Recherches Opérationnelles

Processus Stochastiques et Files d'Attente

Paul MINCHELLA

Octobre 2025

Table des matières

Introduction générale : pourquoi modéliser les files d'attente ?

Motivation : pourquoi étudier les files d'attente ?

Les files d'attente sont omniprésentes dans les systèmes réels :

- dans un **commerce**, les clients attendent à une caisse ou à un guichet ;
- dans une **clinique**, les patients attendent un médecin ou un résultat d'examen ;
- dans un **centre d'appels**, les requêtes attendent un opérateur disponible ;
- dans un **système informatique**, les tâches attendent un processeur ou une ressource mémoire.

Dans tous ces contextes, l'ingénieur cherche à répondre à des questions concrètes :

- Combien de serveurs faut-il pour que le délai moyen reste acceptable ?
- Quel est le risque que le système soit saturé ?
- Comment concilier *coût* et *qualité de service (SLA)* ?

L'objectif du cours est donc de donner un cadre mathématique à ces problèmes. Une fois le modèle posé, on peut calculer les quantités clefs : *temps d'attente moyen*, *taux d'occupation*, *probabilité de refus*, etc. et ainsi optimiser les dispositions.

Pourquoi la loi exponentielle est-elle si intuitive ?

La loi exponentielle, caractérisée par son paramètre λ, possède une propriété exceptionnelle :

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t),$$

c'est-à-dire **l'absence de mémoire**. En d'autres termes, le système "oublie" combien de temps il a déjà attendu : si un client n'est pas encore arrivé, sa probabilité d'arriver dans la prochaine minute ne dépend pas du temps déjà écoulé.

- Cette propriété rend le processus de Poisson **naturelement markovien**.
- Elle permet une modélisation **sans historique** : tout l'avenir dépend seulement de l'état courant, pas du passé.
- C'est cette simplicité qui rend les modèles de files M/M/1, M/M/k, etc., analytiquement solvables.

Ainsi, le couple *processus de Poisson + loi exponentielle* fournit un cadre minimaliste mais réaliste pour décrire des arrivées et services aléatoires.

Pourquoi la loi de Poisson ?

La *loi de Poisson* occupe une place centrale dans la modélisation des phénomènes aléatoires discrets. Elle décrit le nombre d'événements se produisant dans un intervalle de temps (ou d'espace) lorsque ces événements :

— surviennent indépendamment les uns des autres,

— se produisent à un *taux moyen constant* λ,

— et sont suffisamment rares pour que deux événements simultanés soient négligeables.

Intérêt pratique. Cette loi intervient naturellement dans les systèmes où les arrivées sont indépendantes : appels téléphoniques, arrivées de clients, défaillances de machines, ou paquets dans un réseau. Elle sert de base au *processus de Poisson homogène*, qui modélise le flux temporel d'événements dans les files d'attente.

En résumé : la loi de Poisson est la passerelle entre la *comptabilisation d'événements discrets* et la *modélisation dynamique des arrivées* dans le temps.

Pourquoi modéliser les arrivées par un processus de Poisson ?

Le **processus de Poisson homogène** est la pierre angulaire de la modélisation stochastique des arrivées aléatoires. Il repose sur trois hypothèses simples, mais extraordinairement puissantes :

1. Les événements (arrivées) sont **rares** et indépendants à petite échelle ;
2. Les incréments sont **stationnaires** : le comportement statistique ne dépend pas du moment de la journée ;
3. Les incréments sont **indépendants** : ce qu'il se passe sur une période n'influence pas la suivante.

Ces hypothèses mènent naturellement à la loi de Poisson pour le nombre d'événements dans un intervalle, et à la loi **exponentielle** pour les temps inter-arrivées. Ce modèle est à la fois *mathématiquement simple* et *empiriquement très fidèle* à de nombreux systèmes réels (trafic réseau, appels téléphoniques, clients d'un magasin, etc.).

Conclusion : de l'aléatoire au dimensionnement optimal

Étudier les files d'attente, c'est comprendre comment :

— les phénomènes aléatoires influencent les performances d'un système,

— on peut passer de l'observation (taux d'arrivée, taux de service) à une **politique de gestion optimale**,

— et comment la **modélisation probabiliste** éclaire les décisions concrètes (embauche, allocation de serveurs, investissements, etc.).

Le fil conducteur reste le même : partir du hasard, modéliser rigoureusement, et aboutir à des **résultats quantitatifs utiles** à l'ingénieur.

Chapitre 1

Rappels sur les probabilités et les fonctions de répartition

1.1 Espace de probabilité, variables aléatoires et lois usuelles

Définition 1.1.1 (Espace de probabilité)**.** *Un* espace de probabilité *est un triplet* $(\Omega, \mathcal{F}, \mathbb{P})$ *où :*

- Ω *est l'****ensemble des issues possibles*** *(ou espace des mondes possibles) ;*
- $\mathcal{F}$ *est une* σ***-algèbre*** *de sous-ensembles de* Ω*, représentant les événements observables, c'est-à-dire ceux auxquels on peut attribuer une probabilité*
  - *(i)* $\varnothing \in \mathcal{F}$ *et* $\Omega \in \mathcal{F}$ *;*
  - *(ii) si* $A \in \mathcal{F}$*, alors* $A^c \in \mathcal{F}$ (stabilité par complémentaire) *;*
  - *(iii) si* $(A_n)_{n\geq 1} \subset \mathcal{F}$*, alors* $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ (stabilité par union dénombrable) *;*
- $\mathbb{P}$ *est une* ***mesure de probabilité*** *sur* $(\Omega, \mathcal{F})$*, c'est-à-dire une application*

$$\mathbb{P} : \mathcal{F} \longrightarrow [0,1]$$

satisfaisant les trois axiomes fondamentaux :

- *(i)* ***Positivité :*** $\mathbb{P}(A) \geq 0$ *pour tout événement* $A \in \mathcal{F}$ *;*
- *(ii)* ***Normalisation :*** $\mathbb{P}(\Omega) = 1$ *;*
- *(iii)* σ***-additivité :*** *pour toute suite* $(A_i)_{i\geq 1}$ *d'événements deux à deux disjoints,*

$$\mathbb{P}\left(\bigcup_{i\geq 1} A_i\right) = \sum_{i\geq 1} \mathbb{P}(A_i).$$

Les trois propriétés font de $\mathbb{P}$ une *mesure de probabilité*. Elles garantissent que la probabilité est cohérente avec l'intuition d'une mesure de "taille" des événements possibles.

Exemple 1.1.1 (Lancer d'un dé)**.** *On peut modéliser un dé à six faces par :* $\Omega = \{1,2,3,4,5,6\}$, $\mathcal{F} = 2^{\Omega}$, $\mathbb{P}(\{i\}) = \frac{1}{6}$, $i = 1, \ldots, 6$. *Ainsi,* Ω *décrit les issues possibles,* $\mathcal{F}$ *les événements mesurables (par exemple "nombre pair"), et* $\mathbb{P}$ *la probabilité qui leur est associée.*

Définition 1.1.2 (Variable aléatoire)**.** *Une* variable aléatoire réelle *est une application mesurable*

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})),$$

qui associe à chaque issue $\omega \in \Omega$ une valeur réelle $X(\omega)$. Sa loi *est la mesure image $\mathbb{P}_X$ définie par :*

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A), \qquad A \in \mathcal{B}(\mathbb{R}).$$

Remarque 1 (Ce qui nous intéresse vraiment). Dans la pratique, $(\Omega, \mathcal{F})$ sert de cadre théorique, mais on ne le manipule presque jamais directement. L'essentiel est la **loi de** X, c'est-à-dire la distribution des valeurs réelles prises par X. Autrement dit, on s'intéresse au comportement de $X(\omega)$ plutôt qu'aux issues ω elles-mêmes.

C'est cette loi $\mathbb{P}_X$ que l'on calcule, simule ou modélise : elle contient toute l'information probabiliste utile. Ainsi, dans la suite du cours, nous raisonnerons presque toujours sur la distribution de X plutôt que sur l'espace abstrait $(\Omega, \mathcal{F}, \mathbb{P})$. On doit retenir que :

(i) la notion de mesurabilité garantit que les événements du type $\{X \in A\}$ sont bien définis ;

(ii) la loi d'une variable aléatoire est une probabilité sur $\mathbb{R}$.

En pratique, on manipule rarement directement $(\Omega, \mathcal{F})$, mais on raisonne toujours sur la loi de X.

Définition 1.1.3 (Densité d'une variable aléatoire réelle)**.** *Soit X une variable aléatoire réelle définie sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. On dit que X admet une* densité f_X *si, pour tout $A \subset \mathbb{R}$ borélien,*

$$\mathbb{P}(X \in A) = \int_A f_X(x)\,\mathrm{d}x.$$

Dans ce cas, la fonction f_X est ***positive*** *et vérifie :*

$$\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1.$$

Intuition.

— La densité f_X décrit la "concentration de probabilité" sur la droite réelle : elle ne donne pas directement une probabilité, mais une densité locale.

— Pour obtenir une probabilité sur un intervalle, on intègre la densité sur cet intervalle.

— Toutes les lois continues usuelles (exponentielle, gaussienne, uniforme, etc.) possèdent une densité.

Définition 1.1.4 (Fonction de répartition)**.** *La* fonction de répartition *d'une variable aléatoire réelle X est la fonction $F_X : \mathbb{R} \to [0,1]$ définie par :*

$$F_X(x) = \mathbb{P}(X \leq x).$$

Si X admet une densité f_X, alors :

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t, \quad \textit{et} \quad f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) \text{ p.s.}$$

Propriétés.

— C'est une fonction **croissante**, **continue à droite**, et telle que :

$$\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to +\infty} F_X(x) = 1.$$

Définition 1.1.5 (Espérance)**.** *Soit X une variable aléatoire réelle. L'*espérance *de X, si elle existe, est*

$$\mathbb{E}[X] \;=\; \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega).$$

Dans le cas où X admet une densité f,

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f(x)\,\mathrm{d}x.$$

Théorème 1.1.1 (Théorème du transfert)**.** *Soit X une variable aléatoire réelle de loi μ_X et soit $g : \mathbb{R} \to \mathbb{R}$ une fonction mesurable telle que $g(X)$ soit intégrable. Alors*

$$\mathbb{E}[g(X)] = \int_\Omega g(X(\omega))\,\mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}} g(x)\,\mu_X(\mathrm{d}x).$$

Remarque. Ce théorème signifie que pour calculer l'espérance d'une fonction $g(X)$, on peut se contenter de travailler directement avec la loi de X :

— si X est discrète, alors $\mathbb{E}[g(X)] = \sum_x g(x)\,\mathbb{P}(X = x)$;

— si X est continue de densité f, alors $\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f(x)\,\mathrm{d}x$.

C'est l'outil fondamental pour manipuler les espérances : on applique g sur la variable aléatoire et on intègre par rapport à sa loi. C'est un indispensable pour calculer moments, transformations linéaires et lois de variables composées.

Définition 1.1.6 (Moment d'ordre k)**.** *Le* moment d'ordre k *est défini par $\mathbb{E}[X^k]$, lorsque l'intégrale est finie.*

Définition 1.1.7 (Variance)**.** *La* variance *de X, variable aléatoire réelle, est notée* $\mathrm{Var}(X)$ *et s'exprime par :*

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Complément. L'espérance représente la valeur moyenne de X, la "tendance centrale" tandis que sa variance mesure la dispersion des valeurs autour de cette moyenne. Il faut savoir calculer $\mathbb{E}[X]$ et $\mathrm{Var}(X)$ à partir d'une densité ou d'une loi de probabilité. Retenir que $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Définition 1.1.8 (Fonctions génératrices et caractéristiques)**.** *Soit X une variable aléatoire réelle.*

— *La* ***fonction génératrice des moments*** *(si elle existe) est :*

$$M_X(t) = \mathbb{E}[e^{tX}], \qquad t \in \mathbb{R}.$$

Elle est utile pour calculer les moments par dérivation : $\mathbb{E}[X^n] = M_X^{(n)}(0)$.

— *La* ***fonction caractéristique*** *est définie pour tout $t \in \mathbb{R}$ par :*

$$\varphi_X(t) = \mathbb{E}[e^{itX}],$$

où $i^2 = -1$. Elle existe toujours (car $|e^{itX}| = 1$) et détermine complètement la loi de X.

Intuition.

- $M_X(t)$ et $\varphi_X(t)$ traduisent la loi de X dans le "domaine des transformées".
- $M_X(t)$ met en évidence les moments (espérance, variance, etc.), tandis que $\varphi_X(t)$ est utilisée pour prouver des convergences (théorème central limite, etc.).
- Ces outils sont les analogues continus des transformées de Laplace ou de Fourier.

Théorème 1.1.2 (Loi faible des grands nombres)**.** *Soient $(X_n)_{n\geq 1}$ des variables aléatoires indépendantes et identiquement distribuées, d'espérance finie $\mathbb{E}[X_1] = m$. Alors, pour tout $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - m\right| > \varepsilon\right) = 0.$$

Autrement dit, la moyenne empirique converge en probabilité vers la moyenne théorique :

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{\mathbb{P}} m.$$

Théorème 1.1.3 (Loi forte des grands nombres)**.** *Sous les mêmes hypothèses que ci-dessus, on a la convergence presque sûre :*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{p.s.} m.$$

Autrement dit, avec probabilité 1*, la moyenne observée sur un grand nombre d'essais tend vers l'espérance. C'est la formalisation mathématique de la* stabilisation des fréquences *observée expérimentalement.*

Théorème 1.1.4 (Théorème central limite (TCL))**.** *Soient $(X_n)_{n\geq 1}$ des variables i.i.d. d'espérance $\mathbb{E}[X_1] = m$ et de variance $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Alors,*

$$\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{k=1}^{n} X_k - m\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

où $\xrightarrow{\mathcal{L}}$ désigne la convergence en loi. Autrement dit, la moyenne empirique est asymptotiquement normale :

$$\frac{1}{n}\sum_{k=1}^{n} X_k \approx \mathcal{N}\left(m, \frac{\sigma^2}{n}\right) \quad \textit{pour } n \textit{ grand.}$$

Commentaires.

- La **loi faible** exprime la convergence en probabilité - utile pour les grandes masses de données ou les simulations. La **loi forte** est quant à elle plus profonde : elle garantit une stabilisation presque sûre, indépendante du hasard résiduel.
- Le **théorème central limite** décrit la fluctuation aléatoire autour de la moyenne : quelle que soit la loi d'origine, la moyenne centrée et normalisée devient gaussienne.

1.2 Notion de probabilité et variable aléatoire

Nous l'avons vu, une **probabilité** $\mathbb{P}$ est bien une application qui, à tout événement aléatoire, associe un nombre réel entre 0 et 1, représentant la *chance de réalisation* de cet événement. Une **variable aléatoire** X est une grandeur réelle (ou vectorielle) dont la valeur dépend du hasard. On note souvent :

$$\begin{aligned} X : \Omega &\longrightarrow \mathbb{R}, \\ \omega &\mapsto X(\omega) \end{aligned}$$

où Ω désigne l'ensemble des issues possibles de l'expérience aléatoire.

La loi de X décrit la manière dont ses valeurs se répartissent dans $\mathbb{R}$. Pour la décrire, on introduit la notion de fonction de répartition.

1.3 Fonction de répartition

Définition 1.3.1. *La **fonction de répartition** d'une variable aléatoire réelle X est définie par :*

$$F_X(t) = \mathbb{P}(X \leq t), \qquad t \in \mathbb{R}.$$

Elle possède les propriétés fondamentales suivantes :

- F_X est **croissante** : si $t_1 < t_2$, alors $F_X(t_1) \leq F_X(t_2)$;
- $\lim\limits_{t \to -\infty} F_X(t) = 0$ et $\lim\limits_{t \to +\infty} F_X(t) = 1$;
- F_X est **continue à droite**.

Propriété 1.3.1. *Deux variables aléatoires X et Y ont même loi si et seulement si elles ont la même fonction de répartition, c'est-à-dire que pour tout $t \in \mathbb{R}$:*

$$X \overset{\mathcal{L}}{=} Y \quad \Longleftrightarrow \quad F_X(t) = F_Y(t).$$

*Ainsi, la fonction de répartition **caractérise complètement la loi** d'une variable aléatoire réelle.*

1.4 Fonction de survie

Définition 1.4.1. *La **fonction de survie** associée à X est définie par :*

$$G_X(t) = \mathbb{P}(X > t) = 1 - F_X(t).$$

Elle vérifie les propriétés suivantes :

- G_X est **décroissante** sur $\mathbb{R}$;
- $G_X(t) \in [0, 1]$ pour tout t ;
- $\lim\limits_{t \to -\infty} G_X(t) = 1$ et $\lim\limits_{t \to +\infty} G_X(t) = 0$.

Dans le cas d'une variable aléatoire à densité, la fonction de survie s'exprime comme l'intégrale de cette dernière :

$$G_X(t) = \int_t^{+\infty} f_X(u)\,\mathrm{d}u.$$

Elle représente la probabilité que l'événement étudié *n'ait pas encore eu lieu* à la date t.

1.5 Quelques lois usuelles importantes

Définition 1.5.1 (Lois usuelles)**.** *On utilisera fréquemment :*

— ***Exponentielle*** $X \sim \mathrm{Exp}(\lambda)$ *:*

$$f(x) \;=\; \lambda e^{-\lambda x}\, \mathbb{1}_{\{x\geq 0\}}.$$

Sa fonction de répartition est :

$$F_X(t) = \mathbb{P}(X \leq t) = \begin{cases} 0, & t < 0, \\ 1 - e^{-\lambda t}, & t \geq 0, \end{cases}$$

et sa fonction de survie :

$$G_X(t) = \mathbb{P}(X > t) = e^{-\lambda t}, \quad t \geq 0.$$

Remarque : $G_X(t)$ traduit la ***propriété de mémoire nulle*** *:*

$$\mathbb{P}(X > s + t \mid X > s) = \frac{G_X(s+t)}{G_X(s)} = e^{-\lambda t}.$$

— ***Gamma/Erlang*** $X \sim \mathrm{Gamma}(k, \lambda)$ *:*

$$f(x) \;=\; \frac{\lambda^k}{\Gamma(k)}\, x^{k-1} e^{-\lambda x}\, \mathbb{1}_{\{x\geq 0\}}.$$

Remarque : si $k \in \mathbb{N}$ (cas Erlang), alors $\Gamma(k) = (k-1)!$.

