Procesos Estocásticos y Probabilidad

Cadenas de Markov en Tiempo Continuo

Master en Modelización e Investigación Matemática, Estadística y Computación

Josu Doncel Vicente

Departamento de Matemáticas

Universidad del País Vasco, EHU

Estos apuntes presentan los contenidos de la parte de cadenas de Markov en tiempo continuo, de la asignatura Procesos Estocásticos y Probabilidad del Máster en Modelización e Investigación Matemática, Estadística y Computación. Para cualquier asunto relacionado con estos apuntes, no dudeis en contactar con el profesor responsable.

- Nombre: Josu Doncel Vicente
- Dirección: Universidad del Pais Vasco, EHU. Facultad de Ciencia y Tecnología. Departamento de Matemáticas. Barrio Sarriena n/n. 48940, Leioa
- Teléfono: (+34) 94 601 78 95
- Email: josu.doncel@ehu.eus
- Web: http://josudoncel.github.io/

Índice general

1. Introducción a las Cadenas de Markov en Tiempo Continuo **1**

1.1. Cadenas de Markov en Tiempo Continuo 1

1.1.1. Procesos estocásticos 1

1.1.2. Procesos estocásticos en tiempo continuo 2

1.1.3. Propiedad de Markov 2

1.1.4. Definición de cadena de Markov en tiempo continuo 3

1.2. Procesos de nacimiento y muerte 5

1.2.1. Procesos de nacimiento puro y muerte pura 6

1.2.2. Ejemplos 7

1.3. Ecuaciones de Kolmogorov 8

1.3.1. Ecuaciones backward 8

1.3.2. Ecuaciones forward 9

1.3.3. Ejemplo 10

1.4. Análisis en régimen estacionario 11

1.4.1. Teorema ergódico 12

1.4.2. Ejemplo: proceso de nacimiento y muerte 13

1.5. Cadenas Encajadas y Uniformizadas 15

1.5.1. Cadenas Encajadas 15

1.5.2. Cadenas uniformizadas 16

2. Análisis de una Única Cola **17**

2.1. Componentes fundamentales de un sistema de colas 17

2.1.1. Llegadas 17

2.1.2. Servicio 18

2.1.3. Sala de espera o cola 18

2.1.4. Servidores 18

2.2. Notación de Kendall 18

2.3. Ley De Little 19

2.4. Propiedad PASTA 20

2.5. El modelo $M/M/1$ 20

2.5.1. Ecuaciones de balance y distribución estacionaria 21

2.5.2. Condición de estabilidad 22

2.5.3. Comparación de sistemas 23

2.6. El modelo $M/M/1/K$ . . . 24
2.6.1. Distribución estacionaria . . . 25
2.6.2. Probabilidad de bloqueo . . . 25
2.7. Los modelos $M/M/K$ y $M/M/K/K$ . . . 26
2.7.1. El modelo $M/M/K$ . . . 26
2.7.2. El modelo $M/M/K/K$ . . . 28
2.7.3. Aplicación al diseño de un call center . . . 30
2.8. El modelo $M/M/\infty$ . . . 32
2.8.1. Distribución estacionaria . . . 33

3. Análisis de Redes de Colas **35**
3.1. Colas en tándem . . . 35
3.1.1. Transiciones de la cadena . . . 36
3.1.2. Condición de estabilidad . . . 37
3.1.3. Distribución estacionaria . . . 37
3.1.4. Medidas de rendimiento . . . 38
3.1.5. Ejemplo . . . 39
3.2. Redes acíclicas de colas . . . 42
3.2.1. Descripción general del modelo . . . 43
3.2.2. Ecuaciones de tráfico . . . 44
3.2.3. Distribución estacionaria . . . 44
3.2.4. Medidas de rendimiento . . . 45
3.2.5. Ejemplo: una red acíclica de dos estaciones . . . 45
3.3. Redes abiertas de Jackson . . . 50
3.3.1. Descripción general del modelo . . . 50
3.3.2. Transiciones de la cadena . . . 51
3.3.3. Teorema de Jackson . . . 53
3.3.4. Medidas de rendimiento . . . 53
3.3.5. Ejemplo: una red de Jackson de dos estaciones . . . 54
3.4. Redes cerradas de colas . . . 60
3.4.1. Descripción general del modelo . . . 60
3.4.2. Transiciones de la cadena . . . 61
3.4.3. Razones de visita . . . 62
3.4.4. Teorema de Gordon–Newell . . . 62
3.4.5. Existencia de la distribución estacionaria . . . 63
3.4.6. Medidas de rendimiento . . . 63
3.4.7. Ejemplo: una red cerrada de dos estaciones . . . 64

