9

Files d'attente

De nombreuses entreprises sous-traitent leurs centres d'appels téléphoniques à des sociétés de services. Une de ces sociétés de services annonce dans sa publicité sur internet qu'elle peut dimensionner le nombre d'opérateurs du centre d'appel en fonction :

- du nombre moyen d'appels par unité de temps, λ,
- de la durée moyenne des appels, $1/\mu$,
- et d'un seuil d'attente, compté en nombre de sonneries, tel que la probabilité d'attendre plus longtemps que ce seuil avant qu'un opérateur ne réponde soit inférieur à 20 %.

L'étude des lois des temps d'attentes, entre autres, dans les files d'attente remonte aux résultats d'Erlang en 1917 [5], qui travaillait pour la Compagnie de Téléphone de Copenhague. Depuis, la modélisation des files d'attente et des réseaux a connu un essor considérable dû en particulier à la diversité et à l'omniprésence des files d'attente : caisses d'une grande surface, guichets des compagnies de transport, réseaux téléphoniques, réseaux informatiques, requêtes des microprocesseurs, etc.

Le but de ce chapitre est de présenter des modèles élémentaires de files d'attente et de réseaux, mais qui sont en fait représentatifs des phénomènes observés. Plus précisément nous introduisons, dans le paragraphe 9.1, des modèles de chaînes de Markov à temps continu, vues au Chap. 8, pour les files d'attentes à K serveurs. Après avoir exhibé le générateur infinitésimal, nous calculons, au paragraphe 9.2 pour un serveur, et au paragraphe 9.3 pour K serveurs, la probabilité invariante quand elle existe, le nombre moyen de personnes dans la file d'attente, la loi asymptotique du temps d'attente pour un client arrivant dans le système. Le calcul de la loi du temps d'attente permet de comprendre comment dimensionner le nombre de serveurs à partir du nombre moyen d'appels par unité de temps, λ, de la durée moyenne des appels, $1/\mu$, et d'une borne sur le temps moyen d'attente dans la file. On peut également comparer, selon plusieurs critères, les files d'attente à un et deux serveurs, quand le serveur de la première file d'attente est aussi efficace que les deux serveurs de l'autre file d'attente. Le paragraphe 9.4 aborde un exemple

élémentaire de réseaux. Finalement, le paragraphe 9.5 présente le lien entre certaines files d'attente et les processus de Galton-Watson, vus au Chap. 4.

Une vaste littérature sur le domaine est disponible. Nous renvoyons aux ouvrages suivants : Bougerol [3] et Brémaud [4] pour une présentation élémentaire, Asmussen [1], Baccelli et Brémaud [2], et Robert [7] pour une étude plus approfondie.

9.1 Introduction

9.1.1 Modélisation des files d'attente

Nous présentons d'abord la terminologie usuelle pour la description des files d'attente, puis nous donnerons le générateur infinitésimal de la chaîne de Markov correspondant aux files $M/M/K$.

Une file d'attente est décrite par un processus d'arrivée des clients, un modèle pour les temps de services des requêtes exprimées et le mode de gestion des requêtes.

1. Le processus des temps d'arrivées : on utilise la notation GI si les temps entre deux arrivées successives de clients, ou temps d'inter-arrivées, sont des variables aléatoires indépendantes et de même loi. Le processus d'arrivée est alors la fonction de comptage associé aux temps d'inter-arrivées. Si de plus la loi est une loi exponentielle, alors le processus des arrivées est un processus de Poisson, et on utilise la notation M pour souligner le caractère markovien.
2. Les temps de service : on utilise la notation GI si les temps de services sont des variables aléatoires indépendantes de même loi et indépendantes du processus d'arrivée. Si de plus la loi des temps de services est une loi exponentielle, on utilise la notation M. Nous verrons que dans ce cas, si le processus d'arrivée est un processus de Poisson, alors l'évolution de la taille de la file d'attente est une chaîne de Markov.
3. Gestion des requêtes : on distingue plusieurs éléments.
   a) Le nombre de guichets ou serveurs est noté $K \in \mathbb{N}^* \cup \{+\infty\}$.
   b) La taille de la salle d'attente notée $k \in \mathbb{N} \cup \{+\infty\}$. Quand on téléphone à un centre d'appel, si tous les opérateurs sont déjà occupés, alors l'appel est mis en attente. Au delà d'un certain nombre d'appels en attente, correspondant à la taille de la salle d'attente, la connexion peut être refusée. Pour les réseaux téléphoniques, si tous les serveurs sont occupés, ce qui correspond à un réseau saturé, alors la connexion est refusée. Dans ce cas, la taille de la salle d'attente est donc nulle. Enfin, on suppose que la salle d'attente est commune à tous les serveurs. Ce n'est pas le cas par exemple aux caisses d'une grande surface, où chaque serveur a une file d'attente.

c) Il existe plusieurs politiques de gestion des requêtes.
   - FIFO (« First In First Out ») : les clients sont servis suivant leur ordre d'arrivée.
   - LIFO (« Last In First Out ») : le dernier client arrivé est le premier servi. On distingue suivant que le service en cours est terminé avant de servir le dernier client arrivé ou interrompu (LIFO avec préemption), puis repris lorsque le service du ou des derniers clients arrivés est terminé. Cette dernière stratégie est intéressante si les requêtes sont de types très différents : pour un serveur informatique elle permet par exemple de traiter rapidement les courriels courts pour lesquels la probabilité d'interruption est faible, et d'accepter en contrepartie que l'envoi de courriels incluant par exemple de gros fichiers soit en partie ralenti.
   - Requêtes partagées : tous les clients sont servis en même temps, mais avec une vitesse inversement proportionnelle au nombre de clients.
   - Aléatoire : le prochain client de la file d'attente à être servi est choisi au hasard.
   - SPT (« shortest processing time ») : le prochain client de la file d'attente à être servi est celui qui a la requête la plus courte.
   - ...

Dans ce qui suit on ne considère que la stratégie FIFO. On utilise la notation conventionnelle de Kendall $M/GI/K/k$ pour décrire une file d'attente dont les temps d'inter-arrivées sont indépendants de même loi exponentielle, les temps de services sont indépendants de loi quelconque, le nombre de serveurs est K et la taille de la salle d'attente est k. On omet généralement k lorsque $k = \infty$.

9.1.2 Présentation des files $M/M/K$

On considère une file $M/M/K$, où les temps d'inter-arrivées suivent la loi exponentielle de paramètre $\lambda > 0$, et où les temps de services suivent la loi exponentielle de paramètre $\mu > 0$. On note $X_t \in \mathbb{N}$ la taille du système à l'instant $t \geq 0$, qui comprend les clients dans la salle d'attente, ainsi que les clients aux guichets.

On utilise la notation $x \wedge y = \min(x, y)$.

Proposition 9.1.1. *Le processus $X = (X_t, t \geq 0)$ est une chaîne de Markov à temps continu homogène irréductible de générateur infinitésimal $A = (A(i,j), i \geq 0, j \geq 0)$ dont les termes non nuls hors de la diagonale sont $A(i, i+1) = \lambda$ et $A(i+1, i) = (i \wedge K)\mu$ pour $i \in \mathbb{N}$, soit*

$$A = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \dots \\ 0 & (2 \wedge K)\mu & -(\lambda + (2 \wedge K)\mu) & \lambda & \dots \\ & & \vdots & & \end{pmatrix}. \tag{9.1}$$

Les différentes politiques de gestion des requêtes ne changent pas le temps de travail du serveur, mais les temps d'attente des clients. La loi de X_t reste inchangée si on regarde la stratégie LIFO sans préemption, ou la stratégie des requêtes partagées. Enfin les files d'attentes $M/M/K$ exhibent des comportements que l'on retrouve pour les files d'attente plus générales.

Démonstration. On considère les états successifs du système $(Z_n, n \geq 0)$ et $(V_n, n \geq 1)$ les temps entre les changements d'état du système.

Si $Z_0 = 0$, le prochain événement est l'arrivée d'un client. Donc la loi de V_1 conditionnellement à $\{Z_0 = 0\}$ est la loi exponentielle de paramètre λ.

Si $Z_0 = r \geq 1$, le prochain événement est soit l'arrivée d'un nouveau client, à la date T, soit la fin de service de l'un des $r \wedge K$ clients, aux dates $S_1, \ldots, S_{r\wedge K}$. Il arrive donc à l'instant $V_1 = \min\{T, S_1, \ldots, S_{r\wedge K}\}$. D'après le lemme 7.4.4, la loi de V_1 est alors la loi exponentielle de paramètre $\lambda + (r \wedge K)\mu$. De plus on a

$$\mathbb{P}(Z_1 = r+1 | Z_0 = r) = \mathbb{P}(V_1 = T | Z_0 = r) = \frac{\lambda}{\lambda + (r \wedge K)\mu}$$

et

$$\mathbb{P}(Z_1 = r-1 | Z_0 = r) = \mathbb{P}(V_1 \neq T | Z_0 = r) = \frac{(r \wedge K)\mu}{\lambda + (r \wedge K)\mu},$$

ainsi que $\mathbb{P}(|Z_1 - r| \neq 1 | Z_0 = r) = 0$. Le taux de transition, voir la remarque 8.2.5, de r vers $r+1$ est égal à λ et de $r \geq 1$ vers $r-1$ à $(r \wedge K)\mu$. Enfin, en généralisant le lemme 8.1.5 à plusieurs variables exponentielles indépendantes, on obtient que conditionnellement à $V_1 = T$ (resp. $V_1 = S_i$), les variables $S_1 - V_1, \ldots, S_{r\wedge K} - V_1$ (resp. $T - V_1$ et $S_j - V_1$ pour $1 \leq j \leq r \wedge K$ et $j \neq i$) sont indépendantes et de loi exponentielle de paramètre μ (resp. exponentielle de paramètre λ pour $T - V_1$ et exponentielle de paramètre μ pour les autres). En particulier, les lois de $(Z_k, k \geq 2)$ et de $(V_k, k \geq 2)$ conditionnellement à Z_0, V_1 et Z_1, ne dépendent que de Z_1.