— ***Poisson*** $N \sim \mathrm{Poisson}(\theta)$*, $\theta \geq 0$:*

$$\mathbb{P}(N = n) \;=\; e^{-\theta}\, \frac{\theta^n}{n!}\, \mathbb{1}_{\{n\in\mathbb{N}\}}.$$

— ***Géométrique*** $G \sim \mathrm{Geom}(p)$*, $p \in (0,1)$ (sur $\mathbb{N} = \{0, 1, 2, \dots\}$) :*

$$\mathbb{P}(G = n) \;=\; (1-p)^n p\, \mathbb{1}_{\{n\geq 0\}}.$$

Remarque 2. Pour rappel, la fonction Γ généralise la factorielle aux réels strictement positifs :

$$\Gamma(t) = \int_0^{+\infty} x^{t-1} e^{-x}\, dx, \qquad t > 0.$$

Elle vérifie la relation de récurrence :

$$\Gamma(t+1) = t\,\Gamma(t),$$

et pour tout entier $n \geq 1$,

$$\Gamma(n) = (n-1)!.$$

Quelques aspects remarquables.

— Les lois **exponentielle** et **géométrique** sont les seules à présenter la *propriété d'absence de mémoire*.

— La **loi de Poisson** sert à modéliser des *comptages* - c'est-à-dire le nombre d'événements survenant dans un intervalle de temps donné.

— La **loi Gamma/Erlang** apparaît comme la somme de plusieurs exponentielles : elle décrit par exemple le temps d'occurrence du k-ième événement dans un processus de Poisson.

Propriété 1.5.1 (Absence de mémoire). *La loi exponentielle et la loi géométrique vérifient l'absence de mémoire :*

$$\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t), \qquad s,t \geq 0, \quad X \sim \mathrm{Exp}(\lambda),$$

et $\mathbb{P}(G > m+n \mid G > m) = \mathbb{P}(G > n)$ *pour* $G \sim \mathrm{Geom}(p)$.

Vu en TD. Exercices typiques : montrer que cette propriété caractérise ces lois. Application : "l'attente résiduelle" dans un système où les arrivées sont exponentielles.

Proposition 1.5.1 (Discrétisation d'une exponentielle). *Soit* $X \sim \mathrm{Exp}(\lambda)$ *et* $Y = \lfloor X \rfloor$ *la partie entière de* X. *Alors* Y *suit une loi géométrique sur* $\mathbb{N}$ *de paramètre* $p = 1 - e^{-\lambda}$ *:*

$$\mathbb{P}(Y = n) = (1 - e^{-\lambda})\, e^{-\lambda n}, \qquad n = 0, 1, 2, \ldots$$

Indications de preuves. Par propriété de la partie entière, $\{Y = n\} = \{n \leq X < n+1\}$. Ainsi, exprimer le lien entre $\mathbb{P}(Y = n)$ et la fonction de répartition de X, notée F_X. Par opérations, reconnaître la densité d'une loi géométrique. □

Théorème 1.5.1 (Caractérisation de la loi exponentielle par l'absence de mémoire). *Soit* X *une variable aléatoire réelle positive telle que, pour tout* $t, h \geq 0$,

$$\mathbb{P}(X > t+h \mid X > t) = \mathbb{P}(X > h). \tag{1.1}$$

Alors X *suit nécessairement une loi exponentielle de paramètre* $\lambda > 0$, *c'est-à-dire :*

$$\mathbb{P}(X > t) = e^{-\lambda t}, \qquad t \geq 0.$$

Preuve. On note $G(t) := \mathbb{P}(X > t)$ la fonction de survie. Puisque nous avons, par formule de Bayes, l'égalité suivante :

$$\mathbb{P}(X > t+h \mid X > t) = \frac{\mathbb{P}(X > t+h,\, X > t)}{\mathbb{P}(X > t)} = \frac{\mathbb{P}(X > t+h)}{\mathbb{P}(X > t)},$$

alors l'hypothèse (1.1) s'écrit :

$$\frac{G(t+h)}{G(t)} = G(h), \quad \text{pour tout } t, h \geq 0,$$

soit encore

$$G(t+h) = G(t)\, G(h). \tag{1.2}$$

1. Monotonicité. G étant le complément de la fonction de répartition (*i.e.*, $G = 1 - F$) alors G est *décroissante*. On a de plus $G(0) = 1$ et $\lim_{t \to \infty} G(t) = 0$.

2. Passage à une équation différentielle. On retranche à (1.2) le terme $G(t)$ pour faire apparaître de par et d'autre de l'égalité un taux d'accroissement. Ainsi :

$$(1.2) \iff G(t+h) - G(t) = G(t)G(h) - G(t).$$

En divisant par $h \neq 0$ et en factorisant par $G(t)$ le membre à droite, on obtient :

$$\frac{G(t+h) - G(t)}{h} = G(t)\frac{G(h) - 1}{h} \tag{1.3}$$

Or, il faut relever que

$$G'(0) = \lim_{h \to 0^+} \frac{G(h) - 1}{h} = \lim_{h \to 0^+} \frac{G(h) - G(0)}{h}.$$

Puisque G est décroissante, on sait donc que ce nombre dérivé sera négatif. Tant qu'à faire, autant poser :

$$G'(0) := -\lambda, \quad \text{avec } \lambda \geq 0.$$

Par passage à la limite $h \to 0$:

$$\lim_{h\to 0} \frac{G(t+h) - G(t)}{h} = \lim_{h\to 0} G(t)\frac{G(h) - G(0)}{h},$$

et par définition du nombre dérivé (limite du taux d'accroissement), l'équation (1.3) donne :

$$G'(t) = G(t)\, G(0).$$

On en vient donc à résoudre l'équation différentielle avec condition de bord bien définie :

$$\begin{cases} G'(t) = -\lambda G(t), \\ G(0) = 1. \end{cases} \tag{1.4}$$

3. Résolution et unicité. Par théorème de Cauchy, le système (1.4) admet une unique solution continue, s'exprimant comme :

$$G(t) = e^{-\lambda t}, \quad t \geq 0.$$

On rappelle que la décroissance de G impose $\lambda > 0$! Cette fonction satisfait bien (1.2) :

$$G(t+h) = e^{-\lambda(t+h)} = e^{-\lambda t}e^{-\lambda h} = G(t)G(h).$$

Ainsi, par caractérisation de la fonction de survie (qui découle de la caractérisation par fonction de répartition), X suit la loi $\mathrm{Exp}(\lambda)$.

4. Unicité fonctionnelle. Si l'on suppose G continue, l'équation multiplicative (1.2) n'admet que des solutions exponentielles (théorème de Cauchy multiplicatif). La loi exponentielle est donc la *seule* loi continue vérifiant l'absence de mémoire. □

Remarque 3. Quelques explications :

- $G(t) = \mathbb{P}(X > t)$ est la *fonction de survie* : elle mesure la probabilité que l'événement ne soit pas encore survenu au temps t.
- La relation $G(t+h) = G(t)G(h)$ exprime que *le passé n'influence pas le futur* : la probabilité de "survivre encore h unités de temps" ne dépend que de h, pas de t.
- Le passage à l'équation différentielle montre que la pente relative $\frac{G'(t)}{G(t)}$ est constante, ce qui caractérise les décroissances exponentielles.
- Le paramètre λ correspond au **taux instantané de défaillance constant** : en moyenne, un événement a une probabilité $\lambda h + o(h)$ de se produire durant $[t, t+h]$.
- C'est cette propriété d'"absence de mémoire" qui rend la loi exponentielle **fondamentale** dans la modélisation des files d'attente et des processus de Poisson.

Chapitre 2

Rappels essentiels sur les chaînes de Markov

2.1 Vue d'ensemble

Définition 2.1.1 (Chaîne de Markov en temps discret)**.** *Une suite de variables aléatoires* $(X_n)_{n\geq 0}$ *à valeurs dans un espace fini ou dénombrable* $\mathcal{S}$ *est dite* chaîne de Markov *si, pour tout* $n \geq 0$ *et pour tous les états* $i_0, i_1, \ldots, i_{n+1} \in \mathcal{S}$ *:*

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n).$$

Autrement dit, le futur ne dépend du passé qu'au travers de l'état présent : c'est la propriété de Markov.

Remarque 4 (Intuition). Un processus de Markov modélise l'évolution d'un système qui "oublie" son histoire. Ce qui compte, c'est l'état actuel : par exemple, dans une file $M/M/1$, la probabilité d'une arrivée ou d'un service ne dépend que du nombre de clients présents, pas du moment d'arrivée de chacun.

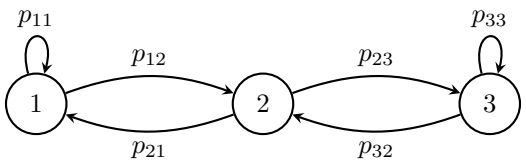
Définition 2.1.2 (Matrice de transition)**.** *Pour une chaîne de Markov à espace d'états* $\mathcal{S} = \{1, \ldots, N\}$*, on définit la matrice stochastique*

$$P = (p_{ij})_{1\leq i,j\leq N}, \qquad p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i).$$

Chaque ligne de P *est une loi de probabilité :*

$$p_{ij} \geq 0, \qquad \sum_j p_{ij} = 1.$$

Un exemple. Pour un espace d'état $\mathcal{S} = \{1, 2, 3\}$, on aura comme représentation graphique :



Chaque flèche correspond à une transition possible entre états, étiquetée par sa probabilité. Ce graphe encode toute la dynamique du système.

Définition 2.1.3 (Loi initiale et évolution). *Si $\mu^{(0)}$ est la loi de X_0 sous forme de vecteur ligne, la loi de X_n est donnée par :*

$$\mu^{(n)} = \mu^{(0)} P^n.$$

Ainsi, P^n contient les probabilités de transition à n pas :

$$(P^n)_{ij} = \mathbb{P}(X_n = j \mid X_0 = i).$$

Définition 2.1.4 (Loi stationnaire). *Une probabilité $\pi = (\pi_i)_{i \in \mathcal{S}}$ est dite* stationnaire *si :*

$$\pi = \pi P, \qquad \text{avec } \sum_i \pi_i = 1,\ \pi_i \geq 0.$$

Autrement dit, si le système est initialement distribué selon π, alors la distribution reste inchangée à chaque pas.

Proposition 2.1.1 (Existence et unicité). *Si la chaîne est* irréductible *(tous les états communiquent) et* récurrente positive*, alors il existe une loi stationnaire unique π, et :*

$$\lim_{n \to \infty} P^n = \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix}.$$

Remarque 5. Les propriétés de régularité reposent sur ces notions clés :

- **Irréductibilité :** tout état est accessible à partir de tout autre.
- **Apériodicité :** un état i est apériodique si le plus grand commun diviseur des temps de retour possibles en i vaut 1.
- **Récurrence positive :** l'espérance du temps de retour en un état est finie.

Ces propriétés garantissent la convergence vers une distribution stationnaire unique.

Remarque 6 (Interprétation temporelle). Pour une chaîne irréductible et apériodique, la probabilité d'être dans un état j au temps n tend vers π_j indépendamment de la condition initiale. C'est une *convergence vers l'équilibre*.

2.2 Propriétés structurelles des chaînes de Markov

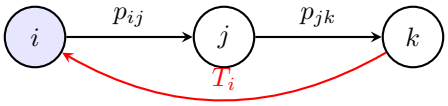
Définition 2.2.1 (Temps d'arrêt). *Soit $(X_n)_{n \geq 0}$ une chaîne de Markov sur un espace d'états $\mathcal{S}$. Une variable aléatoire $T : \Omega \to \mathbb{N} \cup \{\infty\}$ est dite* temps d'arrêt *si, pour tout $n \geq 0$,*

$$\{T = n\} \in \sigma(X_0, X_1, \ldots, X_n).$$

Autrement dit, la connaissance de $(X_0, \ldots, X_n)$ suffit pour savoir si l'arrêt a lieu à l'instant n. Exemple : le ***temps de premier retour*** *en un état i,*

$$T_i = \inf\{n \geq 1 : X_n = i\},$$

est un temps d'arrêt.

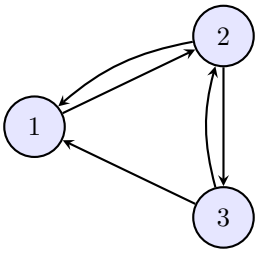


Le temps d'arrêt T_i est le premier
instant où la chaîne revient en i.

Définition 2.2.2 (Irréductibilité)**.** *Une chaîne de Markov est dite* irréductible *si tout état est accessible à partir de tout autre :*

$$\forall\, i, j \in \mathcal{S},\ \exists\, n \geq 0 \text{ tel que } (P^n)_{ij} > 0.$$

Cela signifie que le graphe des transitions est fortement connexe.

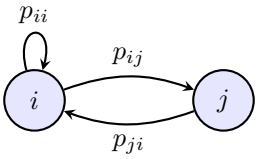


Chaque état peut être atteint depuis n'importe quel autre : le graphe est connexe.

Définition 2.2.3 (Apériodicité)**.** *Un état $i \in \mathcal{S}$ est dit* apériodique *si le plus grand commun diviseur des temps de retour possibles est* 1 *:*

$$\text{pgcd}\{n \geq 1 : (P^n)_{ii} > 0\} = 1.$$

Une chaîne irréductible est apériodique si tous ses états le sont.



Présence de transitions de durées variées (dont des boucles directes) $\Rightarrow$ absence de périodicité.

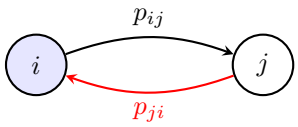
Définition 2.2.4 (Récurrence et récurrence positive)**.** *Un état i est dit* récurrent *si la probabilité de revenir en i est* 1 *:*

$$\mathbb{P}_i(T_i < \infty) = 1.$$

Il est récurrent positif *si l'espérance du temps de retour est finie :*

$$\mathbb{E}_i[T_i] < \infty.$$

Une chaîne est récurrente positive si tous ses états le sont.



L'état i est récurrent si la chaîne revient toujours en i, et positif si le temps moyen de retour est fini.

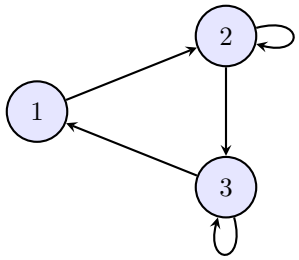
Définition 2.2.5 (Ergodicité)**.** *Une chaîne de Markov est dite* ergodique *si elle est :*

irréductible, apériodique et récurrente positive.

Dans ce cas, il existe une loi stationnaire unique π, et pour tout état j :

$$\forall\, i \in \mathcal{S},\ \lim_{n\to\infty} (P^n)_{ij} = \pi_j.$$

Autrement dit, la distribution de X_n converge vers π indépendamment de la condition initiale.



Une chaîne ergodique converge vers une loi stationnaire unique π.

2.3 Résultats fondamentaux sur les chaînes de Markov

Théorème 2.3.1 (Convergence vers la loi stationnaire)**.** *Soit $(X_n)_{n\geq 0}$ une chaîne de Markov à espace d'états fini $\mathcal{S}$, de matrice de transition P.*
Si la chaîne est ***ergodique*** *(c'est-à-dire irréductible, apériodique et récurrente positive), alors il existe une unique loi stationnaire π telle que :*

$$\pi = \pi P, \qquad \sum_{i\in\mathcal{S}} \pi_i = 1.$$

De plus, pour tout état initial $i \in \mathcal{S}$,

$$\forall\, j \in \mathcal{S},\ \lim_{n\to\infty} (P^n)_{ij} = \pi_j.$$

Autrement dit, la distribution de X_n converge vers π indépendamment de la condition initiale :

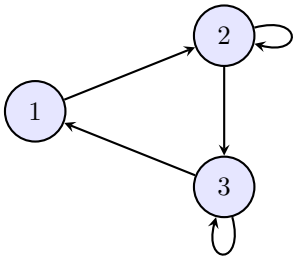
$$\lim_{n\to\infty} \mathbb{P}(X_n = j \mid X_0 = i) = \pi_j.$$

Remarque 7. Le vecteur π est la **mesure invariante à gauche** de P, car :

$$\pi P = \pi.$$

Il est parfois appelé *vecteur propre gauche associé à la valeur propre* 1. Le vecteur colonne $\mathbf{1} = (1,\dots,1)^\top$ est quant à lui le vecteur propre droit associé à la même valeur propre :

$$P\mathbf{1} = \mathbf{1}.$$



Sous ergodicité, la distribution converge vers une mesure stationnaire π, invariante à gauche de la matrice P.

Théorème 2.3.2 (Ergodicité forte — loi des grands nombres de Birkhoff)**.** *Soit $(X_n)_{n\geq 0}$ une chaîne de Markov ergodique, de loi stationnaire π. Alors, pour toute fonction mesurable $f : \mathcal{S} \to \mathbb{R}$ telle que $\mathbb{E}_\pi[|f(X_0)|] < \infty$, on a :*

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \xrightarrow[n\to\infty]{p.s.} \mathbb{E}_\pi[f(X_0)] = \sum_{i\in\mathcal{S}} f(i)\pi_i.$$

Autrement dit, la moyenne empirique des valeurs de $f(X_n)$ converge presque sûrement vers son espérance stationnaire.

Remarque 8 (Interprétation statistique). Ce résultat signifie qu'à long terme, la chaîne "explore" les états selon la distribution stationnaire π. En particulier, la fréquence de visite de l'état i satisfait :

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}_{\{X_k=i\}} \xrightarrow[n\to\infty]{} \pi_i.$$

Théorème 2.3.3 (Existence d'une mesure stationnaire)**.** *Toute matrice de transition P (stochastique) admet au moins une mesure stationnaire π, c'est-à-dire un vecteur de probabilité tel que $\pi P = \pi$.*
Si la chaîne est irréductible et récurrente positive, cette mesure est unique.

Définition 2.3.1 (Réversibilité)**.** *Une chaîne de Markov de matrice de transition P et de loi stationnaire π est dite* réversible *si :*

$$\forall\, i, j \in \mathcal{S}, \quad \pi_i\, p_{ij} = \pi_j\, p_{ji}.$$

Cette relation est appelée ***équilibre détaillé****. Elle traduit le fait que, en régime stationnaire, le flux de probabilité entre i et j est équilibré dans les deux sens.*

Exemple 2.3.1 (Chaîne de file d'attente $M/M/1$)**.** *Dans le modèle $M/M/1$, on a :*

$$p_{n,n+1} = \lambda, \qquad p_{n,n-1} = \mu.$$

La loi stationnaire $\pi_n = (1-\rho)\rho^n$ (avec $\rho = \lambda/\mu$) vérifie bien :

$$\pi_n \lambda = \pi_{n+1}\mu.$$

Ainsi, la chaîne est réversible : les flux entre états voisins s'équilibrent.

