TD 8: Chaînes de Markov en temps continu

lionel.rieg@ens-lyon.fr

Exercice 1 (La chaîne M/M/1)

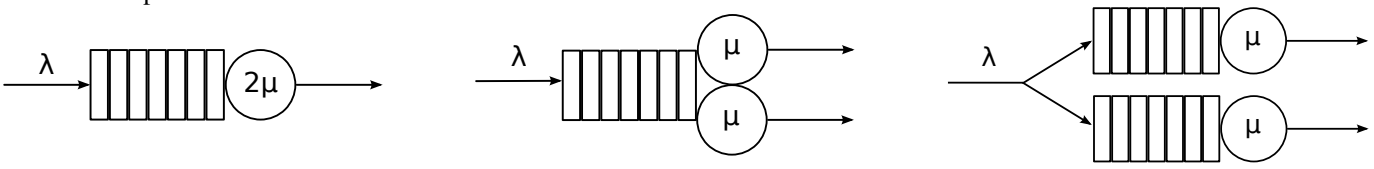
On étudie une file d'attente avec un serveur et une capacité infinie. Les temps d'inter-arrivées et de service sont choisis exponentiels de paramètres respectifs λ et μ. On pose $\rho = \frac{\lambda}{\mu}$ le taux d'utilisation du serveur. Le nombre de clients dans la file, noté $N(t)$, est une chaîne de Markov en temps continu à valeurs dans $\mathbb{N}$.

1. Décrire le schéma de Matthes de cette file.
2. Donner la condition de stabilité de la file et la démontrer à l'aide des théorèmes de Foster.
3. Déterminer le générateur infinitésimal de la chaîne.
4. Calculer l'unique distribution stationnaire de cette chaîne.
5. Donner le nombre moyen de clients dans la file en régime stationnaire.

On peut également calculer le *temps moyen de service* (temps qui sépare une arrivée dans la file et sa sortie) : il vaut $\overline{W} = \frac{1}{\mu - \lambda}$.

Exercice 2 (Comparaison de trois modèles de serveur)

On souhaite comparer les trois architectures suivantes de file d'attente :



1. Intuitivement, quelle configuration semble la meilleure ?
2. Montrez que le processus d'arrivée dans chaque file de la dernière configuration suit une loi de Poisson de paramètre $\frac{\lambda}{2}$.

Pour chacune des trois configurations :

3. Donnez sa condition de stabilité.
4. Écrire son générateur infinitésimal (en précisant les cas limites).
5. Calculer le nombre moyen de clients dans le système.
6. En déduire le temps moyen de réponse.

Où se trouve la différence entre les différentes configurations ? Laquelle préférer ?

Exercice 3 (Serveur et miroir)

On dispose de deux serveurs Web : un serveur principal et un serveur miroir. Les requêtes arrivent à un routeur qui doit essayer de répartir le travail entre le serveur principal ou son miroir, qui est supposé moins performant. Le routeur ne connaissant pas la charge respectives des deux serveurs, il décide d'envoyer chaque requête au serveur principal avec probabilité p ou au serveur miroir avec probabilité $1 - p$. On cherche à optimiser le choix de p.

On suppose que les arrivées suivent une loi de Poisson de paramètre λ. Les temps de service des serveurs sont exponentiels de paramètres respectifs μ pour le serveur principal et ν pour le miroir.

1. Quelles sont les conditions de stabilité du système ?
2. Calculer le nombre de requêtes en attente dans le serveur principal et dans le miroir à l'état stationnaire.
3. Donnez le temps de réponse moyen dans le système global.
4. Déterminez la valeur optimale de p.