En itérant le raisonnement, on obtient que $(Z_n, n \geq 0)$ est une chaîne de Markov de matrice de transition $Q = (Q(i,j), i \geq 0, j \geq 0)$ définie par

$$Q(i,j) = \begin{cases} 0 & \text{si } |i-j| \neq 1, \\ \lambda/[\lambda + (i \wedge K)\mu] & \text{si } j = i+1, \\ (i \wedge K)\mu/[\lambda + (i \wedge K)\mu] & \text{si } i \geq 1 \text{ et } j = i-1. \end{cases}$$

Enfin, conditionnellement à $(Z_n, n \geq 0)$ les variables $(V_n, n \geq 1)$ sont indépendantes, et la loi de V_k est la loi exponentielle de paramètre $\lambda + (Z_{k-1} \wedge K)\mu$.

Si K est fini, alors la condition (8.3) est satisfaite car les taux de sauts sont bornés par $\lambda + K\mu$. Si K est infini, alors la condition (8.3) est satisfaite d'après l'exercice 8.1.4. Par construction, $X = (X_t, t \geq 0)$ est une chaîne de Markov homogène à temps continu de générateur infinitésimal A, défini par (9.1).

Vérifions qu'elle est irréductible. Soit $x \neq y \in \mathbb{N}$. Si $x < y$, on a $\prod_{i=0}^{y-x-1} A(x+i, x+i+1) > 0$ et si $x > y$, on a $\prod_{i=0}^{x-y-1} A(x-i, x-i-1) > 0$. La chaîne est irréductible d'après l'exercice 8.3.5. □

9.2 Étude des files à un serveur : $M/M/1$

Les inter-arrivées des clients suivent la loi exponentielle de paramètre λ, et les temps des services suivent la loi exponentielle de paramètre μ. Le temps moyen de service est $\mathbb{E}[S] = 1/\mu$ et le temps moyen entre deux arrivées de clients est $\mathbb{E}[T] = 1/\lambda$. On définit la densité de trafic par $\boxed{\rho = \lambda/\mu.}$ Elle correspond au temps moyen de service divisé par le temps moyen entre deux arrivées. D'après la proposition 8.4.3, λ représente le nombre moyen de personnes arrivant par unité de temps dans le système. De même μ peut également s'interpréter comme le nombre moyen de personnes servies par unité de temps par un serveur ayant une infinité de clients. Ainsi, la densité de trafic peut aussi s'interpréter comme le nombre moyen de personnes arrivant par unité de temps divisé par le nombre moyen de personnes servies par unité de temps.

On désire étudier le comportement de la file, $X = (X_t, t \geq 0)$, en temps long : nombre moyen de clients dans le système, temps d'attente d'un nouveau client,... Grâce aux théorèmes 8.3.7 et 8.3.12, les limites en temps long correspondent à des moyennes sous la probabilité invariante.

Dans le paragraphe 9.2.1 nous déterminons la probabilité invariante, π, appelée aussi probabilité stationnaire. Puis nous calculons la taille moyenne du système dans le régime stationnaire. Dans les paragraphes 9.2.2 et 9.2.3 nous calculons le temps moyen d'attente d'un client virtuel arrivant dans la file d'attente à l'instant t, avec t grand, ou du n-ième client, avec n grand.

9.2.1 Probabilité invariante

Proposition 9.2.1. *Il existe une (unique) probabilité invariante π pour la chaîne X si et seulement si $\rho < 1$. De plus, si $\rho < 1$, la probabilité invariante $\pi = (\pi_n, n \in \mathbb{N})$ est donnée par*

$$\pi_n = \rho^n(1-\rho), \quad n \in \mathbb{N}. \tag{9.2}$$

Remarquons que si Y est de loi π, alors $1 + Y$ suit une loi géométrique de paramètre $1 - \rho$. Enfin le cas $\rho \geq 1$ est abordé dans la remarque 9.5.2. La figure 9.1 représente des simulations de la chaîne X pour diverses valeurs de ρ.

Démonstration. La chaîne est irréductible. Elle possède donc au plus une probabilité invariante, π déterminée par $\pi A = 0$, où A est le générateur infinitésimal de X donné dans la proposition 9.1.1 avec $K = 1$. On obtient le système suivant : $-\lambda\pi_0 + \mu\pi_1 = 0$ et pour $k \geq 2$,

$$\lambda\pi_{k-2} - (\lambda + \mu)\pi_{k-1} + \mu\pi_k = 0.$$

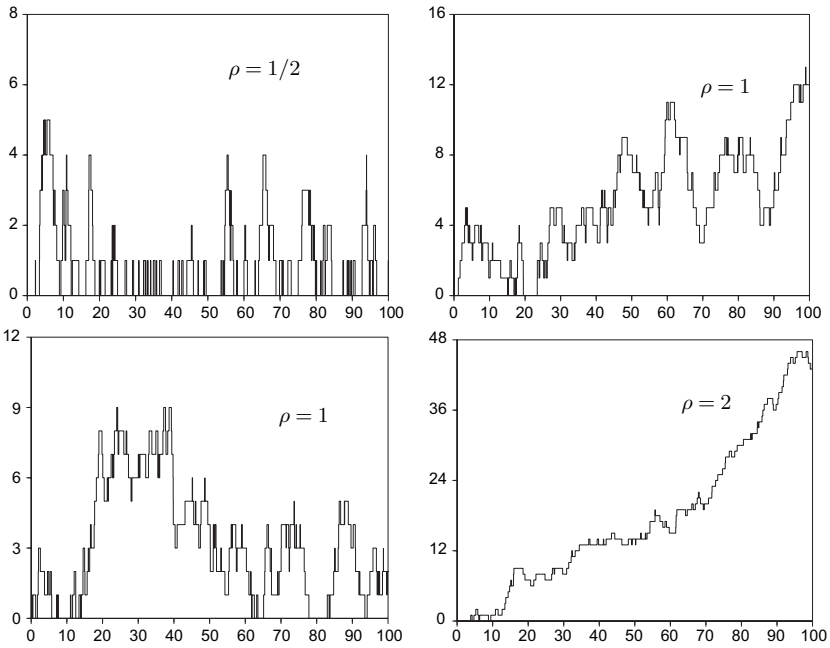

Fig. 9.1. Simulations d'une file $M/M/1$, $t \to X_t$, pour différentes valeurs de ρ, avec $\lambda = 1$

En sommant la première égalité et les égalités ci-dessus pour $k \leq n$ (ce qui revient à sommer les n premières colonnes de A puis à multiplier le résultat à gauche par π), il vient

$$0 = -\lambda\pi_0 + \mu\pi_1 + \sum_{k=2}^{n}(\lambda\pi_{k-2} - (\lambda+\mu)\pi_{k-1} + \mu\pi_k) = -\lambda\pi_{n-1} + \mu\pi_n.$$

On en déduit que $\pi_n = \rho\pi_{n-1}$, puis que $\pi_n = \rho^n\pi_0$. La suite $(\pi_n, n \in \mathbb{N})$ définit une probabilité si et seulement si $\sum_{n\geq 0}\pi_n = 1$, c'est-à-dire $\pi_0 \sum_{n\geq 0}\rho^n = 1$. Ceci n'est réalisable que si $\rho < 1$. Dans ce cas, on a pour $n \geq 0$, $\pi_n = \rho^n(1-\rho)$. □

Remarque 9.2.2. Le théorème 8.3.7 assure, si $\rho < 1$, que la suite des lois des variables X_t converge étroitement vers π quand t tend vers l'infini. Le calcul de $\mathbb{P}(X_t = j \mid X_0 = i)$ est difficile. On peut en trouver une expression dans [1], p. 89 et p. 92, ainsi que l'approximation

$$\mathbb{P}(X_t = j \mid X_0 = i) - \pi(j) \approx C(i,j)t^{-3/2}\,\mathrm{e}^{-\left(\sqrt{\mu}-\sqrt{\lambda}\right)^2 t},$$

où $C(i,j)$ est une constante qui dépend que de i, j, λ et μ. Remarquons que le taux de décroissance dans l'exponentielle est indépendant de i et j. La quantité $\left(\sqrt{\mu}-\sqrt{\lambda}\right)^{-2}$ est souvent appelée temps de relaxation du système. ◊

Proposition 9.2.3. *On suppose $\rho < 1$. En régime stationnaire, on a*

$$\mathbb{E}[X_t] = \sum_{n\geq 0} n\pi_n = \frac{\rho}{1-\rho} \quad \textit{et} \quad \mathrm{Var}(X_t) = \frac{\rho}{(1-\rho)^2}.$$

La figure 9.2 représente des simulations de l'évolution de la moyenne en temps du nombre d'individus dans le système. D'après le théorème ergodique, cette quantité converge vers $\sum_{n\geq 0} n\pi_n = \rho/(1-\rho)$ si $\rho < 1$. On peut démontrer qu'elle diverge si $\rho \geq 1$.

Démonstration. On a vu que si Y est de loi π, alors $1+Y$ suit une loi géométrique de paramètre $1-\rho$. Les résultats découlent alors du paragraphe A.2.1 (voir le tableau A.1 page 396). □

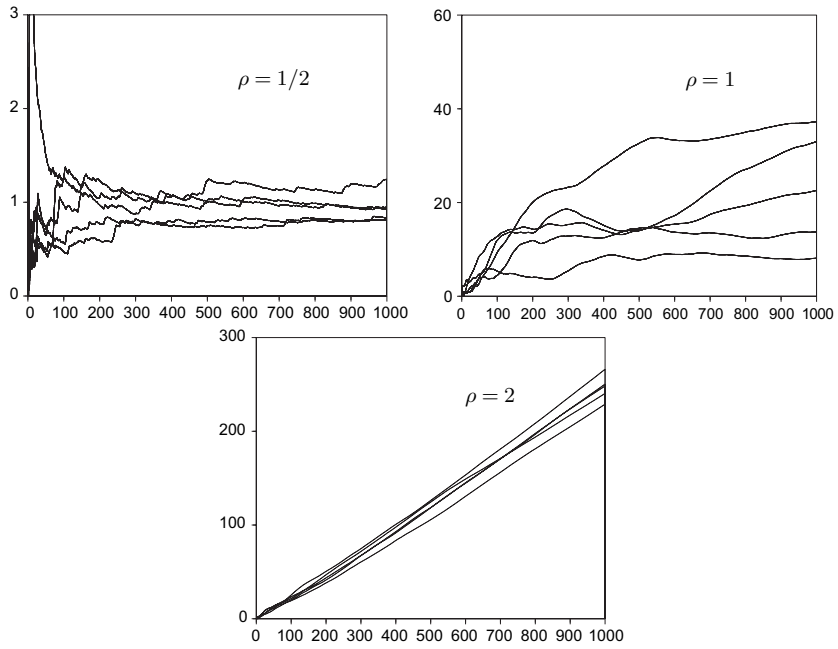

Fig. 9.2. Simulations de $t \to \dfrac{1}{t}\displaystyle\int_0^t X_s\, ds$, pour une file $M/M/1$ pour différentes valeurs de ρ, avec $\lambda = 1$