2.4 Synthèse - Chaîne de Markov

Une **chaîne de Markov** est un processus stochastique $(X_n)_{n\geq 0}$ à temps discret, prenant ses valeurs dans un espace d'états $\mathcal{E}$, et vérifiant la **propriété de Markov** :

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n).$$

Autrement dit, **le futur dépend uniquement du présent, et non du passé**. Ce caractère sans mémoire rend le processus Markovien particulièrement adapté à la modélisation de systèmes dynamiques comme les files d'attente ou les processus de naissance et mort.

L'évolution de la chaîne est décrite par une matrice de transition $P = (p_{ij})_{i,j\in\mathcal{E}}$ où

$$p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i),$$

et la loi à l'instant n est donnée par le vecteur de probabilités $\pi^{(n)} = \pi^{(0)}P^n$.

Mesure stationnaire et comportement asymptotique

Une **mesure stationnaire** (ou **distribution invariante**) est un vecteur de probabilité $\pi = (\pi_i)_{i\in\mathcal{E}}$ satisfaisant :

$$\pi P = \pi, \qquad \sum_{i\in\mathcal{E}} \pi_i = 1.$$

Cela signifie que si X_0 suit la loi π, alors pour tout n, X_n suit la même loi : la distribution est invariante par la dynamique du processus.

Sous certaines conditions, la chaîne converge vers cette loi stationnaire, indépendamment de la condition initiale.

Synthèse des propriétés fondamentales

- **Irréductibilité** ⇒ la chaîne peut atteindre tout état à partir de tout autre état.
- **Récurrence positive** ⇒ existence d'une mesure stationnaire finie π.
- **Apériodicité** ⇒ la chaîne ne présente pas de comportement cyclique (pas d'oscillation périodique).
- **Ergodicité** = combinaison des trois précédentes ⇒

$$\pi^{(n)} \xrightarrow[n\to\infty]{} \pi, \qquad \text{et} \qquad \frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \xrightarrow[n\to\infty]{} \mathbb{E}_\pi[f],$$

pour toute fonction test f.

En résumé, une chaîne de Markov ergodique converge vers une unique loi stationnaire π, solution de l'équation

$$\pi = \pi P,$$

ce qui traduit l'équilibre statistique du système étudié.

Chapitre 3

Chaînes de Markov à temps continu

3.1 Définition et intuition

Définition 3.1.1 (Chaîne de Markov à temps continu (CTMC))**.** *Un* processus stochastique $(X_t)_{t\geq 0}$ *à valeurs dans un ensemble d'états dénombrable* $\mathcal{S}$ *est dit être une* chaîne de Markov à temps continu *si, pour tout* $t, s \geq 0$ *et pour tous* $i, j \in \mathcal{S}$*,*

$$\mathbb{P}(X_{t+s} = j \mid X_s = i, (X_u)_{u<s}) = \mathbb{P}(X_{t+s} = j \mid X_s = i),$$

c'est-à-dire que le futur du processus dépend uniquement de son présent, et non du passé.

Remarque 9. Cette propriété est appelée *propriété de Markov*. Elle formalise la mémoire nulle du système : seul l'état actuel X_t est pertinent pour prédire l'évolution future.

Une CTMC décrit l'évolution d'un système qui se déplace aléatoirement entre des états, en passant un temps aléatoire (généralement exponentiel) dans chaque état avant de sauter vers un autre. Ce cadre est central pour modéliser les *files d'attente*, les *pannes/réparations*, ou les *systèmes à transitions discontinues*.

3.2 Probabilités de transition et matrice de transition

Définition 3.2.1 (Probabilités de transition)**.** *On définit les probabilités de transition par*

$$p_{ij}(t) := \mathbb{P}(X_t = j \mid X_0 = i), \qquad i, j \in \mathcal{S},\ t \geq 0.$$

Les $p_{ij}(t)$ forment une *matrice de transition* $P(t) = (p_{ij}(t))_{i,j\in\mathcal{S}}$, qui satisfait :

$$P(0) = I_d, \qquad P(t+s) = P(t)P(s), \quad t, s \geq 0.$$

Remarque 10. Cette propriété de composition traduit la *mémoire nulle* : l'évolution sur $[0, t+s]$ se décompose en celle sur $[0, s]$ puis sur $[s, t+s]$.

Définition 3.2.2 (Homogénéité)**.** *La chaîne est dite* homogène dans le temps *si* $p_{ij}(t, s) := \mathbb{P}(X_t = j \mid X_s = i)$ *ne dépend que de la durée* $t - s$*, c'est-à-dire :*

$$p_{ij}(t, s) = p_{ij}(t - s).$$

3.3 Générateur infinitésimal

Définition 3.3.1 (Générateur infinitésimal)**.** *On définit le* générateur infinitésimal $Q = (q_{ij})_{i,j\in\mathcal{S}}$ *par :*

$$q_{ij} = \lim_{t\to 0^-} \frac{p_{ij}(t) - \mathbb{1}_{\{i=j\}}}{t}.$$

Propriété 3.3.1. *Deux points sur la structure de Q :*

- $q_{ij} \geq 0$ *pour $i \neq j$ (taux de transition de i vers j).*
- $q_{ii} = -\sum_{j\neq i} q_{ij}$ *(chaque ligne somme à zéro).*

Remarque 11. Les coefficients q_{ij} ont une interprétation physique : ils représentent la *vitesse moyenne de transition* entre états. Ainsi, la probabilité de quitter i dans un court intervalle $\mathrm{d}t$ est $q_i\,\mathrm{d}t$, où $q_i = -q_{ii}$.

3.4 Équations de Chapman–Kolmogorov

Théorème 3.4.1 (Équations de Chapman-Kolmogorov)**.** *Pour un processus markovien homogène $(X_t)_{t\geq 0}$ à espace d'états $\mathcal{S}$, on définit :*

$$p_{ij}(t) = \mathbb{P}(X_t = j \mid X_0 = i), \qquad i,j \in \mathcal{S}.$$

La famille de matrices $P(t) = (p_{ij}(t))_{i,j\in\mathcal{S}}$ constitue alors la matrice de transition *du processus.*

Pour toute chaîne de Markov homogène à temps continu, on a :

$$\forall\, t,h \geq 0,\ P(t+h) = P(t)P(h).$$

Autrement dit, la probabilité de transition sur l'intervalle $[0, t+h]$ s'obtient comme la composition des transitions successives sur $[0,t]$ et $[t,t+h]$.

Preuve. Considérons un processus $(X_t)_{t\geq 0}$ markovien à espace d'états $\mathcal{S}$. Par définition de la propriété de Markov :

$$\mathbb{P}(X_{t+h} = j \mid X_t = k, X_s = i,\ s < t) = \mathbb{P}(X_{t+h} = j \mid X_t = k).$$

En appliquant la formule des probabilités totales sur l'état intermédiaire k, on obtient :

$$\begin{aligned} p_{ij}(t+h) = \mathbb{P}(X_{t+h} = j \mid X_0 = i) &= \sum_{k\in\mathcal{S}} \mathbb{P}(X_{t+h} = j,\ X_t = k \mid X_0 = i) \\ &= \sum_{k\in\mathcal{S}} \mathbb{P}(X_{t+h} = j \mid X_t = k)\,\mathbb{P}(X_t = k \mid X_0 = i) \\ &= \sum_{k\in\mathcal{S}} p_{ik}(t)\,p_{kj}(h). \end{aligned}$$

En notation matricielle, cela s'écrit :

$$P(t+h) = P(t)P(h),$$

ce qui établit la propriété annoncée. □

Remarque 12 (Sur les hypothèses utilisées). Deux propriétés fondamentales justifient les simplifications faites dans la preuve :

(i) **Propriété de Markov :**

$$\mathbb{P}(X_{t+h} = j \mid X_t = k, X_0 = i) = \mathbb{P}(X_{t+h} = j \mid X_t = k),$$

car le futur ne dépend du passé que via l'état présent.

(ii) **Homogénéité temporelle :**

$$\mathbb{P}(X_{t+h} = j \mid X_t = k) = \mathbb{P}(X_h = j \mid X_0 = k) = p_{kj}(h),$$

c'est-à-dire que les probabilités de transition ne dépendent que de la durée écoulée h, et non des instants absolus.

Ces deux hypothèses permettent d'écrire la relation de Chapman–Kolmogorov :

$$p_{ij}(t+h) = \sum_{k \in \mathcal{S}} p_{ik}(t) p_{kj}(h), \quad \text{soit matriciellement} \quad P(t+h) = P(t)P(h).$$

Théorème 3.4.2 (Équations différentielles de Kolmogorov)**.** *Les probabilités de transition vérifient :*

$$\frac{\mathrm{d}}{\mathrm{d}t} P(t) = P(t)Q = QP(t),$$

et donc, pour chaque (i, j)*,*

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{ij}(t) = \sum_{k \in \mathcal{S}} p_{ik}(t) q_{kj}.$$

Preuve.
On part de la **relation de Chapman–Kolmogorov** :

$$\forall\, t, h \geq 0,\ P(t+h) = P(t)P(h),$$

où $P(t) = (p_{ij}(t))$ désigne la matrice des probabilités de transition.

Par définition du générateur infinitésimal Q :

$$Q = \lim_{h \to 0^+} \frac{P(h) - I}{h},$$

soit, pour h petit :

$$P(h) = I + hQ + o(h).$$

En reportant dans Chapman–Kolmogorov :

$$P(t+h) = P(t)(I + hQ + o(h)) = P(t) + hP(t)Q + o(h).$$

On obtient alors :

$$\frac{P(t+h) - P(t)}{h} = P(t)Q + \frac{o(h)}{h}.$$

En faisant tendre $h \to 0$, on en déduit l'**équation différentielle de Kolmogorov avant** :

$$\frac{\mathrm{d}}{\mathrm{d}t} P(t) = P(t)Q.$$

De manière analogue, en partant de $P(t+h) = P(h)P(t)$, on obtient l'**équation de Kolmogorov arrière** :

$$\frac{\mathrm{d}}{\mathrm{d}t} P(t) = QP(t).$$

Ces deux équations expriment que la dynamique de la matrice de transition $P(t)$ est entièrement gouvernée par le générateur infinitésimal Q. La première décrit la *propagation des probabilités vers le futur* (forward equation), la seconde vers le passé (backward equation). □

Remarque 13. La dérivée de $P(t)$ n'est pas égale à Q, car Q ne donne que la **vitesse instantanée de transition** au temps initial $t = 0$. Or, à un instant $t > 0$, les probabilités d'être dans tel ou tel état dépendent de la combinaison des transitions passées, et non seulement de Q.

C'est pourquoi la dérivée doit s'écrire sous forme d'une **convolution matricielle** :

$$\frac{\mathrm{d}}{\mathrm{d}t}P(t) = P(t)Q,$$

ce qui conduit directement aux **équations différentielles de Kolmogorov**.

Remarque 14. Ces équations sont fondamentales pour relier les probabilités de transition $p_{ij}(t)$ au générateur infinitésimal Q. Elles jouent un rôle clé dans la démonstration des équations stationnaires et dans la modélisation des files d'attente.

3.5 Distribution stationnaire et équilibre

Définition 3.5.1 (Distribution stationnaire)**.** *Une distribution de probabilité $\pi = (\pi_i)_{i\in\mathcal{S}}$ est dite* stationnaire *si elle satisfait :*

$$\pi Q = 0, \qquad \sum_{i\in\mathcal{S}} \pi_i = 1.$$

Théorème 3.5.1 (Caractérisation stationnaire)**.** *Si π est une distribution stationnaire, alors pour tout $t \geq 0$:*

$$\pi P(t) = \pi.$$

Preuve. En dérivant $\pi P(t)$ et en utilisant $\pi Q = 0$ ainsi que $\frac{\mathrm{d}}{\mathrm{d}t}P(t) = QP(t)$, on obtient que $\frac{\mathrm{d}}{\mathrm{d}t}(\pi P(t)) = 0$, donc $\pi P(t) = \pi$. □

Remarque 15. La distribution stationnaire correspond à l'équilibre probabiliste du système : si le processus démarre selon π, sa loi ne change plus au cours du temps.

Remarque 16 (Recherche pratique d'une mesure stationnaire)*.* En pratique, la détermination d'une mesure stationnaire π repose rarement sur le calcul direct de $P(t) = e^{tQ}$, car cette exponentielle de matrice est généralement intractable dès que l'espace d'états est de taille moyenne. On exploite donc la relation stationnaire $\pi Q = 0$, qui découle de la dérivée de l'équation $\pi P(t) = \pi$. Ainsi, le problème se ramène à la résolution d'un système linéaire homogène sous contrainte de normalisation :

$$\pi Q = 0, \qquad \sum_{i\in\mathcal{S}} \pi_i = 1.$$

Cette approche est à la fois plus simple numériquement et plus interprétable : chaque équation de la forme

$$\sum_j \pi_j q_{ji} = 0$$

exprime l'*équilibre des flux* entre les états. Autrement dit, dans un état stationnaire, le flux moyen de probabilité entrant dans chaque état est égal au flux sortant.

3.6 Propriétés de convergence et interprétation

Théorème 3.6.1 (Convergence vers l'équilibre)**.** *Si la chaîne est irréductible, apériodique et positive récurrente, alors :*

$$\lim_{t\to\infty} P(t) = \mathbb{1}\pi,$$

c'est-à-dire que, quelle que soit la loi initiale μ,

$$\lim_{t\to\infty} \mu P(t) = \pi.$$

Remarque 17. Cela signifie qu'à long terme, le système "oublie" son état initial : les proportions de temps passées dans chaque état convergent vers les poids stationnaires π_i.

3.7 Temps d'arrêt et propriétés fondamentales de Markov

Pour conclure cette section, nous introduisons deux notions essentielles à la théorie des chaînes de Markov : le *temps d'arrêt*, qui formalise la notion de "moment aléatoire d'observation", et les *propriétés de Markov*, faible et forte, qui généralisent l'idée d'absence de mémoire. Ces concepts jouent un rôle central dans l'étude des processus de sauts, comme le processus de Poisson homogène, et dans l'analyse des temps de retour ou de franchissement d'un état.

Définition 3.7.1 (Temps d'arrêt)**.** *Soit $(X_t)_{t\geq 0}$ un processus stochastique défini sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, et $(\mathcal{F}_t)_{t\geq 0}$ une filtration associée représentant l'information disponible jusqu'au temps t. Une variable aléatoire $T : \Omega \to [0, +\infty]$ est appelée* temps d'arrêt *(ou* temps d'arrêt aléatoire*) si, pour tout $t \geq 0$,*

$$\{T \leq t\} \in \mathcal{F}_t.$$

Autrement dit, au temps t, on peut savoir si l'événement "le temps d'arrêt T s'est produit" est déjà survenu ou non.

Exemples.

- Le premier instant où un processus de file d'attente devient vide : $T = \inf\{t > 0 : N_t = 0\}$.
- Le premier temps de dépassement d'un seuil fixé : $T = \inf\{t \geq 0 : X_t > a\}$.

Propriété 3.7.1 (Propriété de Markov (faible et forte))**.** *Soit $(X_t)_{t\geq 0}$ un processus stochastique à valeurs dans un espace d'états $\mathcal{S}$, défini sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$.*

(i) ***Propriété faible de Markov.*** *On dit que (X_t) vérifie la* propriété de Markov faible *si, pour tout $s, t \geq 0$ et tout état $x \in \mathcal{S}$,*

$$\mathbb{P}(X_{t+s} \in A \mid \mathcal{F}_t) = \mathbb{P}(X_{t+s} \in A \mid X_t), \qquad \forall A \subseteq \mathcal{S}.$$

Autrement dit, le futur du processus ne dépend du passé qu'à travers l'état présent X_t. *En notation transitionnelle, cela se traduit par :*

$$\mathbb{P}(X_{t+s} \in A \mid X_t = x) = P_s(x, A),$$

où $(P_s)_{s\geq 0}$ désigne la famille des probabilités de transition.

(ii) ***Propriété forte de Markov.*** *Soit τ un* temps d'arrêt *pour la filtration naturelle $(\mathcal{F}_t)_{t\geq 0}$. On dit que (X_t) vérifie la* propriété forte de Markov *si, pour tout $t \geq 0$ et tout ensemble mesurable $A \subseteq \mathcal{S}$,*

$$\mathbb{P}\big(X_{\tau+t} \in A \mid \mathcal{F}_\tau\big) = \mathbb{P}\big(X_{\tau+t} \in A \mid X_\tau\big), \quad \text{presque sûrement.}$$

Ainsi, la propriété de Markov reste vraie même lorsqu'on redémarre le processus au temps aléatoire τ.

Intuitions.

- La propriété *faible* formalise l'idée d'absence de mémoire : connaître l'état actuel suffit à décrire le futur.
- La propriété *forte* généralise cette idée à des temps d'arrêt aléatoires, ce qui est crucial pour l'étude des processus de sauts (comme le processus de Poisson) et des temps de retour dans les chaînes de Markov.

Chapitre 4

Processus aléatoires pour file d'attente

4.1 Processus ponctuels et temps de sauts

Définition 4.1.1 (Processus stochastique)**.** *On appelle* processus stochastique *une suite (ou famille continue) de variables aléatoires indexées par le temps, qui décrit l'évolution aléatoire d'un système. Étant donné le temps t, on écrit*

$$(X_t)_{t\geq 0},$$

définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans un espace d'états $\mathcal{S}$.

Définition 4.1.2 (Processus de comptage et temps de saut)**.** *Un* processus de comptage $(N_t)_{t\geq 0}$ *est un processus stochastique à valeurs entières vérifiant les propriétés suivantes :*

(i) $N_t \in \mathbb{N}$ *pour tout* $t \geq 0$,

(ii) $N_0 = 0$,

(iii) les trajectoires $t \mapsto N_t(\omega)$ *sont* càdlàg *(continues à droite, à limites à gauche),*

(iv) le processus est ***croissant par sauts unitaires****, c'est-à-dire :*

$$N_t - N_{t^-} \in \{0, 1\}, \quad \textit{et} \quad N_t \textit{ est non décroissant en } t.$$

On définit les temps de sauts $(T_n)_{n\geq 1}$ *par*

$$T_n := \inf\{t \geq 0 : N_t \geq n\}, \qquad n \geq 1,$$

avec la convention $\inf \varnothing = +\infty$*. Réciproquement,*

$$N_t \;=\; \sum_{j\geq 1} \mathbb{1}_{\{T_j \leq t\}}, \qquad t \geq 0.$$

Intuition derrière de telles définitions.