1

Introducción a las Cadenas de Markov en Tiempo Continuo

1.1. Cadenas de Markov en Tiempo Continuo

Antes de definir una cadena de Markov en tiempo continuo, introducimos primero los conceptos de proceso estocástico y proceso estocástico en tiempo continuo. A lo largo de estos apuntes vamos a suponer que $(\Omega, \mathcal{F}, \mathbb{P})$ es un espacio de probabilidad, es decir, que Ω es un conjunto, $\mathcal{F}$ es una σ-algebra sobre Ω y $\mathbb{P}$ es una medida de probabilidad.

1.1.1. Procesos estocásticos

Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad y sea S un conjunto, llamado *espacio de estados*. Un *proceso estocástico* con espacio de estados S es una familia de variables aleatorias

$$\{X_t\}_{t \in T},$$

definidas sobre $(\Omega, \mathcal{F}, \mathbb{P})$ y con valores en S, donde T es un conjunto de índices, que habitualmente representa el tiempo.

Es decir, para cada $t \in T$, la aplicación

$$X_t : \Omega \longrightarrow S$$

es una variable aleatoria. La interpretación es que X_t describe el estado del sistema en el instante t.

Dependiendo del conjunto de índices T, distinguimos dos casos principales:

- Si $T = \{0, 1, 2, \dots\}$ o, más generalmente, $T \subseteq \mathbb{Z}$, hablamos de un *proceso estocástico en tiempo discreto*.

- Si $T = [0, \infty)$ o, más generalmente, $T \subseteq \mathbb{R}$, hablamos de un *proceso estocástico en tiempo continuo*.

A lo largo de estos apuntes nos interesarán principalmente procesos en tiempo continuo con espacio de estados discreto o numerable.

1.1.2. Procesos estocásticos en tiempo continuo

Un *proceso estocástico en tiempo continuo* es un proceso estocástico indexado por un parámetro temporal continuo, esto es, una familia

$$\{X_t\}_{t\geq 0}$$

de variables aleatorias definidas sobre un mismo espacio de probabilidad y con valores en un espacio de estados S.

La diferencia esencial con los procesos en tiempo discreto es que el sistema puede evolucionar en cualquier instante $t \geq 0$, y no únicamente en tiempos enteros o en una sucesión discreta de instantes. En consecuencia, para describir completamente el comportamiento del proceso es necesario conocer las distribuciones conjuntas de las variables

$$X_{t_1}, X_{t_2}, \dots, X_{t_n}, \qquad 0 \leq t_1 < t_2 < \cdots < t_n.$$

En muchos modelos de interés, el espacio de estados S es finito o numerable. En ese caso, el proceso describe la evolución aleatoria de un sistema que va cambiando de estado a lo largo del tiempo.

1.1.3. Propiedad de Markov

Un proceso estocástico en tiempo continuo $\{X_t\}_{t\geq 0}$ con espacio de estados S se dice que satisface la *propiedad de Markov* si, para cualesquiera instantes

$$0 \leq t_0 < t_1 < \cdots < t_n < t$$

y cualesquiera estados $i_0, i_1, \dots, i_n, j \in S$, se verifica que

$$\mathbb{P}\big(X_t = j \mid X_{t_n} = i_n, X_{t_{n-1}} = i_{n-1}, \dots, X_{t_0} = i_0\big) = \mathbb{P}\big(X_t = j \mid X_{t_n} = i_n\big),$$

siempre que la probabilidad condicionante esté bien definida.

Esta propiedad significa que, dado el estado actual del sistema, su evolución futura no depende del pasado. En otras palabras, el presente contiene toda la información relevante para predecir el futuro del proceso.

