MA 411 : Modélisation et analyse des processus stochastiques

Chaînes de Markov à temps continu (CMTC)
Séance de TD du 20 mai 2020

Vous trouverez ci-après l'énoncé et le corrigé de l'exercice 29 de la séance de TD consacrée aux files et réseaux de files d'attente. La correction a été rédigée dans le but de vous aider si vous êtes bloqué ou pour vérifier votre propre travail. Il se peut qu'elle contienne elle-même des erreurs. Si tel est le cas, elles seront corrigées au fur et à mesure qu'elles sont détectées. La version en ligne sur *https://chamilo.grenoble-inp.fr/courses/MA332* sera mise à jour de manière à intégrer ces corrections. Dans de nombreux exercices, il existe plusieurs méthodes pour aboutir au résultat. Si vous avez des doutes sur la méthode que vous avez vous-même employée, n'hésitez pas à m'en faire part (*laurent.lefevre@lcis.grenoble-inp.fr*).

Exercice 29
Dans cet exercice, on s'intéresse en particulier à l'effet du routage dynamique dans des files d'attente aux guichets des administrations, dans les grands magasins, etc. On dispose de deux serveurs (par exemple deux employés) qui traitent des requêtes similaires avec la même efficacité. Le traitement des requêtes (de durée variée, en général) nécessite un temps de service modélisé par une exponentielle de paramètre μ.

1. Dans un premier temps, on cherche à comparer les deux organisations réprésentées à la figure 1. Dans l'organisation *à l'américaine* (à gauche), les clients attendent dans un buffer unique et la discipline de service est *first come, first served*. Dans la file *à la française* (à droite), les clients se répartissent aléatoirement dans deux files, de manière équiprobable. On posera pour simplifier les notations dans les calculs :
$$\rho := \frac{\lambda}{\mu}$$
   (a) donner les conditions de stabilité pour les deux organisations
   (b) calculer pour les deux organisations le nombre moyen de clients présents et le temps moyen de séjour

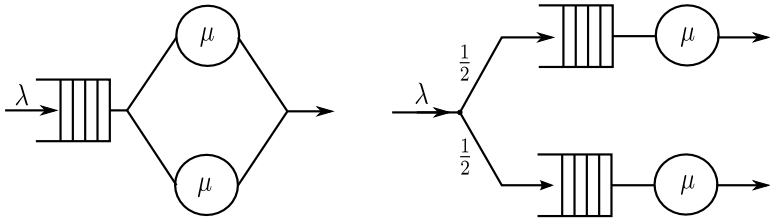

Figure 1: Les deux organisations de files étudiées dans l'exercice 29 (premier point). Elles sont parfois appelées respectivement – improprement – *à l'américaine* (à gauche) et *à la française* (à droite)

(c) comparer les performances de ces deux organisations en terme de nombre de clients et de temps de séjour

(d) calculer et comparer la proportion du temps où au moins un employé est inactif, dans les deux organisations

2. Dans un deuxième temps, on cherche à représenter plus fidèlement l'organisation *à la française*. Dans la réalité, les clients ne se répartissent pas aléatoirement mais choisissent, en arrivant, de se ranger dans la file qui contient le moins de clients[1] . Ce type d'organisation est réprésentée à la figure 2. Le routage devient donc dynamique : la

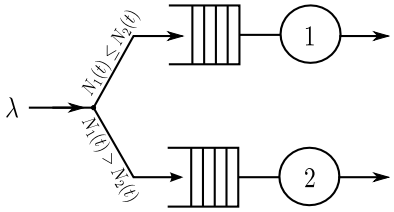

Figure 2: Une organisations de file *à la française* avec un routage dynamique

route choisie par un client qui arrive au temps t dépend de l'état du système $(N_1(t), N_2(t))$ à cet instant.