— Les temps (T_n) représentent les **instants exacts** où ces événements se produisent : T_1 est le temps de la première arrivée, T_2 celui de la deuxième, etc.

— Le processus (N_t) **compte le nombre d'événements** (ou de "tops") survenus jusqu'à l'instant t. Par exemple : nombre de clients arrivés dans une file d'attente, nombre de pannes dans une machine, nombre d'appels reçus.

— Cette correspondance entre (N_t) et (T_n) est fondamentale : elle permet de passer d'une vision "en temps continu" (les événements arrivent au fil du temps) à une vision "par événements" (on observe la suite de leurs instants d'apparition).

Proposition 4.1.1 (Équivalences de base entre N et (T_n))**.** *Pour tout $t \geq 0$ et $0 \leq s < t$:*

$$\{N_t \geq n\} = \{T_n \leq t\},$$
$$\{N_t = n\} = \{T_n \leq t < T_{n+1}\},$$
$$\{N_t \geq n > N_s\} = \{s < T_n \leq t\}.$$

Intuition. Ces identités montrent que la donnée de la loi de N est équivalente à celle de la loi des temps de sauts (T_n) : connaître l'un, c'est connaître l'autre.

Preuve. Par définition de T_n, l'événement $\{N_t \geq n\}$ signifie que le n-ième saut a eu lieu avant t, d'où la première équivalence. La seconde découle de $\{N_t = n\} = \{T_n \leq t\} \cap \{T_{n+1} > t\}$. Enfin, $\{N_t \geq n > N_s\}$ signifie qu'entre s et t on a franchi le niveau n, i.e. $\{s < T_n \leq t\}$. □

Exercice 4.1.1 (Reconstruction par les temps d'inter-arrivées). Poser $S_n := T_n - T_{n-1}$ ($T_0 = 0$). Montrer que $N_t = \max\{n : S_1 + \cdots + S_n \leq t\}$ et que la connaissance de la loi des (S_n) i.i.d détermine la loi de N (processus de renouvellement).

4.2 Processus de Poisson homogène

Définition 4.2.1 (Processus de Poisson homogène (PPH))**.** *Un processus de comptage $(N_t)_{t\geq 0}$ est un* PPH de taux $\lambda > 0$ *si*

(i) $N_0 = 0$ *;*

(ii) les incréments *sont* ***stationnaires*** *et* ***indépendants****, c'est-à-dire :*

$$\forall\, 0 \leq t_1 < t_2 < \cdots < t_n, \quad \textit{les } (N_{t_k} - N_{t_{k-1}}) \textit{ sont indépendants,}$$

et pour tout $s, t \geq 0$,

$$N_{t+s} - N_s \overset{\mathcal{L}}{=} N_t - N_0.$$

Autrement dit, la loi de l'accroissement $N_{t+h} - N_t$ dépend uniquement de la longueur h de l'intervalle, et non de sa position dans le temps.

(iii) pour $h \to 0^-$, $\mathbb{P}(N_h = 1) = \lambda h + o(h)$ et $\mathbb{P}(N_h \geq 2) = o(h)$.

Intuition derrière une telle définition.

— La propriété d'**indépendance des incréments** signifie que le nombre d'événements sur des intervalles disjoints ne dépend pas des autres intervalles. Exemple : le nombre d'arrivées entre 10h et 11h est indépendant de celui entre 11h et 12h.

— La **stationnarité des incréments** veut dire que seule la durée de l'intervalle compte, pas sa position. Ainsi, le nombre d'arrivées dans 5 minutes est distribué de la même façon, qu'on observe de 9h à 9h05 ou de 17h à 17h05.

— Ces deux propriétés combinées modélisent un *flux d'événements aléatoires sans mémoire ni saisonnalité*, c'est-à-dire un processus purement aléatoire mais *régulier en moyenne* : le taux moyen d'arrivées reste constant et chaque sous-intervalle se comporte de façon identique.

— Enfin, la condition (iii) formalise l'idée intuitive qu'en un instant très court, il est pratiquement impossible d'observer plus d'un événement. Cela permet d'obtenir des sauts unitaires (un par un) et une évolution continue par morceaux.

Théorème 4.2.1 (Loi, inter-arrivées, équivalences). *Pour un PPH de taux λ et pour $0 \leq s < t$,*

i) $N_t \sim \mathcal{P}(\lambda t)$,

ii) $N_{t+s} - N_s \sim \mathcal{P}(\lambda t)$,

iii) $N_{t+s} - N_s \perp\!\!\!\perp \sigma(N_u,\, u \leq s)$ (indépendances des accroissements).

où $\sigma(\cdot)$ désigne la tribu engendrée[a].
De plus, les temps inter-arrivées, notés $S_n := T_n - T_{n-1}$, vérifient

$$T_j = \sum_{k=1}^{j} S_k, \quad S_k \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda).$$

Réciproquement, si (S_n) sont i.i.d *exponentielles de paramètre λ, alors (N_t), de termes*

$$N_t \;=\; \sum_{j \geq 1} \mathbb{1}_{\{T_j \leq t\}}$$

pour $t \geq 0$, est un PPH de taux λ.

a. La tribu engendrée par une variable aléatoire X, notée $\sigma(X)$, désigne l'ensemble de tous les événements que l'on peut décrire à partir de la connaissance de X. Elle regroupe tous les événements du type $\{X \in A\}$ pour A mesurable. Dans notre contexte, $\sigma(N_u,\, u \leq s)$ représente toute l'information contenue dans le *passé du processus* jusqu'à l'instant s.

Preuve. La caractérisation locale (iii) entraîne les équations de Kolmogorov pour $p_n(t) = \mathbb{P}(N_t = n)$, d'où $p_n(t) = e^{-\lambda t}(\lambda t)^n / n!$. L'indépendance des accroissements provient de l'additivité et de la propriété de Markov forte. Le lien avec les inter-arrivées est classique via Proposition 4.1.1. □

Intuition. Le PPH combine trois idées : incréments stationnaires, indépendants et sauts unitaires rares ; la loi de Poisson et l'exponentielle (absence de mémoire) en découlent.

Proposition 4.2.1 (Amincissement et superposition des processus de Poisson). *On dispose des deux propriétés fondamentales suivantes :*

*(i) (**Amincissement**) Soit $(N_t)_{t \geq 0}$ un processus de Poisson homogène de taux $\lambda > 0$. À chaque saut T_n, on associe une variable aléatoire $B_n \sim$ Bernoulli(p), indépendante des autres B_m et du processus N. On définit alors :*

$$N_t^{(1)} = \sum_{n \geq 1} \mathbb{1}_{\{T_n \leq t,\, B_n = 1\}}, \qquad N_t^{(2)} = \sum_{n \geq 1} \mathbb{1}_{\{T_n \leq t,\, B_n = 0\}}.$$

Alors :

- *$N^{(1)}$ et $N^{(2)}$ sont deux processus de Poisson homogènes indépendants,*
- *de taux respectifs $p\lambda$ et $(1-p)\lambda$.*

*(ii) (**Superposition**) Si $(N_t^{(1)})$ et $(N_t^{(2)})$ sont deux processus de Poisson homogènes indépendants, de taux λ_1 et λ_2, alors leur somme*

$$N_t := N_t^{(1)} + N_t^{(2)}$$

est un processus de Poisson homogène de taux $\lambda_1 + \lambda_2$.

Preuve. (i) Amincissement. Considérons le processus initial (N_t) avec ses temps de sauts $(T_n)_{n \geq 1}$. La suite (B_n) joue le rôle d'un *filtrage indépendant* des événements : chaque saut est gardé avec probabilité p ou rejeté avec probabilité $1 - p$.

Pour tout $t \geq 0$, conditionnellement à $N_t = n$, on a $N_t^{(1)} \sim \text{Binomial}(n, p)$. Ainsi :

$$\mathbb{P}(N_t^{(1)} = k) = \sum_{n=k}^{\infty} \mathbb{P}(N_t^{(1)} = k \mid N_t = n)\, \mathbb{P}(N_t = n) = \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} e^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

On simplifie :

$$\mathbb{P}(N_t^{(1)} = k) = e^{-\lambda t} \frac{(p\lambda t)^k}{k!} \sum_{n=k}^{\infty} \frac{[(1-p)\lambda t]^{n-k}}{(n-k)!} = e^{-p\lambda t} \frac{(p\lambda t)^k}{k!}.$$

Ainsi $N_t^{(1)} \sim \text{Poisson}(p\lambda t)$. De plus, l'indépendance des incréments découle de celle des incréments du processus d'origine et de l'indépendance des B_n. La même démonstration vaut pour $N_t^{(2)}$, et les deux processus sont indépendants car issus d'un découpage basé sur des tirages indépendants.

(ii) Superposition. Soient deux processus de Poisson indépendants $(N_t^{(1)})$ et $(N_t^{(2)})$. Pour tout $t \geq 0$, la somme $N_t = N_t^{(1)} + N_t^{(2)}$ suit une loi :

$$N_t \sim \text{Poisson}(\lambda_1 t) * \text{Poisson}(\lambda_2 t) = \text{Poisson}((\lambda_1 + \lambda_2)t),$$

puisque la somme de deux variables de Poisson indépendantes est encore une Poisson dont le paramètre est la somme des paramètres.

Les incréments de N sont indépendants et stationnaires comme ceux des processus d'origine. Ainsi, N est un processus de Poisson homogène de taux $\lambda_1 + \lambda_2$. □

Ces opérations modélisent ainsi filtrage (admission/rejet) et agrégation de flux indépendants.

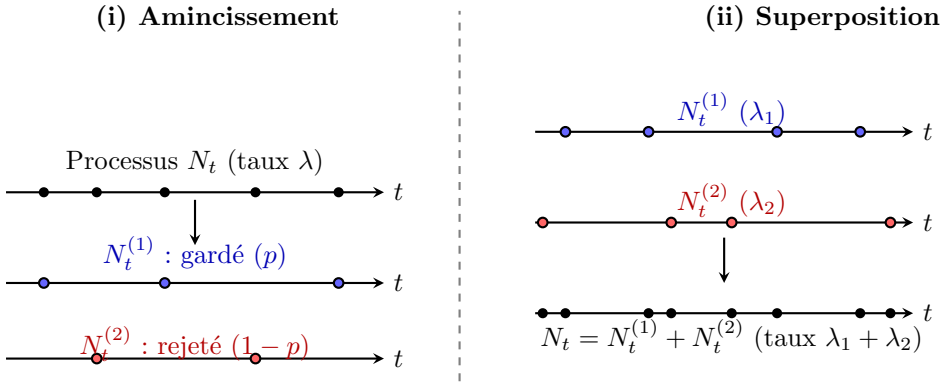


Illustration : à gauche, amincissement (filtrage aléatoire des événements) ; à droite, superposition (fusion de deux processus indépendants).

Exemple 4.2.1 (Formules pratiques). *Pour un PPH de taux λ, $\mathbb{E}[N_t] = \lambda t$, $\text{Var}(N_t) = \lambda t$, et le temps de la k-ième arrivée T_k suit* $\text{Gamma}(k, \lambda)$ *: $\mathbb{E}[T_k] = k/\lambda$.*

Exercice 4.2.1 (Rappel des équivalences $N \leftrightarrow (T_n)$). Utiliser la Proposition 4.1.1 pour démontrer $\mathbb{P}(N_t \geq n > N_s) = \mathbb{P}(s < T_n \leq t)$ et en déduire $\mathbb{P}(N_{t+s} - N_s = k) = \mathbb{P}(T_{n+k} - T_n \leq t < T_{n+k+1} - T_n)$ pour tout n et k (stationnarité des accroissements).

Exemple 4.2.2 (Temps à k arrivées). *Pour un PPH de taux λ, le temps S_k de la k-ième arrivée suit* $\text{Gamma}(k, \lambda)$. *Ainsi, $\mathbb{E}[S_k] = k/\lambda$ et $\text{Var}(S_k) = k/\lambda^2$.*

Chapitre 5

Processus des files d'attente

Motivation. Les systèmes d'attente modélisent l'allocation de ressources partagées : serveurs applicatifs, centres d'appels, *load balancers*, ateliers de production. L'objectif est d'estimer délais, files, pertes, et de *dimensionner* la capacité pour respecter des SLA de délai et de disponibilité.

5.1 Introduction aux files d'attente

5.1.1 Définition et composantes fondamentales

Définition 5.1.1 (File d'attente)**.** *Un* système de file d'attente *est un modèle stochastique décrivant l'évolution de clients (ou tâches) arrivant dans un système pour y recevoir un service assuré par un ou plusieurs serveurs. Le système est défini par :*

- *un* ***processus d'arrivées*** $(A_t)_{t\geq 0}$ *qui décrit le nombre de clients arrivés jusqu'au temps* t *;*
- *une* ***discipline de service*** $(D_t)_{t\geq 0}$ *qui précise l'ordre dans lequel les clients sont servis lorsqu'ils sont en file d'attente ;*
- *Une* ***règle de gestion*** *de la file : FIFO, LIFO, priorité, etc. ;*
- *un* ***nombre*** c ***de serveurs*** *travaillant en parallèle ;*
- *une* ***capacité totale*** K *(optionnelle), qui borne le nombre maximum de clients autorisés dans le système (en attente + en service) ;*
- *une* ***population totale*** L *(optionnelle) de clients potentiels.*

- Dans le langage courant, une file d'attente peut représenter des clients à une caisse, des paquets sur un routeur, des tâches sur un processeur ou encore des appels téléphoniques entrant sur un serveur.
- Le modèle mathématique abstrait permet de :
  - prévoir le temps moyen d'attente,
  - estimer l'occupation des serveurs,
  - optimiser le dimensionnement pour atteindre une contrainte de qualité de service (SLA).

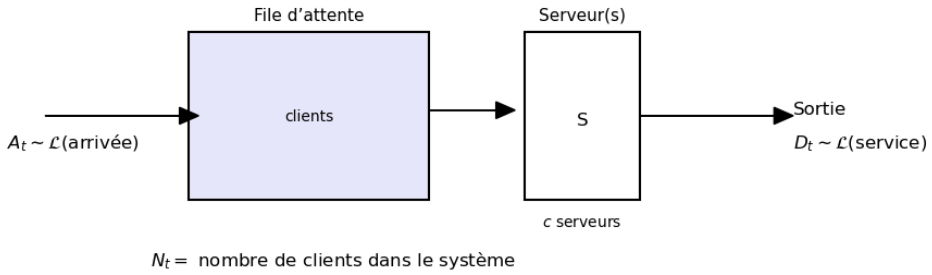
À retenir. Une file d'attente est toujours caractérisée par :

- un *flux d'arrivées*,
- un *flux de services*,
- une *règle d'organisation*,
- et des *ressources finies* (serveurs, capacité).

5.1.2 Notations possibles et motivations

Il existe plusieurs façons de décrire une file d'attente :

— **Par un schéma descriptif.** Exemple : arrivée de clients à un guichet avec une file FIFO illimitée. Ce type de schéma est très utile pour une première modélisation.



— **Par un processus stochastique.** On peut suivre $(N_t)_{t\geq 0}$, le nombre de clients présents dans le système à l'instant t. Ce processus prend ses valeurs dans $\mathbb{N}$ et évolue par sauts unitaires liés aux arrivées et aux départs.

— **Par une notation symbolique.** C'est le rôle de la *notation de Kendall*, qui fournit une description normalisée et compacte du système (voir Section suivante).

À savoir faire.

1. Identifier les composantes d'une file d'attente dans un contexte réel (par ex. serveur informatique, caisse de supermarché, réseau télécom).
2. Traduire cette situation soit par un schéma explicite, soit par un processus (N_t), soit via la notation compacte de Kendall.
3. Comprendre que ces notations ne sont pas redondantes mais complémentaires : l'une pour l'intuition, l'autre pour l'analyse mathématique.

5.2 Notation de Kendall, disciplines de service et métriques

5.2.1 Notation de Kendall

Définition 5.2.1 (Notation de Kendall)**.** *Une file d'attente est décrite par la* notation de Kendall *:*

$$A/B/C\ (/K/Z/L),$$

où :

— *A : loi des arrivées (distribution des temps inter-arrivées) ;*

— *B : loi des services (distribution des durées de service) ;*

— *C : nombre de serveurs en parallèle ;*

— *K (optionnel) : capacité totale maximale du système (en service + en attente) ;*

— *Z (optionnel) : discipline de service (FIFO, LIFO, Priorité, etc.) ;*

— *L (optionnel) : taille de la population totale potentielle.*

Cette notation compacte permet de représenter une grande diversité de systèmes. Pour les lois d'arrivées et de services, la lettre M **désigne "Markovien"** (temps exponentiels), D désigne des durées déterministes, G des durées générales. Il est impératif de savoir reconnaître les composantes d'un système réel et les traduire en notation de Kendall.

5.2.2 Disciplines de service

Définition 5.2.2 (Disciplines courantes)**.** *La* discipline de service *décrit l'ordre dans lequel les clients de la file sont pris en charge. Les principales sont :*

- *FIFO (*First In, First Out*) : les clients sont servis dans l'ordre d'arrivée.*
- *LIFO (*Last In, First Out*) : le client arrivé en dernier est servi en premier.*
- *SIRO (*Service In Random Order*) : ordre aléatoire, chaque client en attente est choisi au hasard.*
- *Priorités : certains clients passent avant d'autres, selon un critère (ex. urgences médicales).*

- La discipline de service influence directement le temps d'attente vu par les clients.
- FIFO est le modèle par défaut en pratique (équité).
- Les autres disciplines apparaissent dans des cas particuliers (stocks, systèmes informatiques, routage de paquets).

5.2.3 Métriques principales

Définition 5.2.3 (Grandeurs de performance)**.** *Soit* $(N_t)_{t\geq 0}$ *un* processus de comptage *représentant le* ***nombre de clients présents dans le système*** *(en attente ou en service) au temps* t*. On définit les mesures de performance suivantes :*

- ***Nombre moyen de clients dans le système*** *:*

$$\mathbb{E}[N] = \lim_{t\to\infty} \mathbb{E}[N_t],$$

  lorsqu'une distribution stationnaire existe.