Exercice 9.2.4. Pour $\rho > 0$, calculer le temps moyen de repos du serveur par unité de temps, c'est-à-dire $\lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbf{1}_{\{X_s=0\}} ds$. ⧫

Exercice 9.2.5. Calculer, en régime stationnaire, la loi du premier temps de sortie d'un client de la file d'attente. Voir le paragraphe 9.4.2 pour plus d'information sur le processus de sortie. ⧫

9.2.2 Temps passé dans la file d'attente : client virtuel

On suppose $\rho < 1$. On considère un client virtuel arrivant à l'instant t dans la file d'attente. On note $W(t)$ le temps passé dans la file d'attente par ce client avant que ne débute son service, et $U(t)$, le temps total passé dans le système. Bien sûr on a $U(t) = W(t) + S$, où S, qui représente le temps de service du client virtuel, est une variable aléatoire indépendante de $W(t)$ de loi exponentielle de paramètre μ.

Proposition 9.2.6. *On suppose $\rho < 1$. Les variables aléatoires $(W(t), t \geq 0)$ convergent en loi quand t tend vers l'infini. Si W suit la loi limite, alors on a $\mathbb{P}(W = 0) = (1-\rho)$ et pour $s \geq 0$, $\mathbb{P}(W > s) = \rho\, \mathrm{e}^{-(\mu-\lambda)s}$. Enfin en régime stationnaire la loi de $W(t)$ est égale à la loi de W.*

Remarquons que $\mathbb{P}(W > s \mid W > 0) = \mathrm{e}^{-(\mu-\lambda)s}$. La loi de W conditionnellement à $\{W > 0\}$ est la loi exponentielle de paramètre $\mu - \lambda$. En résumé, quand un client virtuel arrive à l'instant t, t grand, alors avec probabilité $1-\rho$, il est servi tout de suite, et avec probabilité ρ, il doit attendre un temps aléatoire de loi exponentielle de paramètre $\mu - \lambda$.

Démonstration. Pour $X_t = 0$, on a $W(t) = 0$. Pour $X_t = k \geq 1$, le système comporte k clients avant l'arrivée du client virtuel. Le service de ce dernier débutera à l'instant $S_1 + S_2 + \ldots + S_k$, où S_1 représente le temps résiduel de service du premier client du système et S_i, $i \geq 2$, représente la durée de service du $i-1$-ème client dans la file d'attente. Par construction les variables $(S_i, 1 \leq k)$ sont indépendantes, les variables $(S_i, 2 \leq i \leq k)$ suivent des lois exponentielles de paramètre μ. Par la propriété sans mémoire des lois exponentielles, le temps résiduel de service, S_1 suit également une loi exponentielle de paramètre μ. Si on note ν_t la loi de X_t, on obtient pour $s \geq 0$,

$$\mathbb{P}(W(t) > s) = \sum_{k\geq 0} \mathbb{P}(W(t) > s, X_t = k) = \sum_{k\geq 1} \nu_t(k) \mathbb{P}\Big(\sum_{i=1}^{k} S_i > s\Big).$$

D'après le théorème 8.3.7, la chaîne $(X_t, t \geq 0)$ converge en loi vers une variable de loi π. En particulier, (ν_t, f) converge vers (π, f) pour toute fonction f bornée. Comme $\pi_n = \rho^n(1-\rho)$, on a donc

$$\begin{aligned}\lim_{t\to\infty}\mathbb{P}(W(t)>s) &= \sum_{k\geq 1}\pi_k\mathbb{P}\Big(\sum_{i=1}^{k}S_i>s\Big)\\ &= \sum_{k\geq 1}\rho^k(1-\rho)\,\frac{1}{(k-1)!}\int_s^\infty \mu^k u^{k-1}\,\mathrm{e}^{-\mu u}\;du\\ &= \rho(1-\rho)\mu\int_s^\infty \mathrm{e}^{\rho\mu u}\,\mathrm{e}^{-\mu u}\;du\\ &= \rho\,\mathrm{e}^{-(\mu-\lambda)s},\end{aligned}$$

où pour la deuxième égalité, on a utilisé que la somme de k variables exponentielles de même paramètre, μ, indépendantes suit une loi gamma de paramètre (μ, k).

Enfin on a $\mathbb{P}(W(t)=0)=\mathbb{P}(X_t=0)$, qui converge vers $\pi_0=(1-\rho)$ quand t tend vers l'infini. On en déduit que les fonctions de répartition de $(W(t), t\geq 0)$ convergent vers la fonction de répartition de la variable W, définie par $\mathbb{P}(W\leq s)=1-\mathbb{P}(W>s)=1-\rho\,\mathrm{e}^{-(\mu-\lambda)s}$ pour $s\geq 0$. Ceci implique la convergence en loi de la suite $(W(t), t\geq 0)$ vers W quand t tend vers l'infini.

Enfin, en régime stationnaire, on a $\nu_t=\pi$, ce qui assure que $W(t)$ a même loi que W. □

Exercice 9.2.7. On suppose $\rho>1$. Soit U une variable aléatoire exponentielle de paramètre $\mu-\lambda>0$.

1. Montrer, en utilisant la proposition 9.2.6 et les fonctions caractéristiques, que $(U(t), t\geq 0)$ converge en loi vers U.
2. Vérifier qu'en régime stationnaire $U(t)$ a même loi que U.
3. Montrer et interpréter les égalités suivantes :

$$\mathbb{E}[W]=\frac{1}{\mu}\,\frac{\rho}{1-\rho},\quad \mathbb{E}[U]=\frac{1}{\mu}\,\frac{1}{1-\rho}.$$

♦

9.2.3 Temps passé dans la file d'attente : client réel

Dans les phénomènes d'attente, il existe parfois des paradoxes. Ainsi le temps d'attente d'un client virtuel arrivant à l'instant t, $W(t)$, est en général différent du temps d'attente, W_n, du n-ième client, dit client réel, arrivé après l'instant initial. Par exemple dans le cas stationnaire, à t fixé, la loi du nombre de clients dans le système juste avant l'arrivée du client virtuel à l'instant t, noté X_{t^-}, est π, mais pour T aléatoire, la loi du nombre de clients dans le système juste avant l'instant T, X_{T^-} est a priori différente de π. En effet, considérons comme temps aléatoire T_1, le temps d'arrivée du premier client après l'instant initial. Entre 0 et T_1, si $X_0>0$, il existe une probabilité non nulle pour que

des clients aient terminé leurs services. Donc, si $X_0 > 0$, avec probabilité strictement positive on a $X_{T_1^-} < X_0$. Ainsi la loi de $X_{T_1^-}$ est différente de π.

Pour étudier la loi asymptotique de W_n, nous regardons l'évolution du système juste avant l'arrivée des nouveaux clients. On note $(T_i, i \geq 1)$ la suite des inter-arrivées des clients dans la file d'attente. Pour $n \geq 1$, on pose $\tau_n = \sum_{i=1}^n T_i$ le temps d'arrivée du n-ième client. Le nombre de clients dans le système juste avant l'arrivée du client n est $X_{(n)} = X_{\tau_n^-} = X_{\tau_n} - 1$. On suppose que l'instant $t = 0$ correspond à l'arrivée d'un nouveau client, de sorte que $X_0 \geq 1$, et on pose $X_{(0)} = X_0 - 1$.

Proposition 9.2.8. *La suite $(X_{(n)}, n \geq 0)$ est une chaîne de Markov à temps discret homogène à valeurs dans $\mathbb{N}$.*

Démonstration. Soit $(Z_n, n \geq 0)$ la chaîne trace associée à la chaîne à temps continu $(X_t, t \geq 0)$. Par hypothèse, on a $Z_0 > 0$. On pose $Z_{(0)} = Z_0 - 1$ et $R_1 = \inf\{k \geq 1\,; Z_k = Z_{k-1} + 1\}$, le nombre d'étapes avant l'arrivée d'un nouveau client. Pour $n \geq 1$, on considère, pour la chaîne trace, la date d'arrivée du n-ième client, $V_n \in \mathbb{N}$, la taille du système juste avant son arrivée, $Z_{(n)}$, et le temps d'inter-arrivée entre ce client et le client suivant, $R_{n+1} \in \mathbb{N}^*$. Plus précisément, on pose $V_0 = 0$ et pour tout $n \geq 1$, on définit par récurrence $V_n = \sum_{k=1}^n R_k$, $Z_{(n)} = Z_{V_n} - 1$ et $R_{n+1} = \inf\{k \geq 1\,; Z_{k+V_n} = Z_{k+V_n-1} + 1\}$. Par construction on a $Z_{(n)} = X_{(n)}$ pour tout $n \in \mathbb{N}$.

