and processing. *Queueing Systems, Theory and Applications 5* (1989), 5–36.

[52] Servi, L. D. Average delay approximation of M/G/1 cyclic service queues with Bernoulli schedules. *IEEE Journal on Selected Areas in Communications SAC-4*, 6 (1986), 813–822. Correction in Vol. Sac-5, No. 3, p. 547, 1987.

[53] Srinivasan, M. M. Nondeterministic polling systems. *Management Science 37*, 6 (1991), 667–681.

[54] Suri, R. Implementation of sensitivity calculations on a Monte Carlo experiment. *Journal of Optimization Theory and Applications 40*, 4 (1983), 625–630.

[55] Suri, R., and Zazanis, M. Perturbation analysis gives strongly consistent sensitivity estimates for the M/G/1 queue. *Management Science 34*, 1 (1988), 39–64.

[56] Takagi, H. Queueing analysis of polling models: an update. In *Stochastic Analysis of Computer and Communication Systems*, H. Takagi, Ed. North Holland, Amsterdam, 1990, pp. 267–318.

[57] Tang, Q.-Y., and Chen, H.-F. Convergence of perturbation analysis based optimization algorithm with fixed number of customer period. *Discrete Event Dynamic Systems: Theory and Applications 4*, 4 (1994), 359–375.

[58] Whittle, P. *Systems in Stochastic Equilibrium.* Wiley, 1986.

[59] Zazanis, M. A., and Suri, R. Estimating first and second derivatives of response time for $G/G/1$ queues from a single sample path. Technical report, Division of Applied Sciences, Harvard University, Cambridge, MA, 1985.