- ***Temps moyen de séjour d'un client*** *(attente + service) :*

$$\mathbb{E}[T],$$

  où T *désigne le* temps total passé dans le système *par un client avant de sortir.*

- ***Taux d'utilisation des serveurs*** *(fraction de temps pendant laquelle un serveur est occupé) :*

$$U = \frac{\lambda}{c\mu},$$

  dans le modèle $M/M/c$*, où :*
  - λ *est le* taux moyen d'arrivées *(clients par unité de temps) ;*
  - μ *est le* taux moyen de service par serveur *;*
  - c *est le* nombre de serveurs*.*

- ***Taux de refus*** *: probabilité qu'un client arrivant soit bloqué car la capacité maximale* K *est atteinte :*

$$P_{\text{refus}} = \mathbb{P}(N = K).$$

- N reflète la charge du système,
- T reflète le temps d'attente perçu par le client,
- U mesure l'efficacité de l'utilisation des serveurs,
- P_{refus} indique le risque de perte dans les systèmes à capacité finie.

5.3 Métriques et stabilité des files d'attente

5.3.1 Grandeurs de performance classiques

Définition 5.3.1 (Taux d'arrivée et de service)**.** *On note :*

- A_t *: nombre de clients arrivés sur l'intervalle* $[0,t]$*,*
- D_t *: nombre de clients partis (servis) sur l'intervalle* $[0,t]$*,*
- $\lambda = \lim_{t\to\infty} \frac{A_t}{t}$ *le* taux moyen d'arrivée*,*
- $\chi = \lim_{t\to\infty} \frac{D_t}{t}$ *le* taux moyen de sortie*,*
- μ *: paramètre de la loi exponentielle de service (cas* M*), interprété comme le* taux moyen de service d'un serveur*.*

Le paramètre λ représente le débit moyen du flux entrant, et μ celui du flux de sortie d'un seul serveur. Ces taux sont les briques de base du dimensionnement.

Définition 5.3.2 (Taux d'utilisation)**.** *On appelle* taux d'utilisation *(ou* intensité de trafic*) :*

$$\rho = \frac{\lambda}{c\mu},$$

où c *est le nombre de serveurs. C'est la proportion de charge par rapport à la capacité totale de service.*

- Si $\rho < 1$, le système a assez de ressources en moyenne pour absorber les arrivées.
- Si $\rho \geq 1$, la file croît indéfiniment : la file est instable.

ρ est la grandeur clé qui **conditionne la stabilité** d'une file.

Définition 5.3.3. *On définit ici temps de séjour et nombre de clients :*

- *Le* temps de séjour T *est le temps total passé dans le système par un client (attente + service).*
- *Le* nombre de clients dans le système *à l'instant* t *est noté* N_t*.*

Ces grandeurs sont fondamentales du point de vue de l'expérience utilisateur (temps d'attente) et du point de vue opérateur (charge système). On les reliera à λ via la loi de Little.

Définition 5.3.4 (Taux de refus)**.** *Dans un système à capacité finie* K*, un client qui arrive lorsque* $N_t = K$ *est refusé. Le* taux de refus *est donné par :*

$$P_{\text{refus}} = \lim_{t\to\infty} \frac{\text{nombre de clients refusés sur } [0,t]}{A_t}.$$

P_{refus} mesure le risque qu'un client ne soit pas servi (panne réseau, saturation d'un guichet, appels perdus). Il s'agit donc de la **probabilité stationnaire que le système soit plein** ($\mathbb{P}(N = K)$ dans les cas Markoviens).

5.3.2 Stabilité du système

Définition 5.3.5 (Condition de stabilité)**.** *Un système de file d'attente est dit* stable *si le débit de sortie égale le débit d'entrée, i.e.*

$$\chi := \lim_{t\to\infty} \frac{D_t}{t} = \lambda.$$

La condition $\lambda = \chi$ garantit que le nombre moyen de clients ne diverge pas. En cas d'instabilité ($\lambda \geq c\mu$), la file croît sans borne et les métriques stationnaires n'ont plus de sens.

Proposition 5.3.1 (Lien avec le taux d'utilisation)**.** *Pour une file $M/M/c$, la stabilité est équivalente à*

$$\rho = \frac{\lambda}{c\mu} < 1.$$

La fraction ρ compare la charge au service maximal. Ce critère simple est utilisé en pratique pour vérifier si un système peut être dimensionné de manière stationnaire.

5.4 Lois de conservation : Loi de Little et PASTA

5.4.1 Loi de Little

Théorème 5.4.1 (Loi de Little)**.** *Dans tout système de file d'attente stable, on a la relation fondamentale :*

$$\mathbb{E}[N] \;=\; \lambda\,\mathbb{E}[T],$$

où :

- $\mathbb{E}[N]$ *est le nombre moyen de clients présents dans le système (en attente + en service),*
- λ *est le taux d'arrivée moyen,*
- $\mathbb{E}[T]$ *est le temps de séjour moyen d'un client dans le système.*

— Cette loi est universelle : elle ne dépend ni de la loi d'arrivée, ni de la loi de service, ni de la discipline de service.

— Elle exprime simplement que le nombre moyen de clients est égal au débit multiplié par le temps de séjour moyen (principe du "stock = débit × temps de séjour").

— C'est ici une loi de conservation valable pour tout système stable.

Exemple 5.4.1 (Application pratique)**.** *Si une station de service traite $\lambda = 20$ clients/heure, et que le temps moyen de séjour est de 6 minutes ($\mathbb{E}[T] = 0.1$ heure), alors le nombre moyen de clients dans la station est*

$$\mathbb{E}[N] = \lambda\,\mathbb{E}[T] = 20 \times 0.1 = 2.$$

À retenir.

— La loi de Little est un outil de dimensionnement pratique : connaître deux des trois grandeurs ($\mathbb{E}[N], \lambda, \mathbb{E}[T]$) permet d'obtenir la troisième.

— On s'en sert aussi pour vérifier des résultats de calculs : si la relation n'est pas respectée, l'analyse est probablement incorrecte.

5.4.2 Principe PASTA

Principe PASTA : lien entre taux de perte et probabilité stationnaire Le principe PASTA (*Poisson Arrivals See Time Averages*) énonce que, dans une file d'attente à arrivées de Poisson, les statistiques observées *par les arrivées* coïncident avec celles observées *dans le temps*. Autrement dit, un client arrivant à un instant aléatoire "voit" le système dans un état typique du régime stationnaire.

1. **Mise en contexte.** Considérons un système $M/M/c/K$ avec :
   — un processus d'arrivées de Poisson $(A_t)_{t\geq 0}$ de taux λ ;
   — un processus de service exponentiel de paramètre μ ;
   — c serveurs en parallèle et une capacité maximale K (file d'attente de taille $K-c$).

   On note $(N_t)_{t\geq 0}$ le nombre de clients présents dans le système au temps t, et $(\pi_n)_{0\leq n\leq K}$ sa distribution stationnaire.
2. **Taux de perte.** On appelle *taux de perte* le rapport :

$$\delta := \lim_{t\to\infty} \frac{\text{nombre de clients refusés sur } [0,t]}{\text{nombre de clients arrivés sur } [0,t]}.$$

   En notant R_t le nombre de clients refusés jusqu'à t, et A_t le nombre total d'arrivées, cela s'écrit :

$$\delta = \lim_{t\to\infty} \frac{R_t}{A_t}.$$

3. **Principe PASTA.** Le principe PASTA stipule que, pour un processus d'arrivées de Poisson :

$$\mathbb{P}(\text{arrivée voit le système plein}) = \mathbb{P}(\text{le système est plein à un instant quelconque}).$$

   Autrement dit :

$$\delta = \pi_K.$$

Les arrivées de Poisson étant sans mémoire, elles n'introduisent aucun biais dans l'observation : un client n'a pas plus de chances d'arriver dans un état particulier qu'un instant choisi au hasard dans le temps. Ainsi, la vision du système "vue par les arrivées" coïncide exactement avec celle "vue dans le temps", ce qui justifie que le taux de refus observé expérimentalement égale la probabilité théorique d'un système saturé.

À retenir.

— PASTA n'est pas une loi générale mais un résultat spécifique aux arrivées Poisson.
— Il est fondamental pour relier les métriques "vues par les arrivées" (ex. probabilité de refus, probabilité d'attente) aux moyennes stationnaires calculées dans le modèle.
— Dans les TD, on s'en sert pour calculer les taux de perte des systèmes à capacité finie (ex. $M/M/1/K$, $M/M/c/c$).

Chapitre 6

Modèle $M/M/1$

6.1 Définition du modèle $M/M/1$

Le modèle $M/M/1$ décrit une file d'attente avec :

— Arrivées selon un processus de Poisson homogène de taux λ (inter-arrivées i.i.d Exp(λ)),

— Temps de service i.i.d de loi Exp(μ),

— Une seule file d'attente (FIFO) et un seul serveur,

— File supposée de capacité infinie.

Définition 6.1.1 (Modèle $M/M/1$: structure probabiliste)**.** *Le modèle $M/M/1$ décrit une file d'attente à un seul serveur avec :*

— ***Arrivées.*** *Le processus d'arrivées $(A_t)_{t\geq 0}$, où A_t est le nombre total de clients arrivés jusqu'au temps t, est un* processus de Poisson homogène *de taux $\lambda > 0$:*

$$A_t \sim \mathrm{PPH}(\lambda), \qquad A_t = \sum_{n\geq 1} \mathbb{1}_{\{T_n \leq t\}},$$

où $(Q_n)_{n\geq 1}$ sont les interarrivées i.i.d. *$Q_n \sim \mathrm{Exp}(\lambda)$ et*

$$T_n = \sum_{k=1}^{n} Q_k \quad \textit{(instants d'arrivée).}$$

— ***Services.*** *Les durées de service $(S_n)_{n\geq 1}$ sont* i.i.d. *et exponentielles de paramètre $\mu > 0$:*

$$S_n \sim \mathrm{Exp}(\mu), \qquad \mathbb{P}(S_n \leq h) = 1 - e^{-\mu h} = \mu h + o(h) \textit{ pour } h \downarrow 0.$$

La loi exponentielle est sans mémoire *: pour tout $s, t \geq 0$,*

$$\mathbb{P}(S_n > s + t \mid S_n > s) = e^{-\mu t}.$$

— ***Démarrage des services et départs.*** *En $M/M/1$ (FIFO, un serveur), le n-ième client commence son service à*

$$B_n = \max(T_n, D_{n-1}), \qquad D_0 := 0,$$

puis quitte *le système à l'instant*

$$D_n = B_n + S_n \quad \textit{(instants de départ).}$$

Le processus de départs *est alors le processus de comptage*

$$D_t = \sum_{n \geq 1} \mathbb{1}_{\{D_n \leq t\}}.$$

— ***Nombre de clients dans le système.*** *Le processus de file* $(N_t)_{t \geq 0}$ *est*

$$N_t = A_t - D_t,$$

c'est-à-dire le nombre de clients présents (en attente ou en service) à l'instant t*.*

Le processus (N_t) *est un* processus de naissance et mort *à taux constants :*

$$\lambda_n = \lambda, \qquad \mu_n = \begin{cases} 0, & n = 0, \\ \mu, & n \geq 1. \end{cases}$$

C'est le modèle de base en théorie des files d'attente : il combine deux processus exponentiels (arrivées et services) et un serveur unique. Il sert de prototype pour comprendre le rôle de $\rho = \lambda/\mu$, appelé *charge du système.*
Interprétation.

— Les arrivées suivent un **flux de Poisson** : indépendance et stationnarité des incréments.

— Les services suivent un **mécanisme exponentiel** : absence de mémoire du serveur.

— Le processus N_t capture la dynamique complète du système : c'est le cœur du modèle markovien.

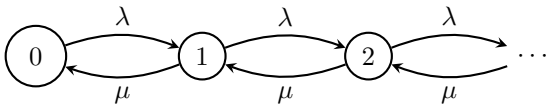
6.2 Représentation par un processus de naissance–mort

On définit $N(t)$ = nombre de clients dans le système (en file + en service) au temps t. Alors $(N(t))_{t \geq 0}$ est une *chaîne de Markov en temps continu* de type naissance–mort avec :

$$\lambda_n = \lambda, \quad \mu_n = \begin{cases} 0, & n = 0, \\ \mu, & n \geq 1, \end{cases}$$

et transitions possibles :

$$n \to n+1 \text{ (arrivée) à taux } \lambda, \qquad n \to n-1 \text{ (départ) à taux } \mu.$$



Intuition. Ce schéma illustre que l'état du système est simplement le nombre de clients. Chaque état n communique avec ses voisins $n-1$ et $n+1$.

6.3 Équations de Chapman–Kolmogorov

Notons $p_n(t) = \mathbb{P}(N(t) = n)$ la probabilité que le système contienne n clients à l'instant t.

Idée générale. Pour établir les équations, on écrit un bilan probabiliste sur un petit intervalle de temps $[t, t+h]$ avec $h \to 0^+$. Durant cet intervalle court :

— une arrivée survient avec probabilité $\lambda h + o(h)$,

— un service se termine (si $n \geq 1$) avec probabilité $\mu h + o(h)$,

— aucun événement ne se produit avec probabilité $1 - (\lambda + \mu)h + o(h)$,

— la probabilité de plusieurs événements simultanés est $o(h)$.

Propriété 6.3.1 (Probabilités infinitésimales sur un court intervalle)**.** *Conditionnellement à $N_t = n$, pour $h \to 0^-$:*

- *une* ***arrivée*** *survient avec probabilité $\lambda h + o(h)$;*
- *un* ***départ*** *survient (si $n \geq 1$) avec probabilité $\mu h + o(h)$, et avec probabilité 0 si $n = 0$;*
- ***aucun événement*** *ne se produit avec probabilité $1 - (\lambda + \mu)h + o(h)$ (pour $n \geq 1$) et $1 - \lambda h + o(h)$ (pour $n = 0$) ;*
- *la probabilité d'****au moins deux événements*** *dans $(t, t + h]$ est $o(h)$ (précisément, en $O(h^2)$).*

Preuve. Soit $(Q_i)_{i\geq 1}$ la suite des temps inter-arrivées, i.i.d. de loi $\text{Exp}(\lambda)$, et $(S_i)_{i\geq 1}$ la suite des durées de service, i.i.d. de loi $\text{Exp}(\mu)$, indépendantes entre elles et du passé. On note A_t le processus de comptage des arrivées et D_t celui des départs.

L'événement "au moins une arrivée dans $(t, t + h]$" s'écrit :

$$\{\text{au moins une arrivée dans } (t, t + h]\} = \{A_{t+h} - A_t \geq 1\}.$$

Or, pour un processus de Poisson de taux λ (cf théorème 4.2.1) :

$$\mathbb{P}(A_{t+h} - A_t \geq 1) = 1 - e^{-\lambda h} \underset{h\to 0}{=} \lambda h + o(h).$$

De même, si $n \geq 1$, le client en service a un temps résiduel exponentiel $\text{Exp}(\mu)$ (par absence de mémoire). Ainsi, l'événement "un départ dans $(t, t + h]$" correspond à :

$$\{\text{un départ dans } (t, t + h]\} = \{S_i \leq h\},$$

où S_i est le temps de service restant du client en cours. On obtient donc :

$$\mathbb{P}\left(\text{un départ dans } (t, t + h] \mid N_t = n\right) = 1 - e^{-\mu h} = \mu h + o(h).$$

Pour $n = 0$, aucun client n'étant en service, cette probabilité vaut 0.

L'événement "aucun changement d'état dans $(t, t + h]$" signifie qu'il n'y a ni arrivée ni départ. Sous indépendance des arrivées et des départs :

$$\begin{aligned}\mathbb{P}(\text{rien dans } (t, t + h] \mid N_t = n) &= \mathbb{P}(\text{aucune arrivée dans } (t, t + h]) \times \mathbb{P}(\text{aucun départ dans } (t, t + h] \mid N_t = n) \\ &= \mathbb{P}(A_{t+h} - A_t = 0) \times \mathbb{P}(S_n > h) \\ &= e^{-\lambda h}\, e^{-\mu h} = e^{-(\lambda+\mu)h} = 1 - (\lambda + \mu)h + o(h),\end{aligned}$$

pour $n \geq 1$. Si $n = 0$, il suffit d'interdire les arrivées : $\mathbb{P}(A_{t+h} - A_t = 0) = e^{-\lambda h} = 1 - \lambda h + o(h)$.

Enfin, la probabilité d'observer *simultanément* une arrivée et un départ pendant $(t, t + h]$ est, par indépendance,

$$\mathbb{P}(\text{arrivée et départ dans } (t, t + h]) \approx (\lambda h)(\mu h) = \lambda\mu h^2 = O(h^2) = o(h).$$

De manière plus générale, pour un processus de Poisson, la probabilité d'avoir au moins deux événements sur un intervalle de longueur h est d'ordre $O(h^2)$:

$$\mathbb{P}(A_{t+h} - A_t \geq 2) = 1 - e^{-\lambda h}(1 + \lambda h) = \frac{\lambda^2 h^2}{2} + o(h^2).$$

Ces contributions sont donc négligeables au premier ordre en h. □

En appliquant la propriété 6.3.1, peut ainsi travailler sur des probabilité linéairisée dans un petit intervalle, ce qui va nous permettre de calculer des taux d'accroissements, et donc d'établir des équations différentielles.

Cas $n = 0$ (file vide). On ne peut qu'avoir :

— un départ n'est pas possible (pas de client),

— une arrivée amène l'état 1 avec probabilité $\lambda h + o(h)$,

— ou rien ne se passe.