Solutions

▸ Exercice 1

1. On a :
   - $\mathcal{E} = \mathbb{N}$ (le nombre de clients en attente dans la file),
   - $\mathcal{S} = \{in, out\}$,
   - $c(\alpha, \iota) \equiv 1$,
   - $\mathcal{A}(0) = \{in\}$ et $\mathcal{A}(n+1) = \{in, out\}$,
   - $p(in, n, n+1) = 1$ et $p(out, n+1, n) = 1$
   - $F_{in}(t) = 1 - \mathrm{e}^{-\lambda t}$ et $F_{out}(t) = 1 - \mathrm{e}^{-\mu t}$
2. La condition est $\rho < 1$. Pour appliquer les théorèmes de Foster, on prend $h = Id$, $F = \{0\}$ ou $F = \varnothing$, $\varepsilon = -\mu(\rho - 1)$ et $M = 1$.
3. le générateur infinitésimal Q est donné par les formules $q_{ii} = -(\lambda + \mu)$, $q_{i+1,i} = \lambda$ et $q_{i-1,i} = \mu$. Il vérifie $1 \cdot Q = 0$.
4. On utilise la méthode de la coupure : à l'équilibre, la probabilité d'être dans les n premiers états est invariante donc ce qui en sort égale ce qui en entre. On en déduit : $\mu\pi_n = \lambda\pi_{n+1}$, *i.e.* $\pi_{n+1} = \rho\pi_n$. C'est valable pour tout n d'où $\pi_n = \rho^n\pi_0$. La condition de normalisation $\sum_{n\in\mathbb{N}} \pi_n = \pi_0 \frac{1}{1-\rho} = 1$ permet d'obtenir $\pi_0 = 1 - \rho$. On vérifie qu'elle est bien stationnaire avec la caractérisation $Q\pi = 0$:

   $$(Q\pi)_i = \sum_{j\in\mathbb{N}} q_{ij}\pi_i = \lambda(1-\rho)\rho^{i-1} - (\lambda+\mu)(1-\rho)\rho^i + \mu(1-\rho)\rho^{i+1} = (1-\rho)\rho^i(\lambda\rho^{-1} - (\lambda+\mu) + \mu\rho) = (1-\rho)\rho^i(\mu - (\lambda+\mu) + \lambda) = 0$$

   Enfin, l'unicité est assurée par l'irréductibilité de la chaîne.
5. Le nombre moyen de clients dans la file est son espérance selon la distribution stationnaire. Pour le calculer, on peut le faire directement en utilisant la fonction génératrice.

   $$G_X(z) = \sum_{n\in\mathbb{N}} \mathbb{P}(X = n)\, z^n = \sum_{n\in\mathbb{N}} (1-\rho)\rho^n z^n = (1-\rho)\sum_{n\in\mathbb{N}} (\rho z)^n = \frac{1-\rho}{1-\rho z}$$

   En effet, l'espérance de X est alors $G'_X(1)$:

   $$G'_N(z) = (1-\rho)\frac{-1}{(1-\rho z)^2}(-\rho) = \frac{\rho(1-\rho)}{(1-\rho z)^2}$$

   d'où $\mathbb{E}(N) = \frac{\rho}{(1-\rho)}$.
   On peut également remarquer qu'il s'agit à un facteur ρ près d'une distribution géométrique dont le terme général est $p(1-p)^{n-1}$ et dont l'espérance vaut $\frac{1}{p}$. Le paramètre de cette loi est $p = 1 - \rho$ et on en déduit le même résultat.

▸ Exercice 2

1. C'est subjectif mais … la deuxième ?
2. On montre tout d'abord que c'est un processus de Poisson, *i.e.* un processus sans mémoire. Pour cela, on note T_n l'arrivée du n^{e} paquet dans le système et $T'_n = T_{2n}$ l'arrivée du n^{e} paquet dans l'une des deux files. On a alors :

   $$\mathbb{P}\Big(T'_{n+1} = a \,\Big|\, T'_n = b_n, \ldots, T'_1 = b_1\Big) = \mathbb{P}\Big(T'_{2n+2} = a \,\Big|\, T'_{2n} = b_n, \ldots, T'_2 = b_1\Big) = \mathbb{P}\Big(T'_{2n+2} = a \,\Big|\, T'_{2n} = b_n\Big) = \mathbb{P}\Big(T'_{n1} = a \,\Big|\, T'_n = b_n\Big)$$

   en utilisant la propriété de Markov forte sur (T_n). il reste à vérifier que son paramètre est $\frac{\lambda}{2}$:

   $$\mathbb{E}\Big(T'_{n+1} - T'_n\Big) = \mathbb{E}(T_{2n+2} - T_{2n}) = \mathbb{E}(T_{2n+2} - T_{2n+1} + T_{2n+1} - T_{2n}) = \mathbb{E}(T_{2n+2} - T_{2n+1}) + \mathbb{E}(T_{2n+1} - T_{2n}) = \frac{2}{\lambda}$$