Cuando, además, las probabilidades de transición dependen únicamente de la diferencia de tiempos y no de los instantes concretos, se dice que el proceso es *homogéneo en el tiempo*. En ese caso,

$$\mathbb{P}(X_{t+s} = j \mid X_s = i) = \mathbb{P}(X_t = j \mid X_0 = i), \qquad s, t \geq 0,\ i, j \in S.$$

1.1.4. Definición de cadena de Markov en tiempo continuo

Una *cadena de Markov en tiempo continuo* (CMTC) es un proceso estocástico en tiempo continuo

$$\{X_t\}_{t \geq 0}$$

que verifica las dos propiedades siguientes:

1. su espacio de estados S es finito o numerable;

2. satisface la propiedad de Markov.

Si además el proceso es homogéneo en el tiempo, entonces para $i, j \in S$ definimos las *probabilidades de transición* por

$$P_{ij}(t) = \mathbb{P}(X_t = j \mid X_0 = i), \qquad t \geq 0.$$

Las cadenas de Markov en tiempo continuo constituyen una generalización natural de las cadenas de Markov en tiempo discreto: el sistema sigue moviéndose entre estados de un conjunto discreto, pero ahora los cambios pueden producirse en cualquier instante del tiempo.

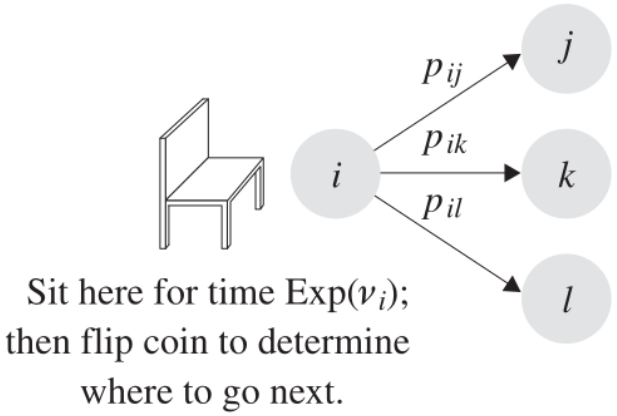
Observación 1.1.1 (Relación con la distribución exponencial)**.** *Si un proceso está durante s instantes de tiempo en el estado i (es decir, no se mueve a otro estado) , cual es la probabilidad de que esté en el mismo estado los siguientes t instantes de tiempo?*
Por la propiedad de Markov, la probabilidad de quedarse en un estado en el intervalo $(s, s+t)$ es la misma que quedarse t unidades de tiempo. Es decir, si τ_i es el tiempo que el proceso se queda en el estado i, la propiedad de Markov dice que

$$\mathbb{P}(\tau_i > s + t | \tau_i > s) = \mathbb{P}(\tau_i > t),$$

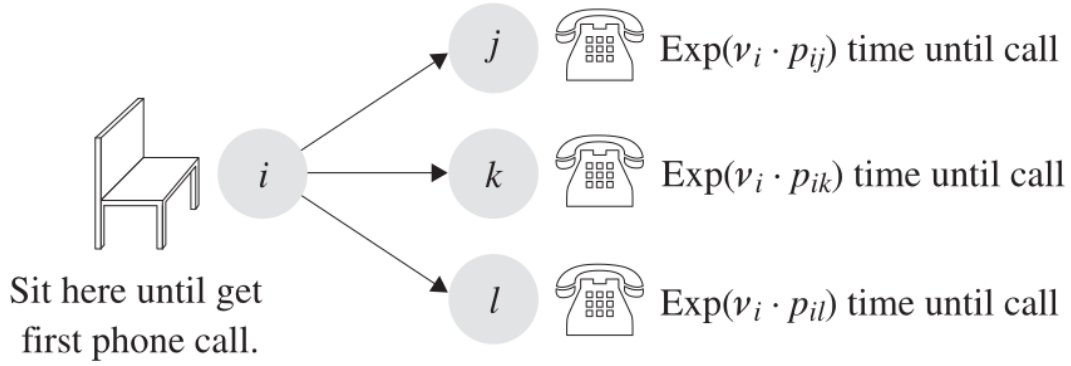
para todo $s, t > 0$. Esto implica que τ_i cumple la condición de pérdida de memoria y, por tanto, τ_i sigue una distribución exponencial.

Como consecuencia de la anterior observación, hay dos formas equivalentes de ver las cadenas de Markov en tiempo continuo:

(1) Al entrar en el estado i, se queda un tiempo exponencial de parámetro ν_i y después va al estado j con probabilidad $p_{i,j}$, con $\sum_{j \neq i} p_{i,j} = 1$.