Nous montrons maintenant que $(Z_{(n)}, n \geq 0)$ est une chaîne de Markov. Par construction, pour tout $n \in \mathbb{N}^*$, on a p.s. $Z_{(n)} \leq Z_{(n-1)} + 1$, avec égalité si aucun client n'a fini son service entre l'arrivée du $(n-1)$-ième et du n-ième client. Entre les instants $V_{n-1} + 1$ et $V_n - 1$, on n'observe pour la chaîne trace que des sorties de clients. On en déduit que pour $k \in \{0, R_n - 1\}$ on a $Z_{V_{n-1}+k} = Z_{(n-1)} + 1 - k$, ainsi que $R_n = Z_{(n-1)} + 2 - Z_{(n)}$ et donc $V_n = Z_{(0)} + 2n - Z_{(n)}$. On en déduit donc que pour un chemin donné, $n \in \mathbb{N}^*$, $x_0, \ldots, x_n$ avec $x_{k+1} \leq x_k + 1$ et $0 \leq k \leq n-1$, l'événement $\{Z_{(0)} = x_0, \ldots, Z_{(n)} = x_n\}$ détermine complètement les valeurs de $(Z_k, 0 \leq k \leq x_0 + 2n - x_n)$. En particulier, si on pose $v_{n-1} = x_0 + 2(n-1) - x_{n-1}$ et $r_n = x_{n-1} + 2 - x_n$, l'événement $\{Z_{(n)} = x_n, \ldots, Z_{(0)} = x_0\}$ est égal à l'intersection de

$$\{Z_{v_{n-1}+r_n} = x_n + 1, (Z_{v_{n-1}+k} = x_{n-1} + 1 - k, 1 \leq k \leq r_n - 1)\}$$

et de $\{Z_{(n-1)} = x_{n-1}, \ldots, Z_{(0)} = x_0\}$. Après avoir remarqué que sur l'événement $\{Z_{(n-1)} = x_{n-1}, \ldots, Z_{(0)} = x_0\}$, on a $V_{n-1} = v_{n-1}$ et $Z_{v_{n-1}} = x_{n-1} + 1$, on déduit de la proposition 1.1.7 appliquée à la chaîne de Markov Z à l'instant v_{n-1}, que pour $n \geq 1$, on a

$$\begin{aligned}
&\mathbb{P}(Z_{(n)} = x_n | Z_{(0)} = x_0, \ldots, Z_{(n-1)} = x_{n-1}) \\
&\quad = \mathbb{P}(Z_{r_n} = x_n + 1, (Z_k = x_{n-1} + 1 - k, 1 \leq k \leq r_n - 1) | Z_0 = x_{n-1} + 1) \\
&\quad = \mathbb{P}(Z_{(1)} = x_n | Z_{(0)} = x_{n-1}).
\end{aligned}$$

Ceci assure que $(Z_{(n)}, n \geq 0)$ est une chaîne de Markov homogène. □

Proposition 9.2.9. *La chaîne de Markov $(X_{(n)}, n \geq 0)$ est irréductible et apériodique. Elle possède une (unique) probabilité invariante si et seulement si $\rho < 1$. La probabilité invariante est alors la probabilité π définie par (9.2).*

Démonstration. On reprend les notations de la démonstration précédente. Déterminons la matrice de transition P.

Pour $k \geq 1$, $l \geq 0$, on remarque que si $Z_{(0)} = k + l$ et $Z_{(1)} = k$, alors à l'instant $t = 0$, $k + l + 1$ clients sont dans le système. De plus, quand arrive un nouveau client, $l + 1$ clients ont été servis et ont quitté le système, et le service du $l + 2$-ième client a débuté, mais n'est pas terminé. Les temps de service $S_1, \ldots, S_{l+1}, S_{l+2}$ de ces $l + 2$ clients sont des variables indépendantes de loi exponentielle de paramètre μ, indépendantes du temps d'arrivée, T du nouveau client. On en déduit que pour $k \geq 1$, $l \geq 0$,

$$\mathbb{P}(Z_{(1)} = k \mid Z_{(0)} = k + l) = \mathbb{P}\left(\sum_{i=1}^{l+1} S_i < T \leq \sum_{i=1}^{l+2} S_i\right).$$

En utilisant l'indépendance de T avec les variables $S_1, \ldots, S_{l+2}$, on déduit de (8.4)

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{l+1} S_i < T \leq \sum_{i=1}^{l+2} S_i\right) &= \mathbb{P}\left(\sum_{i=1}^{l+1} S_i < T\right) - \mathbb{P}\left(\sum_{i=1}^{l+2} S_i < T\right) \\
&= \mathbb{E}\Big[\,\mathrm{e}^{-\lambda \sum_{i=1}^{l+1} S_i}\,\Big] - \mathbb{E}\Big[\,\mathrm{e}^{-\lambda \sum_{i=1}^{l+2} S_i}\,\Big] \\
&= \mathbb{E}\Big[\,\mathrm{e}^{-\lambda S_1}\,\Big]^{l+1} - \mathbb{E}\Big[\,\mathrm{e}^{-\lambda S_1}\,\Big]^{l+2} \\
&= \left(\frac{\mu}{\lambda+\mu}\right)^{l+1} - \left(\frac{\mu}{\lambda+\mu}\right)^{l+2}.
\end{aligned}$$

On obtient pour $k \geq 1$, $l \geq 0$,

$$\mathbb{P}(Z_{(1)} = k \mid Z_{(0)} = k + l) = \frac{\lambda \mu^{l+1}}{(\lambda+\mu)^{l+2}}.$$

Pour $l = -1$, un nouveau client arrive avant que le premier service soit terminé. On a alors $\mathbb{P}(Z_{(1)} = k \mid Z_{(0)} = k-1) = \lambda/(\lambda+\mu)$. Comme $Z_{(1)} \leq Z_{(0)} + 1$, on en déduit que $\mathbb{P}(Z_{(1)} = k \mid Z_{(0)} = k + l) = 0$ si $-k \leq l < -1$.

Enfin si $Z_{(0)} = l$, pour $l \geq 0$, et $Z_{(1)} = 0$, cela signifie que $l+1$ clients sont dans le système à l'instant $t = 0$, et qu'ils ont tous été servis avant l'arrivée du nouveau client. On déduit de (8.4) que

$$\mathbb{P}(Z_{(1)} = 0 \mid Z_{(0)} = l) = \mathbb{P}\left(\sum_{i=1}^{l+1} S_i < T\right) = \mathbb{E}[\mathrm{e}^{-\lambda \sum_{i=1}^{l+1} S_i}] = \frac{\mu^{l+1}}{(\lambda+\mu)^{l+1}}.$$

On en déduit que les termes $P(i,j)$ de la matrice de transition sont nuls pour $j > i+1$, et

$$P(i,j) = \frac{\mu^{i-j+1}}{(\lambda+\mu)^{i-j+1}} \left[\mathbf{1}_{\{j=0\}} + \frac{\lambda}{\lambda+\mu} \mathbf{1}_{\{j>0\}} \right], \quad \text{pour } i+1 \geq j \geq 0,\ i \geq 0.$$

Comme pour tout $i \in \mathbb{N}$, on a $P(i,i+1) = \lambda/(\lambda+\mu) > 0$ et $P(i,0) > 0$, on en déduit que la chaîne est irréductible et apériodique.

Supposons $\rho < 1$, et vérifions que π définie par (9.2) est une probabilité invariante. Soit $j \geq 1$, on a

$$\begin{aligned}
\sum_{i\geq 0} \pi_i P(i,j) &= \sum_{i \geq j-1} \rho^i (1-\rho) \frac{\lambda \mu^{i-j+1}}{(\lambda+\mu)^{i-j+2}} \\
&= \rho^{j-1}(1-\rho) \frac{\lambda}{\lambda+\mu} \sum_{l\geq 0} \rho^l \frac{\mu^l}{(\lambda+\mu)^l} \\
&= \rho^j (1-\rho) \\
&= \pi_j,
\end{aligned}$$

où on a posé $l = i-j+1$ dans la deuxième égalité. En sommant ces égalités sur $j \geq 1$, il vient $\sum_{i\geq 0} \pi_i (1 - P(i,0)) = 1 - \pi_0$, soit $\sum_{i\geq 0} \pi_i P(i,0) = \pi_0$. La probabilité π est donc une probabilité invariante. Comme la chaîne est irréductible, c'est la seule.

Supposons $\rho \geq 1$. Si la chaîne trace, Z, possédait une probabilité invariante, alors d'après l'exercice 8.3.11, comme les taux de sauts sont minorés par λ, la chaîne X possèderait également une probabilité invariante. Comme ce n'est pas le cas d'après la proposition 9.2.1, on en déduit que la chaîne trace ne possède pas de probabilité invariante. Rappelons que V_k désigne le temps d'arrivée du k-ième client pour la chaîne trace. Remarquons que si $Z_{(k)} = 0$ pour $k \geq 1$, alors il existe $j \geq 1$ tel que $V_k = j$ et $Z_{j-1} = 0$. On en déduit que

$$\sum_{k=1}^{n} \mathbf{1}_{\{Z_{(k)}=0\}} \leq \sum_{j=0}^{V_n - 1} \mathbf{1}_{\{Z_j=0\}}.$$

On a vu que $V_n = Z_{(0)} + 2n - Z_{(n)}$. Ceci assure que $\limsup_{n\to\infty} V_n/n \leq 2$, et donc

$$0 \leq \limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{Z_{(k)}=0\}} \leq \limsup_{n\to\infty} (V_n/n) \limsup_{n\to\infty} \frac{1}{V_n} \sum_{j=0}^{V_n-1} \mathbf{1}_{\{Z_j=0\}}.$$

Comme $\lim_{n\to\infty} V_n = +\infty$, le théorème ergodique 8.3.12 implique que le terme de droite est nul. Donc on a p.s. $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{Z_{(k)}=0\}} = 0$. Comme la chaîne $(Z_{(n)}, n \geq 1)$ est irréductible, ceci implique, d'après la proposition 8.3.9, qu'elle ne possède pas de probabilité invariante. □

Donc, pour $\rho < 1$, si $X_0 - 1$ suit la loi π, alors la chaîne $(X_{(n)}, n \geq 0)$ est stationnaire. Dans ce régime stationnaire, le nombre de clients dans le système juste avant l'arrivée d'un nouveau client est distribué suivant la probabilité π. On retrouve la même loi que pour le nombre de clients dans le système avant l'arrivée d'un client virtuel. Il ne s'agit toutefois pas des mêmes régimes stationnaires. Dans le cas du client virtuel, on regarde le régime stationnaire qui apparaît après un temps long, dans le cas du client réel, on regarde le régime stationnaire qui apparaît après un grand nombre d'arrivées de clients (le temps long est ici aléatoire).

On considère le n-ième client, et on note W_n le temps passé dans la file d'attente par ce client avant que ne débute son service. La démonstration de la proposition suivante est analogue à celle de la proposition 9.2.6 pour le client virtuel.