(a) donner une condition de stabilité pour cette organisation

(b) donner un modèle de ce système sous la forme d'une chaîne de Markov à temps continu

(c) écrire les équations d'équilibre qui permettent de calcule la distribution stationnaire de probabilité (sans les résoudre)

[1] un observateur attentif de l'organisation des files *à la française* pourrait critiquer cette représentation simpliste et affirmer – non sans raison – que dans ce type d'organisation, les clients changent de plus de file lorsqu'ils ont l'impression que l'autre file avance plus vite ...

(d) donner les expressions du nombre moyen de clients présents et du temps moyen de séjour des clients, en fonction des composantes (non calculées) de la distribution stationnaire de probabilité

Correction de l'Exercice 29

1. L'organisation "à l'américaine" correspond à une file $M/M/2$ qui est stable pour $\rho < 2$. Pour ce type de file (voir les résultats du cours sur la file $M/M/C$), on obtient :

$$\begin{aligned} \boldsymbol{\pi}_0 &= \frac{2-\rho}{2+\rho} \\ R_a &= \frac{\rho^2}{\mu(2-\rho)^2}\boldsymbol{\pi}_0 + \frac{1}{\mu} = \frac{1}{\mu}\frac{4}{4-\rho^2} \\ Q_a &= \frac{\rho^3}{(2-\rho)^2}\boldsymbol{\pi}_0 + \rho = \frac{4\rho}{4-\rho^2} \end{aligned}$$

On vérifie bien sûr la loi de Little $Q_a = R_a\lambda$ qui nous permet de calculer (par exemple) Q_a à partir de R_a (ou le contraire). L'organisation "à la française" est celle déjà examinée dans l'exercice précédent (trois imprimantes en parallèle, sans spooler d'impression). En effet, le routage probabiliste entraîne des taux d'arrivées $\lambda \cdot \frac{1}{2} = \frac{\lambda}{2}$ dans chacune des deux files en parallèle. En notant Q_1 et Q_2 les nombres moyens de clients respectivement dans les files 1 et 2, et en notant de même X_1 et X_2 les débits moyens respectivement dans les files 1, et 2, on obtient, par conservation des clients :

$$\begin{aligned} Q_f &= Q_1 + Q_2 \\ X &= X_1 + X_2 \end{aligned}$$

Les deux files simples sont stables si $\lambda/2 < \mu$, c'est-à-dire si $\rho < 2$ (même condition que dans l'organisation "à l'américaine"). Dans ce cas, on obtient pour les paramètres de performance asymptotiques des deux files simples de type $M/M/1$:

$$\begin{aligned} Q_1 = Q_2 &= \frac{\rho}{2-\rho} \\ X_1 = X_2 &= \frac{\lambda}{2} \end{aligned}$$

La relation de Little $Q_f = R_f X$ permet alors d'obtenir

$$R_f = \frac{1}{\mu}\frac{2}{2-\rho}$$

Nous pouvons maintenant répondre aux questions du point 1.

(a) les deux organisations sont stables avec la même condition, à savoir $\lambda < 2\mu$

(b) les nombres moyens de clients ont été calculés ci-dessus. On observe en particulier que

$$Q_a = \frac{4\rho}{4-\rho^2} = Q_f \cdot \frac{2}{2+\rho}$$

et

$$R_a = \frac{1}{\mu}\frac{4}{4-\rho^2} = R_f \cdot \frac{2}{2+\rho}$$

(c) La file *américaine* est donc plus performante, aussi bien en ce qui concerne le nombre de client dans le système, qu'en ce qui concerne le temps moyen de séjour. Le facteur $\alpha := 1 + \rho/2$, d'amélioration de performance entre les deux types d'organisation, croit linéairement depuis $\alpha = 1$ (pour le cas limite $\rho = 0$ où il n'y a pas de client ou que le temps de service est nul) jusqu'à $\alpha = 2$ (à la limite de saturation $\rho \to 2$ du système)