3. La condition de stabilité est toujours $\lambda < 2\mu$.
4. – selon le premier exercice : $q_{i+1,i} = \lambda$, $q_{i-1,i} = 2\mu$, $q_{ii} = -(\lambda + 2\mu)$
   – l'analyse donne $q_{i+1,i} = \lambda$, $q_{i-1,i} = 2\mu$, $q_{ii} = -(\lambda + 2\mu)$
   – chaque file a un générateur donné par $q_{i+1,i} = \frac{\lambda}{2}$, $q_{i-1,i} = \mu$, $q_{ii} = -(\frac{\lambda}{2} + \mu)$ donc par linéarité de la dérivation, le générateur final est $q_{i+1,i} = \lambda$, $q_{i-1,i} = 2\mu$, $q_{ii} = -(\lambda + 2\mu)$
5. – $\frac{\lambda}{2\mu-\lambda}$

- $\frac{\lambda}{2\mu-\lambda}$
- $2\frac{\frac{\lambda}{2}}{\mu-\frac{\lambda}{2}} = 2\frac{\lambda}{2\mu-\lambda}$

6. – d'après l'exercice 1 : $\frac{1}{2\mu-\lambda}$
   – presque pareil (même temps d'attente mais la loi de traitement par les serveurs est plus faible)
   – le temps moyen est celui de chaque file : $\frac{1}{\mu-\frac{\lambda}{2}} = \frac{2}{2\mu-\lambda}$

La différence entre les deux premières configurations se fait lorsqu'il n'y a qu'un client dans le système : il est servi à vitesse double dans la première. La meilleure configuration semble être la première mais il y a d'autres facteurs à prendre en compte dans la vie réelle : le coût, la résistance aux pannes, …

► **Exercice 3**

1. Les arrivées vers les deux serveurs sont des processus de Poisson de paramètres $p\lambda$ et $(1-p)\lambda$. En effet, il s'agit de la composition de deux processus sans mémoire : un processus de Poisson et une variable de Bernoulli. Les composées sont donc des processus de Poisson. Leur espérance sur un intervalle $[0, t]$ est intuitivement $p \times \mathbb{E}(A(t))$ où $A(t)$ est le nombre de requêtes déjà arrivées. Plus formellement, si $A^*(t)$ dénote la variable aléatoire donnant le nombre d'arrivée dans la file du serveur principal dans l'intervalle $[0, t]$, en utilisant la formule des probabilités totales on obtient :

$$\mathbb{E}(A^*(t)) = \sum_{k\in\mathbb{N}} \mathbb{E}(A^*(t) \,|\, A(t) = k) \cdot \mathbb{P}(A(t) = k) = \sum_{k\in\mathbb{N}} kp \cdot \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!} = p\sum_{k\in\mathbb{N}} k\mathrm{e}^{-\lambda}\frac{\lambda^k}{k!} = p\mathbb{E}(A(t)) = p\lambda t$$

   L'égalité $\mathbb{E}(A^*(t) \,|\, A(t) = k) = kp$ vient du fait que les k arrivées ont indépendamment tiré une variable de Bernoulli de paramètre p, $\mathbb{E}(A^*(t) \,|\, A(t))$ suit donc une loi binomiale de paramètres k et p dont l'espérance est kp. Ainsi, le paramètre du processus résultant est $p\lambda$.

   Les deux conditions sont alors $p\lambda < \mu$ et $(1-p)\lambda < \nu$.

2. D'après l'étude du premier exercice, le nombre moyen de requêtes dans le serveur principal est $\frac{p\lambda}{\mu-p\lambda}$ et $\frac{(1-p)\lambda}{\nu-(1-p)\lambda}$ dans le miroir.

3. Le temps de réponse moyen dans le serveur principal est $\frac{1}{\mu-p\lambda}$ et dans le serveur secondaire $\frac{1}{\nu-(1-p)\lambda}$. Au total, on décompose par la formule des probabilités totales, d'où $\overline{W} = p\frac{1}{\mu-p\lambda} + (1-p)\frac{1}{\nu-(1-p)\lambda}$.

4. Elle vaut :

$$p_{opt} = \frac{\mu\lambda - 2\mu\nu + \sqrt{-2\mu^2\lambda\nu + 2\mu^2\nu^2 - 2\mu\nu^2\lambda + \mu\nu^3 + \nu\mu\lambda^2 + \mu^3\nu}}{(\mu-\nu)\lambda}$$