Proposition 9.2.10. *On suppose $\rho < 1$. La suite $(W_n, n \geq 0)$ converge en loi vers W, définie dans la proposition 9.2.6. Si $X_0 - 1$ suit la loi π, alors la chaîne $(X_{(n)}, n \geq 0)$ est stationnaire et W_n a même loi que W.*

On retrouve donc les mêmes lois pour le client réel et pour le client virtuel en ce qui concerne les temps d'attente et les temps passés dans le système. Ces résultats sont préservés même si les temps de services sont des variables aléatoires indépendantes de même loi quelconque, pourvu que le processus des arrivées soit un processus de Poisson. Cette propriété est connue sous le nom de propriété PASTA (« Poissons Arrivals See Time Average »). En revanche, les lois asymptotiques pour le temps d'attente du client réel et du client virtuel sont en général différentes si le processus d'arrivée n'est plus un processus de Poisson.

9.3 Étude des files à K serveurs : $M/M/K$

On considère une file d'attente avec $K \in \mathbb{N}$ serveurs indépendants. Le processus d'arrivée est un processus de Poisson de paramètre λ. Les temps de services sont des variables exponentielles de paramètre μ. On définit la densité de trafic par $\boxed{\rho = \lambda/(K\mu)}$. Attention, certains auteurs considèrent que la densité de trafic est $\rho = \lambda/\mu$.

9.3.1 Probabilité invariante

Proposition 9.3.1. *Il existe une (unique) probabilité invariante π pour la chaîne X si et seulement si $\rho < 1$. De plus, si $\rho < 1$, la probabilité invariante $\pi = (\pi_n, n \in \mathbb{N})$ est donnée par*

$$\pi_n = \pi_0 \rho^n \frac{K^n}{n!} \quad \text{si } n \leq K, \qquad \pi_n = \pi_0 \rho^n \frac{K^K}{K!} \quad \text{si } n \geq K, \tag{9.3}$$

et π_0 est déterminé par $\sum_{k \geq 0} \pi_n = 1$.

Démonstration. La chaîne est irréductible. Elle possède au plus une probabilité invariante, π, déterminée par $\pi A = 0$, où A est le générateur infinitésimal de X donné dans la proposition 9.1.1. On obtient le système suivant : $-\lambda\pi_0 + \mu\pi_1 = 0$ et pour $k \geq 2$,

$$\lambda\pi_{k-2} - (\lambda + ((k-1) \wedge K)\mu)\pi_{k-1} + (k \wedge K)\mu\pi_k = 0.$$

En sommant la première égalité et les égalités ci-dessus pour $k \leq n$ (ce qui revient à sommer les n premières colonnes de A puis à multiplier le résultat à gauche par π), il vient

$$\begin{aligned} 0 &= -\lambda\pi_0 + \mu\pi_1 + \sum_{k=2}^{n} (\lambda\pi_{k-2} - (\lambda + ((k-1) \wedge K)\mu)\pi_{k-1} + (k \wedge K)\mu\pi_k) \\ &= -\lambda\pi_{n-1} + (n \wedge K)\mu\pi_n. \end{aligned}$$

On en déduit que $\pi_n = \rho \dfrac{K}{n \wedge K}\pi_{n-1}$, puis que $\pi_n = \pi_0\rho^n \displaystyle\prod_{k=1}^{n} \frac{K}{k \wedge K}$, ce qui donne (9.3). La suite $(\pi_n, n \in \mathbb{N})$ définit une probabilité si et seulement si $\sum_{n\geq 0} \pi_n = 1$. Ceci n'est réalisable que si $\rho < 1$. La constante π_0 est alors définie par $\pi_0 \sum_{k\geq 0} \rho^n \prod_{k=1}^{n} \frac{K}{k\wedge K} = 1$. □

Donnons maintenant quelques résultats sur la file d'attente en régime stationnaire.

Proposition 9.3.2. *Soit $\rho < 1$. On suppose que X_0 est distribué suivant la probabilité invariante.*

1. *La probabilité pour que tous les serveurs soient occupés est $\pi_0 \dfrac{\rho^K}{1-\rho}\dfrac{K^K}{K!}$.*
2. *Le nombre moyen de serveurs occupés est $K\rho$.*
3. *Le nombre moyen de clients dans le système est*

$$\mathbb{E}[X_t] = \sum_{n\geq 0} n\pi_n = K\rho + \pi_0 \frac{\rho^{K+1}}{(1-\rho)^2}\frac{K^K}{K!}.$$

Démonstration. 1. La probabilité pour que tous les serveurs soient occupés est $\mathbb{P}(X_t \geq K)$. On a donc

$$\mathbb{P}(X_t \geq K) = \sum_{n\geq K} \pi_n = \sum_{n\geq K} \pi_0\rho^n \frac{K^K}{K!} = \pi_0 \frac{\rho^K}{1-\rho}\frac{K^K}{K!}.$$

2. Le nombre de serveurs occupés est $X_t \wedge K$. On a, en utilisant $\pi_n = \rho\dfrac{K}{n \wedge K}\pi_{n-1}$ pour $n \geq 1$,

$$\mathbb{E}[X_t \wedge K] = \sum_{n\geq 1}(n\wedge K)\pi_n = \rho K \sum_{n\geq 1}\pi_{n-1} = K\rho.$$

3. On a, en utilisant le calcul précédent,

$$\begin{aligned}\mathbb{E}[X_t] &= \sum_{n\geq 0} n\pi_n \\ &= \sum_{n=1}^{K} n\pi_n + \sum_{n\geq K+1} K\pi_n + \sum_{n\geq K+1}(n-K)\pi_n \\ &= K\rho + \pi_0\rho^K\frac{K^K}{K!}\sum_{n\geq K+1}(n-K)\rho^{n-K}.\end{aligned}$$

Remarquons que

$$\sum_{n\geq K+1}(n-K)\rho^{n-K} = \sum_{n\geq 1} n\rho^n = \frac{\rho}{1-\rho}\sum_{n\geq 1} n\rho^{n-1}(1-\rho) = \frac{\rho}{(1-\rho)^2},$$

où pour la dernière égalité on a reconnu dans la somme l'espérance d'une variable aléatoire de loi géométrique de paramètre $1-\rho$. On en déduit que $\mathbb{E}[X_t] = K\rho + \pi_0\rho^K\frac{K^K}{K!}\frac{\rho}{(1-\rho)^2}$. □

9.3.2 Temps passé dans la file d'attente : client virtuel

On suppose $\rho < 1$. Comme dans le paragraphe 9.2.2, on considère un client virtuel arrivant à l'instant t dans la file d'attente. On note $W(t)$ le temps passé dans la file d'attente par ce client avant que ne débute son service, et $U(t)$, le temps total passé dans le système. Bien sûr on a $U(t) = W(t) + S$, où S est une variable aléatoire indépendante de $W(t)$ de loi exponentielle de paramètre μ.

Proposition 9.3.3. *On suppose $\rho < 1$. Les variables aléatoires $(W(t), t \geq 0)$ convergent en loi quand t tend vers l'infini. Si W suit la loi limite, alors on a $\mathbb{P}(W = 0) = 1 - a$, avec $a = \pi_0\frac{\rho^K}{1-\rho}\frac{K^K}{K!}$ et pour $s \geq 0$, $\mathbb{P}(W > s) = a\,\mathrm{e}^{-(K\mu-\lambda)s}$. Enfin en régime stationnaire la loi de $W(t)$ est égale à la loi de W.*

On retrouve en particulier le résultat 1. de la proposition 9.3.2. En effet $\{W > 0\}$ correspond bien au fait que tous les serveurs sont occupés.

Remarquons encore que la loi de W conditionnellement à $\{W > 0\}$ est la loi exponentielle de paramètre $(K\mu - \lambda)$. En résumé, quand un client virtuel arrive à l'instant t, pour t grand, alors avec probabilité $1 - a$, il est servi tout de suite, et avec probabilité a, il doit attendre un temps aléatoire de loi exponentielle de paramètre $K\mu - \lambda$ avant que son service débute. La figure 9.3 représente l'évolution de $\mathbb{E}[W]$ et de $\mathbb{P}(W > t)$ quand K varie, pour deux valeurs de μ.

Remarque 9.3.4. Comme pour les files d'attente $M/M/1$, on obtient les mêmes résultats si on regarde les limites en loi des temps d'attente dans la file pour un client virtuel arrivant à l'instant t, t grand, ou du n-ième client réel, n grand. ◊

Démonstration de la proposition 9.3.3. Pour $X_t \leq K-1$, il reste au moins un serveur libre. Le service du client virtuel débute immédiatement. On a alors $W(t) = 0$.

Pour $X_t = k \geq K$, le système comporte k clients avant l'arrivée du client virtuel. On note V_1 le premier temps de sortie de la file d'attente d'un client après l'instant t : $V_1 = \min\{S_i, 1 \leq i \leq K\}$, où la variable aléatoire S_i désigne le temps résiduel de service du client au guichet i. Les variables S_i sont indépendantes et, grâce à la propriété sans mémoire des lois exponentielles, de même loi exponentielle de paramètre μ. Donc V_1 suit la loi exponentielle de paramètre $K\mu$. On note V_j le temps entre la $(j-1)$-ième et la j-ème sortie d'un client de la file d'attente après l'instant t. Remarquons que le serveur

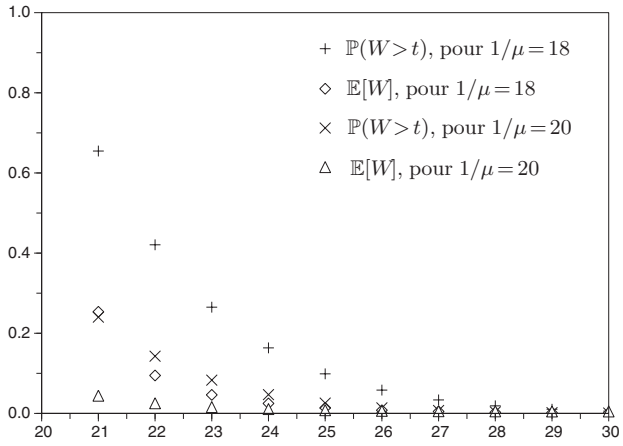

Fig. 9.3. Évolution de $\mathbb{E}[W]$ et de $\mathbb{P}(W > t)$, où $t = 3$ minutes, en fonction de $K \in \{21, \ldots, 30\}$, avec des arrivées de clients toutes les minutes en moyenne ($\lambda = 1$), et des services de 20 et 18 minutes en moyenne

libéré lors de la j-ième sortie d'un client commence à servir un des $k-K$ clients qui attendaient à l'instant t. On en déduit que le service du nouveau client débutera après le temps

$$V_1 + V_2 + \ldots + V_{k-K+1}.$$

En utilisant le caractère sans mémoire des lois exponentielles, on montre que les variables aléatoires $(V_j, 1 \leq j \leq k-K+1)$ sont indépendantes et de même loi exponentielle de paramètre $K\mu$. La loi de $V_1 + V_2 + \ldots + V_{k-K+1}$ est donc une loi gamma de paramètre $(K\mu, k-K+1)$. Si on note ν_t la loi de X_t, on obtient pour $s \geq 0$,

$$\begin{aligned}
\mathbb{P}(W(t) > s) &= \sum_{k \geq K} \mathbb{P}(W(t) > s, X_t = k) \\
&= \sum_{k \geq K} \nu_t(k) \mathbb{P}\left(V_1 + \ldots + V_{k-K+1} > s\right) \\
&= \sum_{k \geq K} \nu_t(k) \int_s^{+\infty} \frac{1}{(k-K)!} (K\mu)^{k-K+1} r^{k-K} \, \mathrm{e}^{-K\mu r} \, dr.
\end{aligned}$$