(d) La proportion du temps où un des deux serveur au moins est inoccupé s'écrit :

$$p := \boldsymbol{\pi}_0 + \boldsymbol{\pi}_1$$

Dans le cas de la "file américaine" $(M/M/2)$, on a :

$$\boldsymbol{\pi}_1 = \frac{\rho^1}{1!}\boldsymbol{\pi}_0 = \rho\frac{2-\rho}{2+\rho}$$

et

$$p_a := \frac{(2-\rho)(1+\rho)}{2+\rho}$$

Dans le cas de la "file française", un des serveur est inactif si une des deux files est vide. Par facilité, on calcule la probabilité de l'évènement complémentaire :

$$P\,(\text{auncune file vide}) = P\,(\text{serveur 1 occupé}) \cdot P\,(\text{serveur 2 occupé}) = U^2$$

où U est le taux d'utilisation du serveur dans une fille $M/M/1$ avec un taux d'arrivée $\lambda/2$ et un taux de service μ. On trouve finalement :

$$p_f := 1 - U^2 = 1 - \frac{\rho^2}{4} = \frac{4-\rho^2}{4}$$

On a donc

$$p_f = \beta p_a$$

avec

$$\beta := \frac{(2+\rho)^2}{4(1+\rho)}$$

La proportion du temps où un serveur est inoccupé est plus grande dans l'organisation "à la française" que dans l'organisation "à l'américaine". Le facteur β d'augmentation croit de manière monotone de $\beta = 1$ (cas limite $\rho = 0$, quand il n'y a pas de client ou que le temps de service est nul) à $\beta = 4/3$ (à la limite de saturation du système)

2. La nouvelle configuration envisagée, avec routage dynamique, de la file "à la française" est représentée à la figure 2.

   (a) La capacité maximale de service (débit moyen maximal de sortie) de ce réseau de deux files est 2μ. Il est atteint lorsque les deux serveurs sont occupés. Or, les deux serveurs sont nécessairement occupés dès qu'il y a plus d'un client dans le système. La file est donc instable si λ (taux global d'arrivée) est supérieur ou égal à 2μ et stable pour $\lambda < 2\mu$.

   (b) L'état du système à un instant donné est un couple de valeurs $n(t) = (n_1(t), n_2(t)) \in \mathbb{N}^2$ où $n_1(t)$ et $n_2(t)$ désignent respectivement le nombre de clients dans la file 1 et dans la file 2, à l'instant $t > 0$. En effet, les probabilité de transition dépendent – à cause du routage dynamique – non seulement du nombre total de clients dans le réseau, mais également de la répartition de ces clients entre les deux files. Le processus stochastique reste néanmoins une chaîne à temps continu puisque le nouvel espace d'état $E := \mathbb{N}^2$ est lui aussi dénombrable. Par ailleurs, les probabilités de transition depuis un état $n = (n_1, n_2)$ vers un autre état ne dépendent que de cet état et de l'état de destination, pas du temps t, ni du chemin parcouru pour arriver à l'état $n = (n_1, n_2)$. Le processus considéré dans son ensemble (i.e. le réseau de deux files simples avec un routage dynamique) reste donc bien une chaîne de Markov homogène à temps continu. Nous pouvons dès lors spécifier le modèle en donnant tous les taux de transition. Une représentation commode du modèle, c'est-à-dire du graphe associé à la CMTC, est celle donnée à la figure 3.

   (c) La CMTC de la figure 3 est irréductible. Sous l'hypothèse de stabilité $\lambda < 2\mu$, la distribution stationnaire existe, est unique, non nulle et égale à la distribution limite de probabilité. On peut donc définir les probabilités stationnaires de la manière suivante :

$$\boldsymbol{\pi}_{m,n} := \lim_{T \to \infty} P\left(N_1(t) = m, N_2(t) = n\right), \ \forall m, n \in \mathbb{N}$$

   Les équations d'équilibre en chaque noeud sont des récurrence qui lient, dans le cas général, la probabilité stationnaire d'un état avec les probabilités d'état de ses quatre voisins. Ces équations sont