(2) Para cada $i \neq j$, puede introducirse un reloj exponencial independiente con parámetro $\nu_i p_{i,j}$ y el próximo salto es al estado cuyo reloj "suena primero".



Se puede demostrar que ambas visiones son equivalentes. Ahora vamos a demostrar que (2) implica (1).

Vamos a demostrar que, si se da (2), entonces: (a) el tiempo de salir del estado i es exponencial de parámetro ν_i, y (b) la probabilidad de moverse del estado i al estado j es $p_{i,j}$.

Empezamos demostrando (a). Sea X_j el tiempo de moverse del estado i al estado j. Por la definición (2), $\min_{j\neq i}\{X_j\}$ es el tiempo de salir del estado i. Como X_i son exponenciales de parámetro $\nu_i p_{i,j}$, se tiene que $\min_{j\neq i}\{X_j\}$ sigue una distribución exponencial de parámetro $\sum_{j\neq i}\nu_i p_{i,j}$. Por tanto, (a) queda demostrado porque

$$\sum_{j\neq i}\nu_i p_{i,j} = \nu_i.$$

Ahora demostramos (b). La probabilidad de moverse al estado k del estado i es:

$$\mathbb{P}(X_k = \min_{j\neq i}\{X_j\}) = \frac{\nu_i p_{i,k}}{\sum_{j\neq i}\nu_i p_{i,j}} = p_{i,k}.$$

Observación 1.1.2. *En la anterior demostración, se han usado propiedades conocidas de la exponencial, como que el minimo de exponenciales de parámetros $\lambda_1, \dots, \lambda_n$ es $\sum_i \lambda_i$ y también que la probabilidad de que el mínimo de exponenciales de parámetros $\lambda_1, \dots, \lambda_n$ sea una de ellas (por ejemplo, la que tiene por parámetro λ_k) es $\lambda_k/(\sum_i \lambda_i)$.*

Definimos el estado i como instantáneo si $\nu_i = \infty$, es decir, si al llegar al estado i se va a otro estado inmediatamente. Vamos a suponer que todos los estados son no-instantáneos, es decir, $\nu_i < \infty$ para todo estado i.

Observación 1.1.3. *En tiempo continuo las cosas pueden cambiar considerablemente con respecto a tiempo discreto, en donde hay una transición por unidad de tiempo (y por tanto, en tiempo finito hay finitas transiciones). Por ejemplo, en tiempo continuo se pueden dar un número infinito de transiciones en tiempo finito. Veamos el siguiente ejemplo.*

Consideramos una cadena de Markov en tiempo continuo con espacio de estados en $\mathbb{Z}$ tal que del estado i sólo se puede mover al estado $i+1$. El tiempo de permanencia en el estado i es exponencial de media $1/i^2$. Entonces, el tiempo esperado de que se transiten todos los estados (es decir, que se den infinitas transiciones) es

$$\sum_i \frac{1}{i^2} < \infty.$$

Definimos una cadena de Markov en tiempo continuo como regular si (casi seguramente) el número de transiciones en cualquier intervalo de tiempo finito es finito. Definimos también la tasa de transición del estado i al estado j de la siguiente manera:

$$q_{i,j} = \nu_i p_{i,j}, \ \forall i \neq j,$$

La matriz generadora de una cadena de Markov en tiempo continuo es $Q = (q_{i,j})_{i,j \in S}$, es decir, es una matrix cuyo elemento (i,j) es la tasa de transición del estado i al estado j, si $i \neq j$ y cuya diagonal viene dada por

$$q_{ii} = -\sum_{j \neq i} q_{i,j}.$$

1.2. Procesos de nacimiento y muerte

Una clase particularmente importante de cadenas de Markov en tiempo continuo es la de los *procesos de nacimiento y muerte*. Se trata de procesos cuyo espacio de estados es

$$S = \{0, 1, 2, \dots\},$$

y en los que, desde cada estado, sólo son posibles transiciones a los estados vecinos.