D'après le théorème 8.3.7, la chaîne $(X_t, t \geq 0)$ converge en loi vers une variable de la loi π. En particulier, (ν_t, f) converge vers (π, f) pour toute fonction f bornée. On a donc

$$\begin{aligned}
\lim_{t \to \infty} \mathbb{P}(W(t) > s) &= \sum_{k \geq K} \pi_k \int_s^{+\infty} \frac{1}{(k-K)!} (K\mu)^{k-K+1} r^{k-K} \, \mathrm{e}^{-K\mu r} \, dr \\
&= \int_s^{+\infty} \sum_{k \geq K} \pi_0 \rho^k \frac{K^K}{K!} \frac{1}{(k-K)!} (K\mu)^{k-K+1} r^{k-K} \, \mathrm{e}^{-K\mu r} \, dr \\
&= \pi_0 \rho^K \frac{K^K}{K!} K\mu \int_s^{+\infty} \mathrm{e}^{-(K\mu - \lambda)r} \, dr \\
&= \pi_0 \frac{\rho^K}{1-\rho} \frac{K^K}{K!} \, \mathrm{e}^{-(K\mu-\lambda)s}.
\end{aligned}$$

On en déduit que $a = \mathbb{P}(W > 0) = \pi_0 \dfrac{\rho^K}{1-\rho} \dfrac{K^K}{K!}$ et $\mathbb{P}(W = 0) = 1 - a$.

Enfin, en régime stationnaire, on a $\nu_t = \pi$, ce qui assure que $W(t)$ a même loi que W. □

Exercice 9.3.5. Montrer que le temps moyen d'attente dans la file est en régime stationnaire :

$$\mathbb{E}[W] = \frac{1}{K\mu}\,\pi_0 \frac{\rho^K}{(1-\rho)^2}\frac{K^K}{K!}.$$

♦

Exercice 9.3.6. On considère la file $M/M/2$ de paramètre (λ, μ).

1. Montrer que $\pi_0 = \dfrac{1-\rho}{1+\rho}$ et qu'en régime stationnaire $\mathbb{E}[X_t] = \dfrac{2\rho}{1-\rho^2}$.
2. Montrer que

$$\mathbb{P}(W=0) = 1 - \frac{2\rho^2}{1+\rho}, \quad \mathbb{E}[W] = \frac{1}{\mu}\,\frac{\rho^2}{1-\rho^2}, \quad \mathbb{E}[U] = \frac{1}{\mu}\,\frac{1}{1-\rho^2}.$$

3. Comparer ces résultats avec une file $M/M/1$ de paramètre $(\lambda, 2\mu)$. Quelle file est préférable pour le client ? Quel critère prendre en compte ?

♦

Exercice 9.3.7. On considère la file $M/M/\infty$, qui comporte une infinité de serveurs.

1. Vérifier, grâce à l'exercice 8.1.4, que le processus $(X_t, t \geq 0)$, où X_t désigne le nombre de personnes dans le système, est une chaîne de Markov à temps continu.
2. Vérifier que la chaîne est homogène irréductible.
3. Calculer et reconnaître la probabilité invariante de la file $M/M/\infty$.
4. Quel est le nombre moyen de clients dans le système en régime stationnaire ?

♦

Exercice 9.3.8. Le but de cet exercice est l'étude d'une file d'attente avec deux serveurs de caractéristiques différentes.

On considère une file d'attente avec deux serveurs A et B. On suppose que les temps de service du serveur A (resp. B) sont des variables aléatoires exponentielles de paramètres μ_A (resp. μ_B), avec $\mu_A \geq \mu_B > 0$. On suppose que le processus d'arrivée est un processus de Poisson de paramètre $\lambda > 0$.

Si les deux serveurs sont libres, on suppose que le client qui arrive choisit le serveur A, qui est en moyenne le plus rapide. Si un seul serveur est libre, on suppose que le client qui arrive va directement à ce serveur. On note X_t l'état du système à l'instant t. Si $X_t = 0$, les deux serveurs sont libres, si $X_t \geq 2$, les deux serveurs sont occupés. Si un seul serveur est occupé, on distingue le cas où A est occupé, on note alors $X_t = A$, et le cas où B est occupé, on note alors $X_t = B$. On note $E = \{0, A, B, 2, \ldots\}$ l'ensemble des valeurs possibles des états du système.

1. Montrer que $X = (X_t, t \geq 0)$ est une chaîne de Markov à temps continu sur E. Donner son générateur infinitésimal et la matrice de transition de la chaîne trace.
2. On pose $\rho = \lambda/(\mu_A + \mu_B)$. On cherche une probabilité invariante $\pi = (\pi_0, \pi_A, \pi_B, \pi_2, \ldots)$ de la chaîne. Expliquer pourquoi si elle existe, alors elle est unique. Montrer que si on pose $\pi_1 = \pi_A + \pi_B$, alors on a $\pi_n = \rho\pi_{n-1}$ pour tout $n \geq 2$ et donc $\pi_n = \rho^{n-1}\pi_1$. Vérifier également que
$$\pi_1 = \pi_0 \frac{1}{1+2\rho} \frac{\lambda}{\mu_A\mu_B} (\lambda + \mu_B).$$
3. En déduire qu'il existe une probabilité invariante si et seulement si $\rho < 1$. Vérifier alors que
$$\frac{1}{\pi_1} = \frac{1}{1-\rho} + (1+2\rho)\frac{\mu_A\mu_B}{\lambda(\lambda+\mu_B)},$$
et
$$\frac{1}{\pi_0} = 1 + \frac{1}{1+2\rho}\frac{1}{1-\rho}\frac{\lambda(\lambda+\mu_B)}{\mu_A\mu_B}.$$
4. On note $\tilde{X}_t$ le nombre de personnes dans le système : $\tilde{X}_t = 1$ si $X_t = A$ ou $X_t = B$, et $\tilde{X}_t = X_t$ sinon. Vérifier que, en régime stationnaire, $\mathbb{E}[\tilde{X}_t] = \dfrac{\pi_1}{(1-\rho)^2}$. En déduire que, à $\mu_A + \mu_B = 2\mu$ constant (taux de service moyen constant), le nombre moyen de personnes dans le système est minimal, de valeur $N_1(\rho)$, pour
$$\mu_B = \lambda\left(\sqrt{1+\frac{1}{\rho}} - 1\right), \quad \text{et} \quad \mu_A = \mu_B\sqrt{1+\frac{1}{\rho}}.$$
5. Soit $N_2(\rho)$ le nombre moyen de personnes dans le système en régime stationnaire pour une file $M/M/2$ de taux de service μ, taux d'arrivée λ et $\rho = \lambda/2\mu$. Vérifier que quand ρ tend vers 1, la différence $N_2(\rho) - N_1(\rho)$ converge vers une limite finie. En déduire que la différence relative entre la file $M/M/2$ et la file optimisée est négligeable quand ρ est proche de 1. On pourra consulter l'ouvrage [6] pour plus de résultats dans cette direction.

♦

9.4 Réseaux de Jackson

9.4.1 Modèle et propriétés

Les réseaux de Jackson introduits en 1957 sont des réseaux constitués de K files d'attentes, chacune associée à un seul serveur, avec plusieurs entrées et plusieurs sorties. Nous reprenons la présentation de Bougerol [3]. Les clients de

la file d'attente i, une fois leur service terminé, se dirigent vers la file d'attente j avec probabilité $p_{i,j}$ et sortent du système avec probabilité β_i où

$$\beta_i + \sum_{j=1}^{K} p_{i,j} = 1.$$

Des clients extérieurs au système arrivent dans la file d'attente i suivant un processus de Poisson de paramètre α_i. Les services fournis par le serveur i sont des variables aléatoires exponentielles indépendantes de paramètre μ_i et indépendantes du processus d'arrivée dans la file i des clients extérieurs. Enfin les temps de service et les processus d'arrivée des clients extérieurs sont indépendants d'un serveur à l'autre.

On note $X_t^{(i)}$ le nombre de personnes dans la file d'attente i à l'instant t, y compris le client au guichet i. L'état du système est entièrement décrit par le vecteur $X_t = (X_t^{(1)}, \ldots, X_t^{(K)})$ à valeurs dans $\mathbb{N}^K$. La proposition suivante généralise la proposition 9.1.1, et sa démonstration est similaire.

Proposition 9.4.1. *Le processus $(X_t, t \geq 0)$ est une chaîne de Markov à temps continu de générateur infinitésimal A dont les termes non nuls hors de la diagonale sont*

$$\begin{aligned} A(n, n + e_i) &= \alpha_i, \\ A(n, n - e_i) &= \beta_i \mu_i \text{ si } n_i > 0, \\ A(n, n - e_i + e_j) &= p_{i,j} \mu_i \text{ si } i \neq j \text{ et } n_i > 0, \end{aligned}$$

où $n = (n_1, \ldots, n_K) \in \mathbb{N}^K$ et e_i est le i-ème vecteur de la base canonique de $\mathbb{R}^K$ ($e_i = (e_i^{(1)}, \ldots, e_i^{(K)})$ avec $e_i^{(l)} = 0$ si $l \neq i$ et $e_i^{(i)} = 1$).