Donc, en écrivant l'évolution :

$$p_0(t+h) = p_0(t)\big(1 - \lambda h\big) + p_1(t)\mu h + o(h).$$

En soustrayant $p_0(t)$ et en divisant par h, puis en passant à la limite $h \to 0$, on obtient :

$$\frac{\mathrm{d}}{\mathrm{d}t} p_0(t) = \mu p_1(t) - \lambda p_0(t).$$

Cas $n \geq 1$ (file vide). On considère un processus de naissance–mort $(N_t)_{t\geq 0}$ décrivant une file d'attente $M/M/1$. On note :

$$p_n(t) := \mathbb{P}(N_t = n), \qquad n \geq 0.$$

Les transitions possibles pendant un intervalle infinitésimal $[t, t+h]$ sont :

$$\begin{cases} N_{t+h} = N_t + 1 & \text{(une arrivée, au taux } \lambda\text{)}, \\ N_{t+h} = N_t - 1 & \text{(un départ, au taux } \mu\text{)}, \\ N_{t+h} = N_t & \text{(aucun événement)}, \\ N_{t+h} - N_t \in \{-2, +2, \dots\} & \text{(événements multiples, négligeables en } o(h)\text{)}. \end{cases}$$

Par **formule des probabilités totales**, pour $n \geq 1$:

$$\begin{aligned} p_n(t+h) &= \mathbb{P}(N_{t+h} = n) \\ &\underset{h\to 0}{=} \mathbb{P}(N_{t+h} - N_t = 1 \mid N_t = n-1)\,\mathbb{P}(N_t = n-1) \\ &\quad + \mathbb{P}(N_{t+h} - N_t = -1 \mid N_t = n+1)\,\mathbb{P}(N_t = n+1) \\ &\quad + \mathbb{P}(N_{t+h} - N_t = 0 \mid N_t = n)\,\mathbb{P}(N_t = n) + o(h). \end{aligned}$$

Justification de chaque terme :

— $\mathbb{P}(N_{t+h} - N_t = 1 \mid N_t = n-1)$ correspond à la probabilité qu'une *arrivée* survienne dans $[t, t+h]$, lorsque le système contient $n-1$ clients à l'instant t. Comme les arrivées suivent un processus de Poisson de taux λ,

$$\mathbb{P}(N_{t+h} - N_t = 1 \mid N_t = n-1) \underset{h\to 0}{=} \lambda h + o(h).$$

— $\mathbb{P}(N_{t+h} - N_t = -1 \mid N_t = n+1)$ correspond à la probabilité qu'un *départ* ait lieu pendant $[t, t+h]$, sachant qu'il y a $n+1$ clients à l'instant t. Comme les services sont exponentiels de paramètre μ, la probabilité qu'un client termine son service est :

$$\mathbb{P}(N_{t+h} - N_t = -1 \mid N_t = n+1) \underset{h\to 0}{=} \mu h + o(h).$$

— $\mathbb{P}(N_{t+h} - N_t = 0 \mid N_t = n)$ représente le cas où *aucun événement* (ni arrivée, ni départ) ne se produit pendant $[t, t+h]$. Comme les deux événements sont indépendants et rares (d'ordre h), la probabilité de ne rien avoir est :

$$\mathbb{P}(N_{t+h} - N_t = 0 \mid N_t = n) \underset{h\to 0}{=} 1 - (\lambda + \mu)h + o(h).$$

Justification de la négligence des autres transitions. Les événements tels que "deux arrivées simultanées" ou "un départ et une arrivée dans le même intervalle" ont des probabilités d'ordre $o(h)$ (en effet, pour un processus de Poisson, la probabilité d'au moins deux événements dans un intervalle de longueur h est $O(h^2)$). Ils sont donc négligés à l'ordre premier.

En regroupant les contributions :

$$\begin{aligned} p_n(t+h) &\underset{h\to 0}{=} p_{n-1}(t)(\lambda h + o(h)) + p_{n+1}(t)(\mu h + o(h)) + p_n(t)(1-(\lambda+\mu)h + o(h)) \\ &\underset{h\to 0}{=} p_n(t) + h\big[\lambda p_{n-1}(t) + \mu p_{n+1}(t) - (\lambda+\mu)p_n(t)\big] + o(h). \end{aligned}$$

En soustrayant $p_n(t)$ et en divisant par h, on obtient à la limite $h \to 0$:

$$\frac{\mathrm{d}}{\mathrm{d}t}p_n(t) = \lambda p_{n-1}(t) + \mu p_{n+1}(t) - (\lambda+\mu)p_n(t), \quad n \geq 1.$$

Synthèse. Les équations de Chapman–Kolmogorov du modèle $M/M/1$ s'écrivent donc :

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}p_0(t) = \mu p_1(t) - \lambda p_0(t), \\ \frac{\mathrm{d}}{\mathrm{d}t}p_n(t) = \lambda p_{n-1}(t) + \mu p_{n+1}(t) - (\lambda+\mu)p_n(t), \quad n \geq 1. \end{cases} \tag{6.1}$$

Intuition. Ces équations traduisent l'égalité fondamentale :

$$\text{Variation instantanée de } p_n(t) = \text{Flux entrants} - \text{Flux sortants}.$$

Chaque état n échange du flux probabiliste uniquement avec ses deux voisins $(n-1,\ n+1)$. Ce raisonnement se généralise à tous les processus de naissance–mort.
Intuition. Ces équations expriment un bilan de flux probabilistes : la dérivée de $p_n(t)$ correspond à la différence entre le flux entrant (provenant des états voisins) et le flux sortant (quittant l'état n).

6.4 Distribution stationnaire

En régime stationnaire (considérer $\frac{\mathrm{d}}{\mathrm{d}t}p_n = 0$ pour (6.1)), en notant conventionnellement les probabilités stationnaire $(\pi_n)_n$, les équations d'équilibre global deviennent :

$$\pi_{n+1}\mu = \pi_n\lambda, \qquad n \geq 0.$$

Donc, par récurrence :

$$\pi_n = \pi_0\rho^n, \qquad \rho = \frac{\lambda}{\mu}.$$

La normalisation $\sum_{n=0}^{\infty} \pi_n = 1$ impose que la série géométrique converge $\iff \rho < 1$. On obtient alors :

$$\pi_n = (1-\rho)\,\rho^n, \qquad n \geq 0.$$

La loi stationnaire suit donc une **loi géométrique de paramètre** $1-\rho$ sur $\mathbb{N}$. La condition de stabilité $\rho < 1$ signifie que la charge du système doit être inférieure à 100%. Sinon, la file diverge et aucune distribution stationnaire n'existe.

6.5 Performances moyennes sous discipline FIFO

— **Nombre moyen de clients dans le système :**

$$\mathbb{E}[N] = \sum_{n=0}^{\infty} n\pi_n = \frac{\rho}{1-\rho}.$$

— **Temps moyen de séjour :** Par la loi de Little :

$$\mathbb{E}[T] = \frac{\mathbb{E}[N]}{\lambda} = \frac{1}{\mu-\lambda}.$$

— **Temps d'attente moyen dans la file :** Temps de séjour − temps de service moyen :

$$\mathbb{E}[W] = \mathbb{E}[T] - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}.$$

— **Taux d'occupation du serveur :**

$$U = \rho = \frac{\lambda}{\mu}.$$

Preuve : Calcul de l'espérance. Comme (π_n) est une loi géométrique,

$$\mathbb{E}[N] = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = (1-\rho)\,\frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho}.$$

□

Intuition.

— La formule $\mathbb{E}[T] = 1/(\mu - \lambda)$ est à retenir absolument. Elle montre l'explosion du temps de séjour quand λ approche μ.

— Le taux d'occupation $U = \rho$ traduit la fraction de temps où le serveur est occupé.

— Ce sont des formules de référence à comparer avec d'autres files ($M/M/c$, $M/M/c/c$, etc.).

Exemple 6.5.1 (Calcul numérique)**.** *Supposons $\lambda = 2$ clients/minute et $\mu = 4$ clients/minute. Alors $\rho = \lambda/\mu = 1/2$,*

$$\mathbb{E}[N] = \frac{\rho}{1-\rho} = \frac{1/2}{1-1/2} = 1,$$

$$\mathbb{E}[T] = \frac{1}{\mu - \lambda} = \frac{1}{4-2} = \tfrac{1}{2}\ min = 30\ s,$$

$$\mathbb{E}[W] = \frac{\rho}{\mu - \lambda} = \frac{1/2}{2} = \tfrac{1}{4}\ min = 15\ s.$$

En moyenne, ***1 client*** *est présent dans le système, chaque client reste* ***30 secondes****, dont environ* ***15 secondes*** *en file.*

6.6 Synthèse sur les files $M/M/1$

Synthèse des propriétés fondamentales du modèle $M/M/1$

— **Arrivées.** Les temps inter-arrivées Q_i sont i.i.d. de loi exponentielle :

$$Q_i \sim \mathrm{Exp}(\lambda),$$

et le processus d'arrivées $(A_t)_{t\geq 0}$ est un **processus de Poisson homogène** :

$$A_t \sim \mathrm{PPH}(\lambda).$$

— **Services.** Les durées de service S_i sont i.i.d. exponentielles :

$$S_i \sim \mathrm{Exp}(\mu),$$

et la discipline est FIFO avec un unique serveur.

— **Taux d'occupation.** Le serveur est occupé une proportion moyenne de temps :

$$U = \rho = \frac{\lambda}{\mu}.$$

La condition de stabilité du système est $\rho < 1$.

— **Équations de Chapman–Kolmogorov.** Pour $n \geq 1$:

$$\frac{\mathrm{d}}{\mathrm{d}t} p_n(t) = \lambda p_{n-1}(t) + \mu p_{n+1}(t) - (\lambda + \mu) p_n(t),$$

et pour $n = 0$:

$$\frac{\mathrm{d}}{\mathrm{d}t} p_0(t) = \mu p_1(t) - \lambda p_0(t).$$

— **Distribution stationnaire.** En régime permanent, les probabilités d'état vérifient :

$$\pi_{n+1}\mu = \pi_n \lambda, \qquad n \geq 0,$$

donc

$$\pi_n = (1-\rho)\rho^n, \qquad \text{loi géométrique sur } \mathbb{N}.$$

— **Valeur moyenne du nombre de clients.**

$$\mathbb{E}[N] = \frac{\rho}{1-\rho}.$$

Cette quantité mesure le *nombre moyen de clients* dans le système (en attente + en service).

— **Loi de Little.** En régime stationnaire :

$$\mathbb{E}[N] = \lambda\, \mathbb{E}[T],$$

d'où :

$$\mathbb{E}[T] = \frac{1}{\mu - \lambda}.$$

— **Temps d'attente moyen dans la file.** En soustrayant le temps de service moyen :

$$\mathbb{E}[W] = \mathbb{E}[T] - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}.$$

Chapitre 7

Extension du modèle de file d'attente

7.1 Modèle $M/M/c$

7.1.1 Structure probabiliste

Le modèle $M/M/c$ généralise le modèle $M/M/1$ à c **serveurs identiques** travaillant en parallèle. Il est utilisé pour modéliser des systèmes multi-serveurs tels que :

— un centre d'appels avec plusieurs opérateurs ;
— un hôpital avec plusieurs médecins disponibles ;
— un serveur informatique avec plusieurs processeurs indépendants.

L'objectif est d'évaluer les performances du système en fonction du nombre de serveurs c, notamment les temps d'attente, la probabilité qu'un client attende, et le taux d'occupation global.

Définition 7.1.1 (Modèle $M/M/c$)**.** *Le modèle $M/M/c$ repose sur les hypothèses suivantes :*

— ***Arrivées.*** *Les clients arrivent selon un processus de Poisson homogène de taux $\lambda > 0$:*

$$A_t = \sum_{n \geq 1} \mathbb{1}_{\{T_n \leq t\}}, \qquad Q_n = T_n - T_{n-1} \sim \mathrm{Exp}(\lambda).$$

Les arrivées sont indépendantes et stationnaires dans le temps.

— ***Services.*** *Chaque serveur offre un service de durée exponentielle indépendante :*

$$S_i \sim \mathrm{Exp}(\mu),$$

et il y a c serveurs fonctionnant en parallèle. Si plus de c clients sont présents, les excédents attendent en file (FIFO).

— ***Processus de file.*** *Le nombre total de clients dans le système (en service + en attente) à l'instant t est noté :*

$$N_t = A_t - D_t.$$

Le processus $(N_t)_{t \geq 0}$ est un ***processus de naissance et mort*** *dont les taux de transition sont :*

$$\lambda_n = \lambda, \qquad \mu_n = \min(n, c)\, \mu.$$

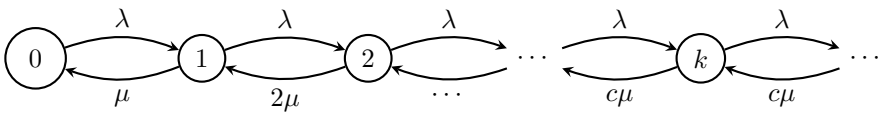
Ainsi :

— tant qu'il y a moins de c clients, tous sont servis et le taux global de service croît linéairement avec n ;
— dès que $n \geq c$, tous les serveurs sont occupés, le taux de service total est alors constant et égal à $c\mu$.

7.1.2 Représentation par un processus de naissance–mort

On note N_t le nombre total de clients dans le système (en file + en service). Le processus $(N_t)_{t\geq 0}$ est un **processus de naissance–mort** de paramètres :

$$\forall n \geq 0, \quad \lambda_n = \lambda,$$
$$\mu_n = \begin{cases} n\mu, & 1 \leq n \leq c, \\ c\mu, & n \geq c. \end{cases}$$



Intuition. Tant que $n < c$, chaque client en service contribue à un taux de départ μ, d'où $\mu_n = n\mu$. Dès que $n \geq c$, tous les serveurs sont occupés : le taux global de sortie devient constant, $\mu_n = c\mu$.

7.1.3 Équations de Chapman–Kolmogorov

Comme $(N_t)_{t\geq 0}$ est une chaîne de naissance–mort, les équations différentielles d'évolution des probabilités d'état sont :

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} p_0(t) = \mu_1 p_1(t) - \lambda p_0(t), \\ \dfrac{\mathrm{d}}{\mathrm{d}t} p_n(t) = \lambda p_{n-1}(t) + \mu_{n+1} p_{n+1}(t) - (\lambda + \mu_n) p_n(t), \quad n \geq 1. \end{cases}$$

Explication. Ces équations s'obtiennent par le même raisonnement que pour le modèle $M/M/1$: sur un intervalle infinitésimal $[t, t+h]$, on écrit le bilan des flux de probabilité associés :

- **naissance** $n \to n+1$ avec taux λ ;
- **mort** $n \to n-1$ avec taux μ_n ;
- aucun événement avec probabilité $1 - (\lambda + \mu_n)h + o(h)$.

Le processus est donc toujours de type **Markovien**, mais les taux de service dépendent désormais du nombre de serveurs actifs.

7.1.4 Distribution stationnaire

En régime permanent, le processus (N_t) admet une distribution stationnaire $(\pi_n)_{n\geq 0}$ dès que la condition de stabilité suivante est vérifiée :

$$\rho = \frac{\lambda}{c\mu} < 1.$$

Les équations d'équilibre global s'écrivent :

$$\lambda\, \pi_n = \mu_{n+1}\, \pi_{n+1}, \qquad n \geq 0,$$

soit, en remplaçant les valeurs de μ_n :

$$\pi_{n+1} = \begin{cases} \dfrac{\lambda}{(n+1)\mu}\, \pi_n, & n < c, \\ \dfrac{\lambda}{c\mu}\, \pi_n, & n \geq c. \end{cases}$$

Par récurrence, on obtient :

$$\pi_n = \begin{cases} \dfrac{1}{n!} \left(\dfrac{\lambda}{\mu}\right)^n \pi_0, & 0 \leq n \leq c, \\ \dfrac{1}{c!} \left(\dfrac{\lambda}{\mu}\right)^c \rho^{\,n-c} \pi_0, & n \geq c, \end{cases}$$

où $\rho = \dfrac{\lambda}{c\mu}$.
La constante de normalisation π_0 est déterminée par la condition $\sum_{n=0}^{\infty} \pi_n = 1$, soit :

$$\pi_0 = \left[\sum_{n=0}^{c-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \frac{1}{c!} \left(\frac{\lambda}{\mu}\right)^c \frac{1}{1-\rho} \right]^{-1}.$$

Cette distribution stationnaire est appelée **loi de file d'attente d'Erlang-C**. Elle décrit la probabilité d'avoir n clients dans le système à long terme, pour un système $M/M/c$ de capacité infinie.

7.1.5 Performances moyennes sous discipline FIFO

À partir de la loi stationnaire, on déduit plusieurs indicateurs de performance :

— **Probabilité qu'un client attende (tous serveurs occupés) :**

$$P_{\text{attente}} = \frac{1}{c!} \left(\frac{\lambda}{\mu}\right)^c \frac{1}{1-\rho}\, \pi_0.$$

— **Nombre moyen de clients dans la file d'attente :**

$$\mathbb{E}[N_q] = P_{\text{attente}} \cdot \frac{\rho}{1-\rho}.$$

— **Nombre moyen total de clients dans le système :**

$$\mathbb{E}[N] = \mathbb{E}[N_q] + \frac{\lambda}{\mu}.$$

— **Temps moyen d'attente dans la file (loi de Little) :**

$$\mathbb{E}[W] = \frac{\mathbb{E}[N_q]}{\lambda}.$$

— **Temps moyen total dans le système :**

$$\mathbb{E}[T] = \mathbb{E}[W] + \frac{1}{\mu}.$$

Commentaires.

— La formule de P_{attente} correspond à la **formule d'Erlang-C**. Elle donne la probabilité qu'un client doive patienter avant d'être servi.

— Lorsque $c = 1$, on retrouve exactement les résultats du modèle $M/M/1$:

$$\pi_n = (1-\rho)\rho^n, \quad \mathbb{E}[N] = \frac{\rho}{1-\rho}, \quad \mathbb{E}[T] = \frac{1}{\mu - \lambda}.$$

7.2 Modèle $M/M/c/K$

Le modèle $M/M/c/K$ généralise encore les modèles précédents en introduisant une **capacité finie** du système. Autrement dit, la file ne peut contenir plus de K clients (en service et en attente). Lorsqu'un client arrive alors que le système est plein ($N_t = K$), il est **perdu** ou **refusé** : on parle alors de *modèle avec pertes*.

7.2.1 Structure probabiliste

Définition 7.2.1 (Modèle $M/M/c/K$)**.** *Le système $M/M/c/K$ est défini par les hypothèses suivantes :*

— ***Arrivées.*** *Les arrivées forment un processus de Poisson homogène de taux $\lambda > 0$. Si le système contient K clients, les nouvelles arrivées sont refusées :*

$$\lambda_n = \begin{cases} \lambda, & 0 \le n < K, \\ 0, & n = K. \end{cases}$$

Autrement dit, le taux de naissance s'annule dès que la capacité maximale est atteinte.

— ***Services.*** *Les durées de service sont i.i.d. exponentielles de paramètre $\mu > 0$, et il y a c serveurs en parallèle. Le taux global de départ dépend donc du nombre de clients présents :*

$$\mu_n = \begin{cases} n\mu, & 1 \le n \le c, \\ c\mu, & c < n \le K. \end{cases}$$

Ainsi, lorsque moins de c clients sont présents, tous sont servis simultanément ; au-delà, les c serveurs sont saturés.