Un *proceso de nacimiento y muerte* es una cadena de Markov en tiempo continuo $\{X_t\}_{t\geq 0}$ con espacio de estados $S = \mathbb{N}_0$ tal que, para cada $i \geq 0$,

$$q_{i,i+1} = \lambda_i, \qquad q_{i,i-1} = \mu_i,$$

donde $\lambda_i \geq 0$ y $\mu_i \geq 0$, con la convención $\mu_0 = 0$, y además

$$q_{ij} = 0 \qquad \text{si } |j - i| > 1.$$

La cantidad λ_i se denomina *tasa de nacimiento* en el estado i, y μ_i se denomina *tasa de muerte* en el estado i.

La diagonal de la matriz generadora viene dada por

$$q_{ii} = -(\lambda_i + \mu_i),$$

de modo que la matriz generadora del proceso es

$$Q = (q_{ij})_{i,j\geq 0}, \qquad q_{ij} = \begin{cases} \lambda_i, & j = i+1, \\ \mu_i, & j = i-1, \\ -(\lambda_i + \mu_i), & j = i, \\ 0, & \text{en otro caso.} \end{cases}$$

Estos procesos aparecen de forma natural en numerosos contextos aplicados: dinámica de poblaciones, teoría de colas, fiabilidad, genética o epidemiología. En todos ellos, el estado representa el tamaño de la población bajo consideración.

1.2.1. Procesos de nacimiento puro y muerte pura

Los procesos de nacimiento y muerte más sencillos se obtienen cuando sólo se permiten transiciones en una dirección.

Un *proceso de nacimiento puro* es un proceso de nacimiento y muerte para el cual

$$\mu_i = 0, \qquad i \geq 0.$$

En consecuencia, desde el estado i el único salto posible es al estado $i + 1$, con tasa λ_i.

En un proceso de nacimiento puro, si el sistema está en el estado i, el tiempo de permanencia en dicho estado es exponencial de parámetro λ_i, y tras ese tiempo el proceso pasa necesariamente al estado $i + 1$.

Un *proceso de muerte pura* es un proceso de nacimiento y muerte para el cual

$$\lambda_i = 0, \qquad i \geq 0.$$

En consecuencia, desde el estado $i \geq 1$ el único salto posible es al estado $i - 1$, con tasa

μ_i.

En un proceso de muerte pura, si el sistema está en el estado $i \geq 1$, el tiempo de permanencia en dicho estado es exponencial de parámetro μ_i, y tras ese tiempo el proceso pasa necesariamente al estado $i-1$. Al llegar al estado 0, el sistema se queda en este estado para siempre.

1.2.2. Ejemplos

- El proceso de Poisson homogéneo de parámetro $\lambda > 0$ puede interpretarse como un proceso de nacimiento puro con tasas constantes

$$\lambda_i = \lambda, \qquad i \geq 0.$$

- El proceso de Yule es un proceso de nacimiento puro en el que cada individuo de la población da lugar a un nuevo individuo de manera independiente, con una tasa constante por individuo.

  El *proceso de Yule* con parámetro $\lambda > 0$ es el proceso de nacimiento puro cuyas tasas de nacimiento son

$$\lambda_i = \lambda i, \qquad i \geq 0.$$

  Así, cuando hay i individuos en la población, la tasa total de nacimiento es proporcional a i, lo que refleja que cada individuo se reproduce independientemente a razón λ.

  Tanto en el proceso de Poisson como el de Yule, el estado i indica que hay i individuos en la población.

- Vamos a analizar ahora un modelo de crecimiento lineal con inmigración. Suponemos que tenemos una población en la que:

  - Cada individuo se reproduce con un tiempo exponencial independiente de parámetro λ
  - Debido a la inmigración, una nueva persona se une a la población con tiempo exponencial de parámetro θ (y este tiempo es independiente de las reproducciones)
  - La muerte de un individuo se da con tiempo exponencial de parámetro μ. Las muertes se dan de manera independiente.

  Este modelo es un proceso de nacimiento y muerte, en el que el estado representa el número de individuos de la población. Para cualquier estado $i > 0$ del proceso, las tasas de nacimiento y muerte en el estado i vienen dadas por

$$\lambda_i = i\lambda + \theta, \quad \mu_i = i\mu,$$

  mientras que $\lambda_0 = \theta$ y $\mu_0 = 0$.

- Vamos a analizar un modelo epidémico simple. Consideramos una población en donde hay dos posibles estados, susceptible o infectado, y en la que los infectados no pueden volver a ser susceptibles. Suponemos que, en el instante inicial, hay un infectado y $m-1$ susceptibles. Cada uno de los infectados infecta a un susceptible con tiempo exponencial de tasa α e independiente de las infecciones de otros.