S'il existe i et j tels que $\alpha_i > 0$ et $\beta_j > 0$, alors on parle de réseaux de Jackson ouverts. Sinon on parle de réseaux de Jackson fermés ou de réseaux de Gordon-Newell. Dans ce qui suit on considère les réseaux de Jackson ouverts.

La chaîne est irréductible dès que les deux conditions suivantes sont satisfaites :

1. Pour tout j, il existe $m \in \mathbb{N}^*$, $i, i_1, \ldots, i_m$ tels que $\alpha_i p_{i,i_1} \ldots p_{i_m,j} > 0$.
2. Pour tout i, il existe $m \in \mathbb{N}^*$, $i_1, \ldots, i_m, j$ tels que $p_{i,i_1} \ldots p_{i_m,j} \beta_j > 0$.

La première condition signifie que pour tout serveur j, il existe une probabilité strictement positive qu'un client entre en i, puis se dirige vers les serveurs $i_1, \ldots$ et enfin j. La deuxième condition assure que pour tout serveur i, il existe une probabilité strictement positive qu'un client sorte de i, se dirige vers les serveurs $i_1, \ldots$, et enfin j d'où il sort du système. Il est clair que ces deux conditions impliquent que de tout état on peut rejoindre tout autre état avec probabilité strictement positive.

On suppose dorénavant que les conditions 1 et 2 sont satisfaites. Supposons que le régime soit stationnaire. Le flux entrant dans la file i, i.e. le nombre

de clients entrant dans la file i par unité de temps, et le flux sortant, i.e. le nombre de clients sortant de la file par unité de temps, de cette même file sont égaux. On note λ_i le flux (entrant ou sortant) du serveur i. Ce raisonnement intuitif conduit aux formules dites « équation de trafic » :

$$\boxed{\text{Pour tout } 1 \leq i \leq K, \quad \alpha_i + \sum_{j=1}^{K} \lambda_j p_{j,i} = \lambda_i.}$$

En sommant les équations ci-dessus pour $i \in \{1, \ldots, K\}$, et en utilisant que $\sum_{i=1}^{K} p_{j,i} = 1 - \beta_j$, on obtient

$$\sum_{i=1}^{K} \alpha_i = \sum_{j=1}^{K} \lambda_j \beta_j. \tag{9.4}$$

Lemme 9.4.2. *L'équation de trafic possède une seule solution* $(\lambda_1, \ldots, \lambda_K) \in \mathbb{R}_+^K$.

Démonstration. On introduit une chaîne de Markov intermédiaire qui représente l'état d'un client : il est soit à un guichet $i \in \{1, \ldots, K\}$ soit hors du système dans l'état 0. On note Q la matrice de transition définie de la manière suivante : si $i \neq 0$ et $j \neq 0$, alors $q_{i,j} = p_{i,j}$; si $i \neq 0$, $q_{i,0} = \beta_i$; si $j \neq 0$, $q_{0,j} = \alpha_j / \sum_{l=1}^{K} \alpha_l$; enfin $q_{0,0} = 0$. Cette chaîne de Markov n'est pas la chaîne trace. Les hypothèses 1 et 2 entraînent que la chaîne de Markov de matrice de transition Q est irréductible. La remarque 1.5.7 implique que la chaîne de Markov possède une unique probabilité invariante $\nu = (\nu_i, 0 \leq i \leq K)$ et de plus $\nu_i > 0$ pour tout i. En particulier, on a $\nu Q = \nu$, ce qui donne : pour $1 \leq i \leq K$,

$$\sum_{j=1}^{K} \nu_j p_{j,i} + \nu_0 \frac{\alpha_i}{\sum_{l=1}^{K} \alpha_l} = \nu_i.$$

Une solution de l'équation de trafic est donc $\lambda_i = \left(\sum_{l=1}^{K} \alpha_l\right) \nu_i / \nu_0$, pour $1 \leq i \leq K$. Enfin si $(\lambda'_1, \ldots, \lambda'_K) \in \mathbb{R}_+^K$ est une autre solution de l'équation de trafic, on vérifie à l'aide de l'équation de trafic et de (9.4) que le vecteur $(\nu'_0, \ldots, \nu'_K)$ où $\nu'_0 = (\sum_{l=1}^{K} \alpha_l)/(\sum_{i=1}^{K} (\lambda'_i + \alpha_i))$ et $\nu'_i = \lambda'_i \nu'_0 / \sum_{j=1}^{K} \alpha_j$ est une probabilité invariante de Q. Par unicité de la probabilité invariante, on a $\nu' = \nu$. Cela implique donc que $\lambda_i = \lambda'_i$ pour $i \in \{1, \ldots, K\}$. La solution positive de l'équation de trafic est donc unique. □

On pose $\boxed{\rho_i = \dfrac{\lambda_i}{\mu_i}}$ pour $1 \leq i \leq K$, qui s'interprète comme une densité de trafic.

Théorème 9.4.3. *Si pour tout $1 \leq i \leq K$, on a $\rho_i < 1$, alors l'unique probabilité invariante du processus $X = (X_t, t \geq 0)$, est la probabilité π définie par*

$$\pi_n = \prod_{i=1}^{K} (1-\rho_i)\rho_i^{n_i}, \quad \text{où} \quad n = (n_1, \ldots, n_K).$$

La probabilité invariante est sous forme de produit. En régime stationnaire, la file d'attente au guichet i est, à l'instant t, indépendante de la file d'attente au guichet $j \neq i$. De plus chaque file d'attente i, se comporte comme une file $M/M/1$ de paramètre (λ_i, μ_i), où $(\lambda_1, \ldots, \lambda_n)$ est solution de l'équation de trafic.

Démonstration. Le processus X étant irréductible, il possède au plus une probabilité invariante. Il suffit de vérifier que $\pi A = 0$ pour affirmer que π est la probabilité invariante. Nous allons donc vérifier que pour tout $n \in \mathbb{N}^K$, $\sum_{m \neq n} \pi_m A(m,n) = -\pi_n A(n,n)$, où on rappelle que

$$\begin{aligned} A(n,n) &= -\sum_{m \neq n} A(n,m) \\ &= -\sum_{i=1}^{K} \left[\alpha_i + \beta_i \mu_i \mathbf{1}_{\{n_i > 0\}} + \sum_{j \neq i} p_{i,j} \mu_i \mathbf{1}_{\{n_i > 0\}} \right] \\ &= -\sum_{i=1}^{K} \left[\alpha_i + (1 - p_{i,i}) \mu_i \mathbf{1}_{\{n_i > 0\}} \right]. \end{aligned}$$

On a pour n fixé,

$$\begin{aligned} \sum_{m \neq n} \pi_m A(m,n) = \sum_{i=1}^{K} \Big[& \pi_{n-e_i} A(n-e_i, n) \mathbf{1}_{\{n_i > 0\}} + \pi_{n+e_i} A(n+e_i, n) \\ & + \sum_{j \neq i} \pi_{n+e_j-e_i} A(n+e_j-e_i, n) \mathbf{1}_{\{n_i > 0\}} \Big]. \end{aligned}$$

En utilisant la forme de π, il vient

$$\begin{aligned} \sum_{m \neq n} \pi_m A(m,n) &= \pi_n \sum_{i=1}^{K} \left[\frac{\alpha_i}{\rho_i} \mathbf{1}_{\{n_i > 0\}} + \rho_i \mu_i \beta_i + \sum_{j \neq i} \frac{\rho_j \mu_j p_{j,i}}{\rho_i} \mathbf{1}_{\{n_i > 0\}} \right] \\ &= \pi_n \sum_{i=1}^{K} \left[\lambda_i \beta_i + \frac{1}{\rho_i} \left(\alpha_i + \sum_{j \neq i} \lambda_j p_{j,i} \right) \mathbf{1}_{\{n_i > 0\}} \right]. \end{aligned}$$

Grâce à l'équation de trafic, on a

$$\begin{aligned}\sum_{m\neq n}\pi_m A(m,n) &= \pi_n \sum_{i=1}^{K}\left[\lambda_i\beta_i + \frac{1}{\rho_i}\,\lambda_i\,(1-p_{i,i})\,\mathbf{1}_{\{n_i>0\}}\right]\\ &= \pi_n \sum_{i=1}^{K}\left[\lambda_i\beta_i + \mu_i\,(1-p_{i,i})\,\mathbf{1}_{\{n_i>0\}}\right]\\ &= -\pi_n A(n,n),\end{aligned}$$

où l'on a utilisé (9.4) pour la dernière égalité. Ceci conclut la démonstration. □

9.4.2 Files en tandem, processus de sortie

On désire étudier un système constitué de deux serveurs en tandem. Quand un client arrive, il est d'abord dirigé vers le serveur 1, et dès que sa requête est servie, il est dirigé vers le serveur 2, où il effectue une nouvelle requête. Le système est décrit par le couple $X_t = (X_t^1, X_t^2)$, où X_t^i est la taille du système i : nombre de clients dans la file d'attente du serveur i, y compris le client encore servi par le serveur i. On suppose que le processus des arrivées est un processus de Poisson de paramètre λ, et les temps de service sont des variables indépendantes entre elles et indépendantes du processus des arrivées. On suppose que les temps de service du serveur i suivent des lois exponentielles de paramètre μ_i. Il s'agit d'un cas particulier des réseaux de Jackson. L'équation de trafic se résume à $\lambda = \lambda_1$ pour le serveur 1 et $\lambda_1 = \lambda_2$ pour le serveur 2. En particulier $(X_t, t \geq 0)$ est une chaîne de Markov homogène sur $\mathbb{N}^2$, qui possède une probabilité invariante si $\rho_1 = \lambda/\mu_1 < 1$ et $\rho_2 = \lambda/\mu_2 < 1$. De plus en régime stationnaire X_t^1 et X_t^2 sont, à t fixé, indépendants, et la loi de X_t^i est la loi $\pi(\rho_i)$ donnée par (9.2) avec $\rho = \rho_i$.