— ***Processus de file.*** *Le processus $(N_t)_{t\ge 0}$, nombre de clients dans le système à l'instant t, est un **processus de naissance–mort** à espace d'états fini :*

$$\mathcal{S} = \{0, 1, 2, \dots, K\}.$$

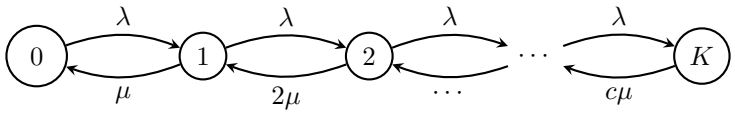
Les transitions possibles sont :

$$n \xrightarrow[\text{taux } \lambda]{\text{arrivée}} n+1 \quad (\text{si } n < K), \qquad n \xrightarrow[\text{taux } \mu_n]{\text{départ}} n-1 \quad (\text{si } n > 0).$$

7.2.2 Processus de naissance-mort associé

On définit N_t comme étant le nombre de clients en service à l'instant t. Les taux de transition sont :

$$\lambda_n = \begin{cases} \lambda, & 0 \le n < K, \\ 0, & n = K \quad (\text{arrivée bloquée}), \end{cases} \qquad \mu_n = \min(n, c)\mu, \quad 1 \le n \le K.$$



La chaîne de Markov est toujours une naissance–mort, mais la différence avec $M/M/c$ est que le taux d'arrivée devient nul en K : **aucune arrivée ne peut entrer** si la capacité maximale est atteinte.

7.2.3 Distribution stationnaire

En régime permanent, le processus (N_t) admet une distribution stationnaire $(\pi_n)_{0\leq n\leq K}$ solution des équations d'équilibre global :

$$\lambda_{n-1}\pi_{n-1} = \mu_n \pi_n, \qquad 1 \leq n \leq K.$$

On en déduit par récurrence :

$$\pi_n = \begin{cases} \dfrac{1}{n!}\left(\dfrac{\lambda}{\mu}\right)^n \pi_0, & 0 \leq n \leq c, \\ \dfrac{1}{c!}\left(\dfrac{\lambda}{\mu}\right)^c \rho^{\,n-c}\pi_0, & c < n \leq K, \end{cases}$$

où $\rho = \dfrac{\lambda}{c\mu}$.
La constante de normalisation π_0 est déterminée par :

$$\pi_0 = \left[\sum_{n=0}^{c-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \frac{1-\rho^{\,K-c+1}}{1-\rho}\right]^{-1}.$$

Cette distribution stationnaire est la **loi d'Erlang-B généralisée** (ou loi de file à capacité finie). Elle décrit la probabilité d'occuper chaque état n, y compris la saturation du système ($n = K$).

7.2.4 Probabilité de perte et performances moyennes

La **probabilité de perte**, notée P_{perte}, correspond à la probabilité que le système soit saturé :

$$P_{\text{perte}} = \pi_K.$$

Un client arrivant alors que $N_t = K$ est immédiatement rejeté.

On définit le **taux effectif d'arrivée** (ou *flux admis*) :

$$\lambda_{\text{eff}} = \lambda(1 - P_{\text{perte}}).$$

Il s'agit du taux moyen de clients effectivement acceptés dans le système.

Nombre moyen de clients dans le système :

$$\mathbb{E}[N] = \sum_{n=0}^{K} n\,\pi_n.$$

Temps moyen passé dans le système : (en appliquant la loi de Little au flux accepté)

$$\mathbb{E}[T] = \frac{\mathbb{E}[N]}{\lambda_{\text{eff}}}.$$

Temps moyen d'attente dans la file :

$$\mathbb{E}[W] = \mathbb{E}[T] - \frac{1}{\mu}.$$

— Le modèle $M/M/c/K$ permet de représenter des files à capacité limitée : parkings, systèmes de communication ou d'appels saturés.

— Lorsque $K \to \infty$, on retrouve le modèle $M/M/c$ classique.

— Lorsque $K = c$, aucune attente n'est possible : on obtient le modèle à pertes $M/M/c/c$ (ou modèle d'Erlang-B).

Chapitre 8

Exercices de synthèse

8.1 Exercice formel

8.1.1 Énoncé

Exercice 8.1.1 (File d'attente $M/M/2/3$ – Exercice de synthèse). On considère une file d'attente de type $M/M/2/3$, modélisée comme un processus de naissance et mort $(N_t)_{t\geq 0}$ représentant le nombre total de clients dans le système (en service ou en attente) à l'instant t.

(1) Signification du modèle. Expliquer le sens de la notation $M/M/2/3$. Que signifient les lettres et les chiffres ? Quelle est la politique (ou discipline) de service ici ?

(2) Processus de comptage et dynamique. On rappelle la représentation générique d'un processus de comptage :

$$N_t = \sum_{n\geq 0} \mathbb{1}_{\{T_n \leq t\}},$$

où (T_n) désignent les temps d'arrivée successifs. Dans le modèle $M/M/2/3$:

(i) Quelle loi suit le processus des inter-arrivées Q_n ? Quel est leur lien avec (T_n) ?

(ii) Quelle loi suit le processus de services S_n ?

(iii) Donner les approximations infinitésimales des probabilités d'événements élémentaires jusqu'à l'ordre 1 en $o(h)$ pour $(N_t)_{t\geq 0}$.

(3) Espace des états. Déterminer l'espace d'états $\mathcal{S}$ de $(N_t)_{t\geq 0}$ et interpréter chaque état.

(4) Équations de Chapman–Kolmogorov. En notant $p_n(t) = \mathbb{P}(N_t = n)$, établir les équations de Chapman–Kolmogorov avant :

$$\frac{\mathrm{d}}{\mathrm{d}t} p_n(t) = \lambda_{n-1} p_{n-1}(t) + \mu_{n+1} p_{n+1}(t) - (\lambda_n + \mu_n) p_n(t),$$

avec les conditions aux bords appropriées.

(5) Relation entre les lois temporelles et la matrice de transition. On note :

$$p(t) = \big(p_0(t),\, p_1(t),\, p_2(t),\, p_3(t)\big) = \big(\mathbb{P}(N_t = 0),\, \mathbb{P}(N_t = 1),\, \mathbb{P}(N_t = 2),\, \mathbb{P}(N_t = 3)\big)$$

le vecteur ligne décrivant la loi de (N_t) à l'instant t, et $p(0)$ la loi initiale. On note également $P(t)$ la *matrice de transition* du processus, telle que :

$$P(t) = \big(p_{ij}(t)\big)_{i,j}, \qquad p_{ij}(t) = \mathbb{P}(N_t = j \mid N_0 = i).$$

(i) Exprimer la relation entre $p(t)$, $p(0)$ et $P(t)$.

(ii) Rappeler l'équation différentielle reliant $P(t)$ à la matrice infinitésimale Λ.

(iii) Donner l'expression formelle de $p(t)$ en fonction de Λ.

(iv) Justifier que la matrice infinitésimale Λ est la dérivée en $t = 0$ de la matrice des transitions $P(t)$ et interpréter cette relation.

(6) Matrice infinitésimale. Écrire la matrice infinitésimale Λ explicitement, puis interpréter :

— le signe de ses coefficients diagonaux ;

— la signification physique des coefficients hors diagonale ;

— pourquoi chaque ligne de Λ somme à 0.

(7) Loi stationnaire et stabilité. Définir la loi stationnaire π du processus (N_t). Énoncer les équations qu'elle satisfait. Discuter la notion de **stabilité** : que signifie $\rho = \frac{\lambda}{c\mu} < 1$ dans le cas $M/M/c$ (ρ désigne le facteur de charge) ? Est-elle pertinente ici (capacité finie $K = 3$) ?

(8) Résolution de la loi stationnaire. Déterminer $\pi = (\pi_0, \pi_1, \pi_2, \pi_3)$ explicitement, en justifiant les récurrences :

$$\pi_n \mu_n = \pi_{n-1} \lambda_{n-1}.$$

Donner la valeur normalisée de π_0, puis exprimer toutes les autres composantes.

(9) Taux de refus et métriques de performance. Calculer :

— la **probabilité de refus** $P_{\text{refus}} = \pi_3$;

— le **taux effectif d'entrée** $\lambda_{\text{eff}} = \lambda(1 - \pi_3)$;

— la **charge du système** $U = \frac{\lambda_{\text{eff}}}{c\mu}$.

Interpréter ces résultats : que se passe-t-il si $\lambda \gg \mu$?

(10) Loi de Little et temps moyen de séjour. Rappeler la *loi de Little*, en notant $\mathbb{E}[N]$ le nombre moyen de client dans le système stationnaire. En déduire le temps moyen de séjour $\mathbb{E}[T]$ dans le système $M/M/2/3$, puis préciser les valeurs de $L = \mathbb{E}[N]$, L_q, $W = \mathbb{E}[T]$ et W_q. Donner la relation entre ces quantités et expliquer leur signification pratique.

(11) Interprétation globale. Discuter l'intérêt du modèle $M/M/2/3$ dans un contexte réel (centre d'appel, guichet, service hospitalier). Quelles grandeurs peuvent être ajustées pour optimiser les performances ?

8.1.2 Solution

(1) Signification du modèle

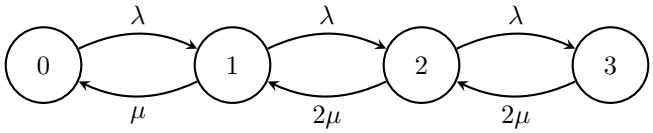
Le modèle **M/M/2/3** se lit selon la **notation de Kendall** :

$$A/B/C/K,$$

où :

- $A = M$ signifie que les arrivées suivent un **processus de Poisson** de paramètre λ, c'est-à-dire que les temps inter-arrivées sont **exponentiels de moyenne** $1/\lambda$;
- $B = M$ indique que les temps de service sont également **exponentiels**, de paramètre μ ;
- $C = 2$ correspond au nombre de **serveurs en parallèle** ;
- $K = 3$ est la **capacité totale du système**, incluant les clients en service et en attente.

Ainsi, le système peut contenir au maximum trois clients : deux en service et un en attente. La **discipline de service** est **FIFO** (First In, First Out), c'est-à-dire que les clients sont servis dans l'ordre d'arrivée.



(2) Processus de comptage et dynamique

On rappelle qu'un processus de comptage $(N_t)_{t\geq 0}$ peut se représenter comme :

$$N_t = \sum_{n\geq 0} \mathbb{1}_{\{T_n \leq t\}},$$

où $(T_n)_{n\geq 0}$ sont définis par $T_n = \sum_{k=1}^{n} Q_k$ désignant la somme des temps inter-arrivées.

(i) **Processus des inter-arrivées.** Dans un système $M/M/2/3$, les arrivées suivent un *processus de Poisson homogène* de taux λ. Par définition, les *inter-arrivées* $Q_n = T_n - T_{n-1}$ sont alors des variables aléatoires **exponentielles de paramètre** λ, i.i.d. On a donc :

$$Q_n \sim \text{Exp}(\lambda), \qquad T_n = \sum_{k=1}^{n} Q_k \sim \text{Gamma}(n, \lambda).$$

Ce lien exprime que le processus (T_n) est le cumul des temps inter-arrivées.

(ii) **Processus de services.** Les durées de service (S_n) sont supposées **i.i.d. exponentielles de paramètre** μ où μ représente le nombre moyen de départs par unité de temps pour un client en service. Chaque serveur fonctionne indépendamment et fournit des temps de service $S_n \sim \text{Exp}(\mu)$. Dans un modèle $M/M/2/3$, jusqu'à deux services peuvent être exécutés simultanément.

(iii) **Approximation infinitésimale des transitions.** Pour un petit intervalle de temps $[t, t+h]$ avec $h \to 0^+$, les probabilités élémentaires de transitions du processus (N_t) s'écrivent, à l'ordre $o(h)$:

$$\begin{cases} \mathbb{P}(N_{t+h} = N_t + 1) = \lambda h + o(h), & \text{(une arrivée)}, \\ \mathbb{P}(N_{t+h} = N_t - 1) = \mu h + o(h), & \text{(un service se termine, si } N_t \geq 1), \\ \mathbb{P}(N_{t+h} = N_t) = 1 - (\lambda + \mu)h + o(h), & \text{(aucun événement)}, \\ \mathbb{P}(\text{plus d'un événement sur } [t, t+h]) = o(h). \end{cases}$$

Ces relations traduisent la *dynamique de naissance–mort* caractéristique du processus $M/M/2/3$, où les arrivées se produisent au taux λ et les départs au taux μ par serveur actif.

(3) Espace des états

L'espace des états du processus (N_t) est :

$$S = \{0, 1, 2, 3\}.$$

— $N_t = 0$: aucun client dans le système ;
— $N_t = 1$: un client en service, un serveur occupé ;
— $N_t = 2$: deux clients en service, serveurs saturés ;
— $N_t = 3$: système plein, un client en attente et deux en service.

(4) Équations de Chapman–Kolmogorov avant

Notons $p_n(t) = \mathbb{P}(N_t = n)$ la probabilité que le système contienne n clients à l'instant t.

Idée générale. On établit les équations d'évolution en effectuant un *bilan probabiliste* sur un petit intervalle $[t, t+h]$ avec $h \to 0^+$. Durant cet intervalle court :

— une **arrivée** survient avec probabilité $\lambda h + o(h)$ (si le système n'est pas plein) ;
— un **départ** (fin de service) survient avec probabilité $\mu_n h + o(h)$;
— aucun événement ne se produit avec probabilité $1 - (\lambda_n + \mu_n)h + o(h)$.

Ici, les taux dépendent de l'état :

$$\lambda_n = \begin{cases} \lambda, & n < 3, \\ 0, & n = 3, \end{cases} \qquad \mu_n = \begin{cases} 0, & n = 0, \\ \min(n, 2)\mu, & n \geq 1. \end{cases}$$

On néglige les événements multiples (probabilité $o(h)$). En appliquant la **formule des probabilités totales** :

$$p_n(t+h) = p_{n-1}(t)(\lambda_{n-1}h + o(h)) + p_{n+1}(t)(\mu_{n+1}h + o(h)) + p_n(t)(1 - (\lambda_n + \mu_n)h + o(h)).$$

En soustrayant $p_n(t)$, divisant par h et en passant à la limite $h \to 0$, on obtient l'équation différentielle de Chapman-Kolmogorov :

$$\frac{\mathrm{d}}{\mathrm{d}t} p_n(t) = \lambda_{n-1} p_{n-1}(t) + \mu_{n+1} p_{n+1}(t) - (\lambda_n + \mu_n) p_n(t), \quad n \in \{0, 1, 2, 3\}.$$

Les conditions aux bords s'écrivent naturellement :

$$\lambda_{-1} = 0, \quad \mu_0 = 0, \quad \lambda_3 = 0.$$

(5) Relation entre lois temporelles et matrice de transition

On note :

$$p(t) = \big(p_0(t), p_1(t), p_2(t), p_3(t)\big).$$

(i) La relation entre la loi à l'instant t et la loi initiale est :

$$p(t) = p(0)P(t),$$

où $P(t) = (p_{ij}(t))_{i,j}$ est la matrice de transition.

(ii) L'évolution de $P(t)$ est régie par :

$$P'(t) = P(t)\Lambda = \Lambda P(t).$$

(iii) La solution formelle est :

$$P(t) = e^{t\Lambda} \quad \text{et donc} \quad p(t) = p(0)e^{t\Lambda}.$$

(iv) Relation entre Λ et $P(t)$.
La matrice des transitions $P(t) = (p_{ij}(t))_{i,j}$ décrit la dynamique du processus :

$$p_{ij}(t) = \mathbb{P}(N_t = j \mid N_0 = i).$$

Pour une chaîne de Markov à temps continu, cette matrice satisfait les **équations de Chapman-Kolmogorov** :

$$\forall\, t, h \geq 0,\ P(t+h) = P(t)P(h),$$

avec la condition initiale $P(0) = I$ (matrice identité).

En dérivant cette relation par rapport à t en $t = 0$, on obtient le **générateur infinitésimal** Λ :

$$\Lambda = \lim_{h \to 0^+} \frac{P(h) - I_d}{h}.$$

Chaque coefficient de Λ possède alors une interprétation probabiliste :

$$\Lambda_{ij} = \lim_{t \to 0^+} \frac{P(t) - \mathbb{1}_{\{i=j\}}}{t} = \begin{cases} \lim\limits_{h \to 0^+} \dfrac{p_{ij}(h)}{h}, & i \neq j, \\ -\sum\limits_{k \neq i} \Lambda_{ik}, & i = j. \end{cases}$$

Ainsi :

— pour $i \neq j$, Λ_{ij} est le **taux instantané de transition** de l'état i vers l'état j ;
— pour $i = j$, la valeur négative Λ_{ii} représente le **taux total de sortie** de l'état i.

Réciproquement, connaissant Λ, la matrice des transitions à l'instant t s'obtient par la **résolution du système matriciel différentiel** :

$$\frac{\mathrm{d}}{\mathrm{d}t} P(t) = P(t)\Lambda, \qquad P(0) = I.$$

La solution formelle s'écrit alors :

$$P(t) = e^{t\Lambda} = I + t\Lambda + \frac{t^2}{2!}\Lambda^2 + \cdots$$

Cette expression montre que la matrice infinitésimale Λ gouverne entièrement la dynamique du processus : elle joue le rôle de **générateur du semi-groupe** $(P(t))_{t \geq 0}$.

Vérification : À partir de $P(t)$, on retrouve bien Λ en dérivant en $t = 0$:

$$\Lambda = P'(0) = \left.\frac{\mathrm{d}}{\mathrm{d}t} P(t)\right|_{t=0}.$$

En pratique, pour une file M/M/2/3, les premières dérivées des coefficients $p_{ij}(t)$ en $t = 0$ donnent exactement les taux de transitions : par exemple,

$$\left.\frac{\mathrm{d}}{\mathrm{d}t} p_{01}(t)\right|_{t=0} = \lambda, \quad \left.\frac{\mathrm{d}}{\mathrm{d}t} p_{10}(t)\right|_{t=0} = \mu, \quad \left.\frac{\mathrm{d}}{\mathrm{d}t} p_{22}(t)\right|_{t=0} = -(\lambda + 2\mu),$$

ce qui confirme la cohérence entre les définitions de $P(t)$ et Λ.

Ainsi, Λ et $P(t)$ sont intimement liées :

$$\Lambda = P'(0) \quad \text{et} \quad P(t) = e^{t\Lambda}.$$

La première décrit la **dynamique instantanée**, la seconde son **évolution temporelle globale**.