  En este modelo, el número de infectados en el instante t es un proceso de nacimiento puro en el que, en el estado i,

  $$\lambda_i = (m-i)i\lambda,$$

  si $i = 1, 2, \dots, m-1$ y $\lambda_i = 0$ en otro caso.

1.3. Ecuaciones de Kolmogorov

Sea $\{X_t\}_{t\geq 0}$ una cadena de Markov en tiempo continuo con espacio de estados S y matriz generadora Q. Recordamos que

$$P_{i,j}(t) := \mathbb{P}(X_t = j \mid X_0 = i),$$

y que

$$q_{i,i} = -\sum_{j\neq i} q_{ij} =: -\nu_i,$$

donde ν_i es la tasa de salida del estado i.

El comportamiento en intervalos pequeños de la cadena de Markov es el siguiente:

- $\mathbb{P}(\text{ningún salto en } [0,h]) = 1 - \nu_i h + o(h)$,
- $\mathbb{P}(\text{salto de } i \text{ a } j \text{ en } [0,h]) = q_{ij}h + o(h)$,
- La probabilidad de más de un salto en $[0,h]$ es $o(h)$.

1.3.1. Ecuaciones backward

Observamos que

$$P_{i,j}(t+h) = \sum_{k\in S} P_{i,k}(h)P_{kj}(t) = P_{i,i}(h)P_{i,j}(t) + \sum_{k\neq i} P_{i,k}(h)P_{k,j}(t) = \\ (1-\nu_i h + o(h))\, P_{i,j}(t) + \sum_{k\neq i}(q_{ik}h + o(h))\, P_{k,j}(t).$$

Restamos $\mathbb{P}_{i,j}(t)$ en ambos lados de la igualdad:

$$P_{i,j}(t+h) - P_{i,j}(t) = -\lambda_i h\, P_{i,j}(t) + \sum_{k\neq i} q_{i,k}h\, P_{k,j}(t) + o(h).$$

Dividimos por h:

$$\frac{P_{i,j}(t+h) - P_{i,j}(t)}{h} = \sum_k q_{i,k} P_{k,j}(t) + o(1).$$

Tomando $h \to 0$, obtenemos la ecuación backwards:

$$\frac{d}{dt} P_{i,j}(t) = \sum_k q_{i,k} P_{k,j}(t)$$

En forma matricial, se puede escribir de la siguiente manera:

$$\mathbb{P}'(t) = Q\mathbb{P}(t), \qquad \mathbb{P}(0) = I.$$

Es decir, la ecuación backward es una ecuación diferencia ordinaria matricial, cuya solución es:

$$\mathbb{P}(t) = e^{tQ} = I + tQ + \frac{t^2}{2!}Q^2 + \cdots .$$

1.3.2. Ecuaciones forward

Usando argumentos similares a la ecuación backward, obtenemos la ecuación forward:

$$\frac{d}{dt} P_{i,j}(t) = \sum_k P_{i,k}(t)\, q_{k,j}.$$

Para ver cómo obtenemos esta ecuación, usaremos

$$P_{i,j}(t+h) = P_{i,j}(h)P_{j,j}(t) + \sum_{k \neq j} P_{ik}(h)P_{k,j}(t).$$

Teniendo en cuenta el comportamiento en tiempos pequeños visto anteriormente, sustituimos:

$$P_{i,j}(t+h) = P_{i,j}(t)(1 - \nu_j h + o(h)) + \sum_{k \neq j} P_{i,k}(t)(q_{k,j}h + o(h)).$$

Restamos $P_{i,j}(t)$ y dividimos por h:

$$\frac{P_{i,j}(t+h) - P_{i,j}(t)}{h} = \sum_k P_{ik}(t) q_{k,j} + o(1).$$

Tomando $h \to 0$ obtenemos la ecuación forward, que en forma matricial es:

$$\mathbb{P}'(t) = \mathbb{P}(t)\, Q, \qquad P(0) = I.$$

La solución de esta ecuación diferencia matricial es también

$$\mathbb{P}(t) = e^{tQ} = I + tQ + \frac{t^2}{2!}Q^2 + \cdots .$$

1.3.3. Ejemplo

Consideremos una máquina que puede estar en dos estados: 1 (en funcionamiento) y 2 (estropeada). La máquina se estropena con un tiempo exponencial de parámetro λ (es decir, se estropean λ veces por unidad de tiempo) y la máquina se repara con un tiempo exponencial de parámetro μ (es decir, se reparan μ máquinas por unidad de tiempo). Esta situación se puede modelizar como una cadena de Markov en tiempo continuo con dos estados. Ver Figura 1.1.