On peut utiliser une autre propriété intéressante des files $M/M/1$ pour retrouver ce résultat. On note N_t, le nombre de clients sortis avant l'instant t d'une file d'attente $M/M/1$ de paramètre ρ. Le processus $(N_t, t \geq 0)$ est appelé processus de sortie de la file d'attente. On peut montrer (cf. [1], proposition 4.4 p. 64) que pour $\rho < 1$, en régime stationnaire, le processus de sortie est un processus de Poisson d'intensité λ. De plus le processus de sortie jusqu'à l'instant t, $(N_u, u \in [0,t])$, est en régime stationnaire indépendant de l'évolution future du système.

Si on considère une file d'attente en tandem, le processus des arrivées des clients au premier serveur est un processus de Poisson de paramètre λ. En régime stationnaire, la loi de $X_t^{(1)}$ est la loi $\pi(\rho_1)$. De plus le processus de sortie de la file 1, est un processus de Poisson de paramètre λ. Ce processus correspond au processus des arrivées pour la file 2. En particulier, la loi de $X_t^{(2)}$ est, en régime stationnaire, la loi $\pi(\rho_2)$. Comme en régime stationnaire, $X_t^{(1)}$ est indépendant du processus de sortie de la file 1 jusqu'à l'instant t, on retrouve que $X_t^{(1)}$ et $X_t^{(2)}$ sont indépendants.

9.5 Explosion et récurrence des files $M/GI/1$

On considère une file $M/GI/1$. Les temps de service $(S_n, n \geq 1)$ des différents clients sont des variables aléatoires indépendantes de même loi. Le processus des temps d'arrivée est un processus de Poisson de paramètre $\lambda > 0$. On note X_t le nombre de clients dans le système à l'instant t. On suppose qu'à l'instant 0, le système comporte un seul client dont le service vient juste de débuter : $X_0 = 1$. On note $\tau = \inf\{t > 0, X_t = 0\}$ le temps de retour à un système vide. On a le résultat suivant.

Théorème 9.5.1. *Si $\lambda\mathbb{E}[S_1] \leq 1$, alors p.s. le système retourne à l'état vide : $\tau < \infty$. Si $\lambda\mathbb{E}[S_1] > 1$, alors on a $\mathbb{P}(\tau = \infty) > 0$.*

En particulier, si $\lambda\mathbb{E}[S_1] \leq 1$, alors la file d'attente revient infiniment à l'état vide presque sûrement. Cet état est donc récurrent. En revanche si $\lambda\mathbb{E}[S_1] > 1$, alors le nombre de retours à l'état vide est p.s. fini Et on peut vérifier que $\lim_{t\to\infty} X_t = +\infty$, i.e. la file d'attente est transiente. Ceci permet de compléter l'étude des files d'attente $M/M/1$.

Remarque 9.5.2. Pour les chaînes $M/M/1$, on a $\rho = \lambda\mathbb{E}[S_1]$. La condition de stabilité est satisfaite si $\rho < 1$. Si $\rho = 1$, alors le système retourne p.s. à l'état vide, mais il ne possède pas de probabilité invariante. Donc pour $\rho = 1$, la chaîne de Markov $(X_t, t \geq 0)$ est récurrente nulle. Si $\rho > 1$, alors le système a une probabilité non nulle de ne pas retourner à l'état vide dès qu'un client arrive. La chaîne est transiente. ◊

Le lemme suivant nous servira pour démontrer le théorème 9.5.1.

Lemme 9.5.3. *Soit $N = (N_t, t \geq 0)$ un processus de Poisson de paramètre $\lambda > 0$ et S une variable aléatoire positive indépendante de N. On a $\mathbb{P}(N_S = n) = \mathbb{E}\left[\dfrac{\lambda^n S^n}{n!}\,\mathrm{e}^{-\lambda S}\right]$.*

Démonstration. Pour $n \geq 1$, on a $\{N_S = n\} = \{\sum_{k=1}^{n} T_k \leq S < \sum_{k=1}^{n+1} T_k\}$, où les variables $(T_k, k \geq 1)$ sont indépendantes de loi exponentielle de paramètre λ, et sont indépendantes de S. On déduit de la proposition A.1.21 que $\mathbb{P}(N_S = n) = \mathbb{E}[\psi(S)]$, où $\psi(s) = \mathbb{E}[\mathbf{1}_{\{\sum_{k=1}^{n} T_k \leq s < \sum_{k=1}^{n+1} T_k\}}] = \mathbb{E}[N_s]$ pour $s \geq 0$. D'après la proposition 8.4.2, N_s suit une loi de Poisson de paramètre λs. On en déduit que $\psi(s) = \frac{(\lambda s)^n}{n!}\,\mathrm{e}^{-\lambda s}$, et donc

$$\mathbb{P}(N_S = n) = \mathbb{E}\left[\frac{(\lambda S)^n}{n!}\,\mathrm{e}^{-\lambda S}\right].$$

Pour $n = 0$, en utilisant (8.4), il vient $\mathbb{P}(N_S = 0) = \mathbb{P}(S < T_1) = \mathbb{E}[\mathrm{e}^{-\lambda S}]$. Ceci termine la démonstration du lemme. □

Démonstration du théorème 9.5.1. On définit ν_k égal
- à zéro, si le système s'est vidé avant l'arrivée du k-ième client,
- au nombre de clients arrivés pendant le service du k-ième client sinon.

Ainsi $U = \inf\{n \geq 1 ; \sum_{k=1}^{n} \nu_k = n-1\}$, avec la convention que $\inf \emptyset = +\infty$, désigne le nombre de clients servis avant que le système ne se soit vidé. On a donc $\{\tau < \infty\} = \{U < \infty\}$.

Pour déterminer la loi de ν_1, on remarque que $\{\nu_1 = n\} = \{N_{S_1} = n\}$, où $N = (N_t, t \geq 0)$ est le processus d'arrivée des clients. Comme N est un processus de Poisson de paramètre λ, on déduit du lemme 9.5.3 que pour $n \in \mathbb{N}$, on a $\mathbb{P}(\nu_1 = n) = \mathbb{E}\left[\frac{\lambda^n S_1^n}{n!} \mathrm{e}^{-\lambda S_1}\right]$.

On calcule ensuite la loi jointe de (ν_1, ν_2). On note $(T_n, n \geq 1)$ la suite des temps d'inter-arrivées. Il s'agit d'une suite de variables aléatoires indépendantes de loi exponentielle de paramètre λ. Pour $n \geq 1$, $k \geq 0$, on a

$$\mathbb{P}(\nu_1 = n, \nu_2 \geq k) = \mathbb{P}\Big(\sum_{k=1}^{n} T_k \leq S_1 < \sum_{k=1}^{n+1} T_k, \sum_{k=1}^{n+k} T_k \leq S_1 + S_2\Big).$$

En posant $R = S_1 - \sum_{i=1}^{n} T_i \geq 0$, on obtient

$$\begin{aligned}
\mathbb{P}(\nu_1 = n, \nu_2 \geq k) &= \mathbb{P}\Big(\sum_{k=1}^{n} T_k \leq S_1 < \sum_{k=1}^{n+1} T_k\Big) \\
&\qquad \mathbb{P}\Big(\sum_{k=n+2}^{n+k} T_k + \Big(T_{n+1} - R\Big) \leq S_2 \Big| T_{n+1} > R \geq 0\Big) \\
&= \mathbb{P}\Big(\sum_{k=1}^{n} T_k \leq S_1 < \sum_{k=1}^{n+1} T_k\Big)\mathbb{P}\Big(\sum_{k=n+2}^{n+k} T_k + T_{n+1} \leq S_2\Big) \\
&= \mathbb{P}(\nu_1 = n)\mathbb{P}(\nu_1 \geq k),
\end{aligned}$$

où l'on a utilisé pour la deuxième égalité le caractère sans mémoire des lois exponentielles, i.e. le lemme 8.1.5 avec $V = T_{n+1}$. Soit $(\nu'_k, k \geq 1)$ une suite de variables indépendantes de même loi que ν_1. On a donc montré que $(\nu_k, k \in \{1, \min(2, U)\})$ a même loi que $(\nu'_k, k \in \{1, \min(2, U')\})$, où $U' = \inf\{n \geq 1 ; \sum_{k=1}^{n} \nu'_k = n-1\}$.

En généralisant cette démonstration, on obtient que $(\nu_n, 1 \leq n \leq U)$ a même loi que $(\nu'_n, 1 \leq n \leq U')$. On déduit du lemme 4.4.6 que U a même loi que la population totale d'un processus de Galton-Watson dont la loi de reproduction est la loi de ν_1.

En particulier le système retourne à l'état vide p.s. si la population totale du processus de Galton-Watson est p.s. finie. On déduit de la proposition 4.1.5 que $\tau < \infty$ p.s. si et seulement si $\mathbb{E}[\nu_1] \leq 1$. Comme

$$\mathbb{E}[\nu_1] = \sum_{n=1}^{\infty} n\mathbb{E}\left[\frac{\lambda^n S_1^n}{n!} \mathrm{e}^{-\lambda S_1}\right] = \mathbb{E}\left[\lambda S_1 \sum_{n=0}^{\infty} \frac{\lambda^n S_1^n}{n!} \mathrm{e}^{-\lambda S_1}\right] = \lambda \mathbb{E}[S_1],$$

on en déduit que $\tau < \infty$ p.s. si $\lambda \mathbb{E}[S_1] \leq 1$. Et si $\lambda \mathbb{E}[S_1] > 1$, alors on a $\mathbb{P}(\tau = \infty) > 0$. □

Références

1. S. Asmussen. *Applied probability and queues*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1987.
2. F. Baccelli et P. Brémaud. *Élements of queueing theory*, volume 26 of *Applications of Mathematics*. Springer-Verlag, Berlin, seconde édition, 2003.
3. P. Bougerol. *Processus de saut et files d'attente*. Cours de Maîtrise, Paris VI, http ://www.proba.jussieu.fr/supports.php, 2002.
4. P. Brémaud. *Markov chains. Gibbs fields, Monte Carlo simulation, and queues*. Springer texts in applied mathematics. Springer, 1998.
5. A. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektrotkeknikeren*, 13, 1917.
6. A. Kaufmann et R. Cruon. *Les phénomènes d'attente. Théorie et applications*. Dunod, Paris, 1961.
7. P. Robert. *Réseaux et files d'attente : méthodes probabilistes*, volume 35 de *Mathématiques & Applications*. Springer-Verlag, Berlin, 2000.