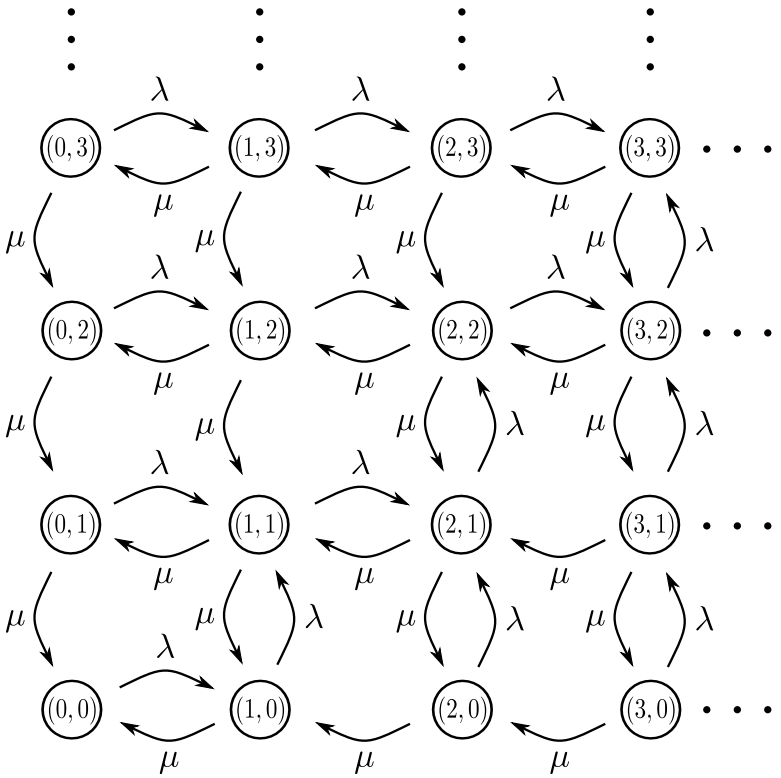

Figure 3: Le graphe de la CMTC associée au réseau de deux files avec routage dynamique

difficiles à écrire, et bien sûr bien davantage encore à résoudre. On pourra toutefois se limiter en première approximation à la résolution d'un problème approché obtenu en limitant l'état à $\bar{E} := \{1, \ldots, N\} \times \{1, \ldots, N\}$ car par hypothèse de stabilité, les probabilités d'état sont décroissantes avec N. Les équations d'équilibre – quelles soient obtenues en exprimant l'équilibre en chaque noeud ou par la méthodes des coupes – devront être complétées par la condition de normalisation :

$$\sum_{m,n \geq 0} \boldsymbol{\pi}_{m,n} = 1$$

(d) Le nombre moyen de client dans le système s'écrit

$$\begin{aligned} Q &:= \sum_{m,n\geq 0} (m+n)\boldsymbol{\pi}_{m,n} \\ &= \sum_{m,n\geq 0} m\boldsymbol{\pi}_{m,n} + \sum_{m,n\geq 0} n\boldsymbol{\pi}_{m,n} \\ &= \sum_{m\geq 0} m \sum_{n\geq 0} \boldsymbol{\pi}_{m,n} + \sum_{n\geq 0} n \sum_{m\geq 0} \boldsymbol{\pi}_{m,n} \end{aligned}$$

La CMTC de la figure 3 n'est pas symétrique – à cause du routage dynamique[2] – et en général $\boldsymbol{\pi}_{m,n} \neq \boldsymbol{\pi}_{n,m}$. Une fois calculé Q, le temps moyen de séjour des clients dans le système s'obtient par la formule de Little :

$$R = \frac{Q}{X} = \frac{Q}{\lambda}$$

[2]pour rétablir la symétrie, il faudrait que lorsque les deux files sont de même taille, le client entrant choisisse de manière équiprobable l'une ou l'autre file