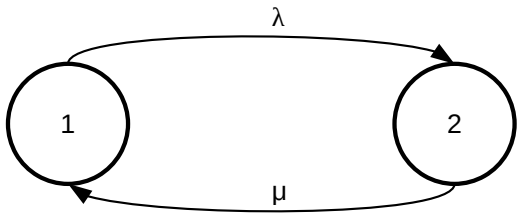


Figura 1.1: Cadena de Markov de dos estados.

La matriz de tasas de transición de esta cadena de Markov es

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}, \qquad \lambda, \mu > 0.$$

Las ecuaciones backward son:

$$\begin{cases} P'_{1,1}(t) = -\lambda P_{1,1}(t) + \mu P_{2,1}(t), \\ P'_{2,1}(t) = \lambda P_{1,1}(t) - \mu P_{2,1}(t), \end{cases}$$

Vamos a suponer $P_{1,1}(0) = 1$ y $P_{2,1}(0) = 0$, es decir que la máquina está en funcionamiento en el instante inicial.

Se puede obtener la solución de la ecuacion e^{tQ} para este caso

$$\mathbb{P}(t) = e^{Qt} = \begin{pmatrix} \pi_1 + \pi_2 e^{-(\lambda+\mu)t} & \pi_2 - \pi_2 e^{-(\lambda+\mu)t} \\ \pi_1 - \pi_1 e^{-(\lambda+\mu)t} & \pi_2 + \pi_1 e^{-(\lambda+\mu)t} \end{pmatrix},$$

donde

$$\pi_1 = \frac{\mu}{\lambda+\mu}, \qquad \pi_2 = \frac{\lambda}{\lambda+\mu}.$$

Es decir,

$$P_{1,1}(t) = \pi_1 + \pi_2 e^{-(\lambda+\mu)t}, \qquad P_{1,2}(t) = \pi_2(1 - e^{-(\lambda+\mu)t}),$$

$$P_{2,1}(t) = \pi_1(1 - e^{-(\lambda+\mu)t}), \qquad P_{2,2}(t) = \pi_2 + \pi_1 e^{-(\lambda+\mu)t}.$$

Como ejemplo, consideremos $\lambda = 1$ y $\mu = 2$. En este caso, tenemos que $\pi_1 = 2/3$ y $\pi_2 = 1/3$. La Figura 1.2, ilustra $P_{1,1}(t)$ para este ejemplo cuando t varía entre 0 y 20.

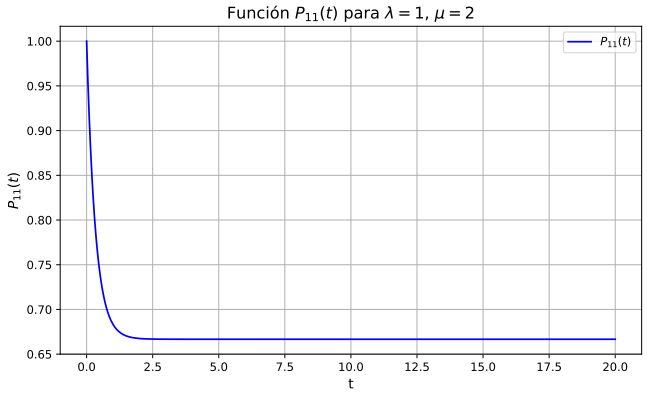


Figura 1.2: Evolución de $P_{1,1}(t)$ cuando t varía entre 0 y 20.

Esta figura muestra que $P_{11}(t)$ tiende a π_1 cuando $t \to \infty$ (esto se puede deducir también de la expresión obtenida) y que esta convergencia es muy rápida; es, de hecho, exponencial.