(6) Matrice infinitésimale

Rappelons que l'expression formelle de la matrice infinitésimale Λ est la suivante :

$$\Lambda = P'(0) \qquad \forall\, i,j \in \mathcal{S},\ \Lambda_{i,j} = \begin{cases} p'_{ij}(0), & i \neq j, \\ -\displaystyle\sum_{k \in \mathcal{S}\setminus\{i\}} p'_{ik}(0), & i = j. \end{cases}$$

Pour interpréter les termes Λ_{ij}, partons de la définition différentielle de $p'_{ij}(0)$:

$$p'_{ij}(0) = \lim_{h \to 0^+} \frac{p_{ij}(h) - \mathbb{1}_{\{i=j\}}}{h}.$$

Or, pour un processus de file d'attente de type $M/M/2/3$, les probabilités élémentaires sur un petit intervalle $[t, t+h]$ vérifient, à l'ordre $o(h)$:

$$\begin{cases} \mathbb{P}(N_{t+h} - N_t = 1 \mid N_t = i) = \lambda_i h + o(h), & \text{(une arrivée, si } i < 3), \\ \mathbb{P}(N_{t+h} - N_t = -1 \mid N_t = i) = \mu_i h + o(h), & \text{(un service terminé, si } i \geq 1), \\ \mathbb{P}(N_{t+h} - N_t = 0 \mid N_t = i) = 1 - (\lambda_i + \mu_i)h + o(h), & \text{(aucun événement)}, \\ \mathbb{P}(\text{plus d'un événement sur } [t, t+h]) = o(h). \end{cases}$$

où les taux dépendent de l'état i selon :

$$\lambda_i = \begin{cases} \lambda, & i < 3, \\ 0, & i = 3, \end{cases} \qquad \mu_i = \begin{cases} 0, & i = 0, \\ \mu, & i = 1, \\ 2\mu, & i = 2, 3. \end{cases}$$

Ainsi, pour $t = 0$:

$$p_{ij}(h) = \begin{cases} \lambda_i h + o(h), & \text{si } j = i+1, \\ \mu_i h + o(h), & \text{si } j = i-1, \\ 1 - (\lambda_i + \mu_i)h + o(h), & \text{si } j = i, \\ o(h), & \text{sinon.} \end{cases}$$

En divisant par h et en faisant tendre $h \to 0$, comme $p_{ij}(0) = 0$ pour $i \neq j$ (aucun changement d'état instantané), on obtient :

$$p'_{ij}(0) = \begin{cases} \lambda_i, & j = i+1, \\ \mu_i, & j = i-1, \\ -(\lambda_i + \mu_i), & j = i, \\ 0, & \text{sinon.} \end{cases}$$

Vue comme la matrice des taux instantanés, la matrice infinitésimale est donc donnée par :

$$\Lambda = \begin{pmatrix} -\lambda & \lambda & 0 & 0 \\ \mu & -(\lambda+\mu) & \lambda & 0 \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda \\ 0 & 0 & 2\mu & -2\mu \end{pmatrix}.$$

Interprétation.

— Les coefficients diagonaux $\Lambda_{ii} < 0$ représentent les **taux de sortie** de l'état i.

— Les coefficients hors diagonale positifs indiquent les **taux de transition** entre états :

$$\Lambda_{i,i+1} = \lambda_i, \quad \Lambda_{i,i-1} = \mu_i.$$

— Chaque ligne somme à zéro :

$$\sum_j \Lambda_{ij} = 0,$$

traduisant la **conservation des probabilités**.

(7) Loi stationnaire et stabilité

Une **loi stationnaire** $\pi = (\pi_0, \pi_1, \pi_2, \pi_3)$ vérifie :

$$\pi\Lambda = 0, \qquad \sum_{n=0}^{3} \pi_n = 1.$$

La **stabilité** d'une file M/M/c s'exprime par le taux de charge :

$$\rho = \frac{\lambda}{c\mu} < 1,$$

ce qui garantit que le système ne diverge pas. Ici, la capacité est finie ($K = 3$) : le système est **toujours stable**, même si $\rho > 1$, car les arrivées excédentaires sont simplement refusées.

(8) Résolution de la loi stationnaire

En régime stationnaire, le **principe de conservation des flux de probabilité** impose, pour chaque état n, l'égalité entre flux entrants et sortants. Ainsi :

$$\text{Flux entrant en } n = \text{Flux sortant de } n.$$

On démarre à $n = 1$, obtenant donc :

$$\lambda_0\pi_0 = \mu_1\pi_1.$$

Pour $n = 2$:

$$(\lambda_1 + \mu_1)\pi_1 = \lambda_0\pi_0 + \mu_2\pi_2 \quad \Longleftrightarrow \quad \mu_2\pi_2 = \lambda_1\pi_1.$$

Et donc, par récurrence immédiate, pour tout $n \geq 1$:

$$\pi_n\mu_n = \pi_{n-1}\lambda_{n-1}.$$

D'où les résultats numériques :

$$\pi_1 = \frac{\lambda}{\mu}\pi_0, \quad \pi_2 = \frac{\lambda^2}{2!\mu^2}\pi_0, \quad \pi_3 = \frac{\lambda^3}{2!2\mu^3}\pi_0 = \frac{\lambda^3}{4\mu^3}\pi_0.$$

En imposant la normalisation $\sum_{n=0}^{3} \pi_n = 1$, on obtient :

$$\pi_0 = \left(1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2} + \frac{\lambda^3}{4\mu^3}\right)^{-1}.$$

Les autres composantes s'en déduisent directement :

$$\pi_n = \pi_0 \cdot \frac{\lambda^n}{\mu^n \cdot \min(n, 2)!}.$$

(9) Taux de refus et métriques de performance

$$P_{\text{refus}} = \pi_3, \qquad \lambda_{\text{eff}} = \lambda(1 - \pi_3), \qquad U = \frac{\lambda_{\text{eff}}}{2\mu}.$$

Interprétation.

- π_3 est la probabilité stationnaire que le système soit plein : une arrivée est alors immédiatement refusée.
- Par le **principe PASTA** (*Poisson Arrivals See Time Averages*), les arrivées de Poisson "voient" le système selon ses proportions stationnaires. Ainsi, la probabilité qu'une arrivée soit bloquée est exactement égale à la probabilité que le système soit saturé :

$$P_{\text{perte}} = P_{\text{refus}} = \pi_3.$$

  Autrement dit, le **taux de perte effectif** est précisément égal au **taux de refus** .
- $\lambda_{\text{eff}} = \lambda \mathbb{P}(N_t < 3) = \lambda(1 - \pi_3)$ est donc le **taux d'entrée effectif** : il correspond au flux d'arrivées réellement acceptées et servies.
- Enfin, $U = \dfrac{\lambda_{\text{eff}}}{2\mu}$ mesure la **charge moyenne par serveur**. Si $\lambda \gg \mu$, la file est presque toujours saturée, donc $\pi_3 \to 1$ et $\lambda_{\text{eff}} \approx 0$: le système refuse la quasi-totalité des clients.

(10) Loi de Little et temps moyens

La **loi de Little** relie les quantités stationnaires :

$$\mathbb{E}[N] = \lambda_{\text{eff}} \mathbb{E}[T] \iff \mathbb{E}[T] = \frac{\mathbb{E}[N]}{\lambda_{\text{eff}}}.$$

Définitions des grandeurs :

- En régime stationnaire, la variable aléatoire N représente **le nombre total de clients présents dans le système (en service ou en attente)** lorsque le processus $(N_t)_{t\geq 0}$ a atteint son équilibre, c'est-à-dire :

$$\mathbb{E}[N] = \lim_{t\to\infty} \mathbb{E}[N_t].$$

- $L = \mathbb{E}[N] = \sum_{n=0}^{3} n\pi_n$: nombre moyen de clients dans le système ;
- $L_q = L - \frac{\lambda_{\text{eff}}}{\mu}$: nombre moyen de clients en file d'attente ; puisque $\frac{\lambda_{\text{eff}}}{\mu}$ représente le **nombre moyen de clients en service**.
- $W = \mathbb{E}[T]$: temps moyen passé dans le système ;
- $W_q = W - \frac{1}{\mu}$: temps moyen d'attente avant service.

Définitions et justification des grandeurs moyennes. En régime stationnaire, la distribution de probabilité du nombre de clients dans le système est donnée par la loi $(\pi_n)_{n=0,\dots,3}$, vérifiant

$$\pi_n = \mathbb{P}(N = n), \qquad \sum_{n=0}^{3} \pi_n = 1.$$

Cette loi représente la proportion de temps que le système passe dans chaque état à long terme. Par le **théorème ergodique**, les moyennes temporelles coïncident avec les moyennes d'espérance sous la loi stationnaire. Ainsi, toute grandeur stationnaire d'intérêt (nombre moyen de clients, temps moyen, etc.) s'obtient par espérance sur (π_n).

On note :

$$L = \mathbb{E}[N] = \sum_{n=0}^{3} n\,\pi_n.$$

C'est le **nombre moyen de clients présents dans le système** (en service ou en attente).

— D'après la relation d'équilibre $\pi_n = \pi_0 \frac{\lambda^n}{\mu^n \min(n,2)!}$, on a :

$$L = \pi_0 \left(0 \cdot 1 + 1\frac{\lambda}{\mu} + 2\frac{\lambda^2}{2\mu^2} + 3\frac{\lambda^3}{4\mu^3} \right)$$
$$= \pi_0 \left(\frac{\lambda}{\mu} + \frac{\lambda^2}{\mu^2} + \frac{3\lambda^3}{4\mu^3} \right),$$

avec

$$\pi_0 = \left(1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2} + \frac{\lambda^3}{4\mu^3} \right)^{-1}.$$

— **Nombre moyen de clients dans la file d'attente** :

$$L_q = L - \frac{\lambda_{\text{eff}}}{\mu}.$$

Le terme $\frac{\lambda_{\text{eff}}}{\mu}$ correspond au nombre moyen de clients en service (loi des flux stationnaires). En remplaçant $\lambda_{\text{eff}} = \lambda(1 - \pi_3)$, on obtient :

$$L_q = L - \frac{\lambda(1 - \pi_3)}{\mu}.$$

— **Temps moyen dans le système (loi de Little)** :

$$W = \frac{L}{\lambda_{\text{eff}}} = \frac{L}{\lambda(1 - \pi_3)}.$$

C'est le **temps d'attente total moyen** (file + service).

— **Temps moyen d'attente avant service** :

$$W_q = W - \frac{1}{\mu} = \frac{L}{\lambda(1 - \pi_3)} - \frac{1}{\mu}.$$

Le terme $\frac{1}{\mu}$ est la durée moyenne de service pour une loi exponentielle de paramètre μ.

Interprétation.

— L mesure la **charge globale du système** : nombre moyen de clients présents, toutes positions confondues.

— L_q quantifie la **congestion de la file** : un indicateur direct de la qualité de service.

— W est le **temps moyen passé par un client** depuis son arrivée jusqu'à sa sortie.

— W_q est le **temps d'attente pur**, excluant la durée de service.

Ces relations constituent les **formules fondamentales de la file d'attente M/M/2/3** : elles relient les performances temporelles (temps d'attente, congestion) aux caractéristiques structurelles du système (taux d'arrivée, taux de service et capacité).

Ces relations permettent d'évaluer le confort de service et le dimensionnement optimal du système.

(11) Interprétation globale

Le modèle **M/M/2/3** illustre une situation concrète où deux serveurs travaillent en parallèle avec une capacité d'attente limitée (par exemple un centre d'appel ou un service hospitalier).

— Si λ augmente, les refus deviennent plus fréquents (π_3 augmente) et la qualité de service diminue.

— Si μ augmente (meilleure efficacité des serveurs), les files se résorbent et le taux de refus décroît.

— Si K ou c sont augmentés, on améliore la capacité d'absorption du système au prix d'un coût supplémentaire.

Ainsi, l'étude de la file M/M/2/3 permet de **trouver un équilibre entre coût et performance**, en quantifiant rigoureusement la charge, les temps d'attente et le taux de refus.

8.2 Projet de synthèse : Optimisation du nombre de serveurs dans un système $M/M/c/K$

8.2.1 Contexte et problématique

Une entreprise de services souhaite dimensionner le nombre optimal de serveurs à embaucher au guichet d'un centre d'accueil. Les arrivées de clients sont modélisées par un processus de Poisson d'intensité variable :

$$\lambda(t) = \begin{cases} \lambda_{\text{matin}}, & 8h \leq t < 17h, \\ \lambda_{\text{soir}}, & 17h \leq t \leq 20h, \end{cases}$$

avec $\lambda_{\text{soir}} > \lambda_{\text{matin}}$.

Les durées de service sont indépendantes et suivent une loi exponentielle de paramètre μ. Le système est modélisé par une file d'attente $M/M/c/K$ où :

— c : nombre de serveurs disponibles ;

— K : capacité maximale du système (clients servis + en attente) ;

— λ : taux d'arrivée moyen ;

— μ : taux de service par serveur.

L'entreprise souhaite **maximiser son gain économique quotidien** en choisissant judicieusement le nombre de serveurs c.

8.2.2 Partie I — Modélisation théorique et optimisation

1. Loi stationnaire et indicateurs de performance. Pour un système $M/M/c/K$, la loi stationnaire $(\pi_n)_{n=0}^{K}$ vérifie :

$$\pi_n = \begin{cases} \dfrac{\rho^n}{n!}\pi_0, & 0 \leq n \leq c, \\ \dfrac{\rho^n}{c!c^{n-c}}\pi_0, & c < n \leq K, \end{cases}$$

avec $\rho = \lambda/\mu$, et π_0 déterminé par la normalisation

$$\sum_{n=0}^{K} \pi_n = 1.$$

On rappelle que le taux d'arrivée effectif est :

$$\lambda_{\text{eff}} = \lambda(1 - \pi_K).$$

Le nombre moyen de clients dans le système :

$$L = \sum_{n=0}^{K} n\pi_n, \qquad L_q = L - \frac{\lambda_{\text{eff}}}{\mu}.$$

2. Fonction économique de coût et gain. Soient :

— r : revenu moyen par client servi ;

— s : salaire horaire d'un serveur ;

— p : pénalité moyenne (ou coût d'opportunité) pour un client refusé.

Alors, le gain économique horaire espéré s'écrit :

$$G(c) = r\lambda_{\text{eff}} - sc - p\lambda\pi_K.$$

L'objectif est d'optimiser :

$$c^* = \underset{1 \leq c \leq c_{\max}}{\operatorname{argmax}}\ G(c)$$

où $c_{\max}$ est le nombre maximal de serveurs embauchables.

3. Discussion. Le paramètre K reflète la **capacité physique du système** :

— si K est petit, les refus augmentent $\Rightarrow$ pertes de revenu ;

— si K est grand, les clients attendent plus longtemps $\Rightarrow$ satisfaction réduite.

Ainsi, une étude de sensibilité sur K est pertinente conjointement à l'optimisation de c.

8.2.3 Partie II — Simulation et implémentation Python

Objectif. Simuler différentes configurations (c, K) pour estimer $G(c)$, puis tracer le gain attendu et déterminer le c^* optimal.

```python
import numpy as np
import matplotlib.pyplot as plt

def pi_stationnaire_MMckK(lam, mu, c, K):
    """Compute stationary distribution  for M/M/c/K."""
    rho = lam / mu
    pi = np.zeros(K+1)
    # Normalizing constant
    sum_terms = sum(rho**n / np.math.factorial(n) for n in range(c+1))
    sum_terms += sum(rho**n / (np.math.factorial(c) * c**(n-c)) for n in range(c+1, K+1))
    pi[0] = 1 / sum_terms
    # Recursive definition
    for n in range(1, K+1):
        if n <= c:
            pi[n] = (rho**n / np.math.factorial(n)) * pi[0]
        else:
            pi[n] = (rho**n / (np.math.factorial(c) * c**(n-c))) * pi[0]
    return pi

def gain_MMckK(lam, mu, c, K, r, s, p):
    """Expected gain function."""
    pi = pi_stationnaire_MMckK(lam, mu, c, K)
    lambda_eff = lam * (1 - pi[K])
    G = r * lambda_eff - s * c - p * lam * pi[K]
    return G, pi, lambda_eff

# Parameters
lam, mu = 8, 5       # rates
r, s, p = 10, 20, 5  # economic parameters
K = 10

c_values = range(1, 8)
gains = [gain_MMckK(lam, mu, c, K, r, s, p)[0] for c in c_values]

# Optimal number of servers
c_opt = c_values[np.argmax(gains)]

plt.plot(c_values, gains, marker='o')
plt.axvline(c_opt, color='r', linestyle='--', label=f'c* = {c_opt}')
plt.xlabel('Number of servers c')
plt.ylabel('Expected gain G(c)')
plt.title('Optimization of number of servers in M/M/c/K system')
plt.legend()
plt.grid()
plt.show()
```

Commentaires.

— La simulation calcule la loi stationnaire puis le gain $G(c)$ pour chaque configuration.

— Le maximum indique le nombre optimal de serveurs à embaucher c^*.

— On peut réitérer l'étude pour différents λ_{matin}, λ_{soir} afin de planifier des embauches dynamiques.

Extension possible. On pourrait enrichir le modèle en rendant $\lambda(t)$ dépendant du temps de façon continue (modèle non stationnaire) et adapter dynamiquement $c(t)$ en conséquence.

Remarque 18. Le package `BirDePy`[1] (*Birth–Death Processes in Python*) permet de modéliser et de simuler aisément des processus de naissance–mort, dont les files d'attente $M/M/c/K$ constituent un cas particulier. Il offre :

— une représentation matricielle directe de la matrice génératrice Λ ;

— le calcul automatique de la loi stationnaire (π_n) par résolution de $\pi\Lambda = 0$;

— des fonctions intégrées de simulation temporelle (trajectoires (N_t)) et de visualisation.

Exemple d'utilisation pratique :

```python
from birdepy import BirthDeathProcess

# Paramètres du modèle M/M/c/K
lam = 8.0      # taux d'arrivée
mu = 5.0       # taux de service par serveur
c, K = 2, 5

# Définition des taux dépendant de l'état
lambda_n = [lam if n < K else 0 for n in range(K+1)]
mu_n = [n*mu if n <= c else c*mu for n in range(K+1)]

# Création du processus
bdp = BirthDeathProcess(lambda_n=lambda_n, mu_n=mu_n)

# Loi stationnaire
pi = bdp.stationary_distribution()
print("Distribution stationnaire:", pi)

# Simulation de trajectoire
trajectory = bdp.simulate(t_max=50)
bdp.plot_trajectory(trajectory)
```

Avantage. L'emploi de `BirDePy` simplifie grandement la partie d'implémentation : il évite les boucles et normalisations manuelles, permet de visualiser les trajectoires et d'estimer rapidement des quantités telles que L, L_q, ou encore le taux de refus π_K.

1. Hautphenne, S., & Patch, B. (2024). Birth-and-death processes in python : The birdepy package. Journal of Statistical Software, 111, 1-54.