Es muy frecuente el análisis de cadenas de Markov en tiempo continuo cuando t tiende a infinito, es decir, en régimen estacionario. Desafortunadamente, el cálculo de $P_{i,j}(t)$ puede ser muy difícil cuando la cadena de Markov es compleja. Por tanto, el análisis en régimen estacionario de una cadena de Markov calculando previamente $P_{i,j}(t)$ no parece ser una buena opción. En el siguiente capítulo vamos a ver un método alternativo.

1.4. Análisis en régimen estacionario

En esta sección estudiamos el comportamiento de una cadena de Markov en tiempo continuo $\{X_t\}_{t\geq 0}$ cuando t tiende a infinito.

En primer lugar, vamos a definir el concepto de distribución límite. Si existe $\pi_j = \lim_{t\to\infty} P_{ij}(t) > 0$ para todo $j \in S$ (que es independiente del estado inicial $i \in S$), tal que $\sum_{j\in S} \pi_j = 1$, entonces $(\pi_j)_{j\in S}$ es la distribución límite.

Nuestro objetivo es describir su distribución límite y las condiciones bajo las cuales dicha distribución existe y es independiente del estado inicial.

Recordemos la ecuación forward: $P'(t) = P(t)Q$, donde $P(t) = (P_{ij}(t))_{i,j\in S}$ y Q es la matriz generadora.

Vamos a usar esta fórmula considerando que $t \to \infty$. Vamos a suponer que $\mathbb{P}'(t) = 0$ cuando $t \to \infty$ (en la Figura 1.2 se ve que se cumple para $P_{1,1}(t)$). En este caso, obtenemos lo siguiente:

$$0 = \vec{\pi} Q.$$

Suponiendo que existe, veremos que ésta es una forma de calcular la distribución límite. Usaremos el ejemplo con dos estados anterior, es decir

$$0 = (\pi_1, \pi_2) \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix} \iff \lambda \pi_1 = \mu \pi_2.$$

Observación 1.4.1. *Hemos obtenido a partir de $0 = \vec{\pi} Q$ la siguiente ecuación: $\lambda \pi_1 = \mu \pi_2$. Esta ecuación se puede interpretar de la siguiente manera: para cada estado, la tasa de salida es igual a la tasa de llegar (es decir, existe una conservación de flujos de entrada/salida en todos los estados).*

- *Estado 1: la tasa de salida es $\lambda \pi_1$ y la tasa de llegada es $\mu \pi_0$*
- *Estado 2: la tasa de llegada es $\lambda \pi_1$ y la tasa de salida es $\mu \pi_0$*

Teniendo en cuenta que $\pi_1 + \pi_2 = 1$ y $\lambda \pi_1 = \mu \pi_2$, se obtiene que $\pi_1 = \frac{\mu}{\lambda + \mu}$ y $\pi_2 = \frac{\lambda}{\lambda + \mu}$. Comparando el resultado obtenido con lo visto en el ejemplo anterior, la distribución límite es la misma que $P(t)$ cuando $t \to \infty$. En el siguiente apartado, vamos a ver las condiciones bajo las cuales este comportamiento se da en casos más generales.

1.4.1. Teorema ergódico

Sea $\{X_t\}_{t \geq 0}$ una cadena de Markov en tiempo continuo con espacio de estados S y matriz de tasas Q. Decimos que la cadena es *ergódica* si cumple las siguientes dos condiciones:

1. **Irreducibilidad:** para todo par de estados $i, j \in S$ existe $t > 0$ tal que $P_{ij}(t) > 0$.
2. **Recurrencia positiva:** todos los estados son recurrentes positivos; es decir, el tiempo medio de retorno al estado i, es finito para cada $i \in S$.

Theorem 1.4.1 (Teorema ergódico para cadenas de Markov en tiempo continuo)**.** *Sea $\{X_t\}_{t \geq 0}$ una cadena de Markov en tiempo continuo irreducible y recurrente positiva (o equivalentemente ergódica). Entonces:*

1. *Existe una única distribución estacionaria $\vec{\pi} = (\pi_j)$ tal que*

$$\vec{\pi} Q = 0, \qquad \sum_j \pi_j = 1.$$

2. Para todo par de estados i, j se cumple la convergencia en régimen estacionario:

$$\lim_{t\to\infty} P_{ij}(t) = \pi_j.$$

3. La convergencia es independiente del estado inicial.

Es decir, la distribución del proceso converge a la distribución estacionaria:

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

Este resultado es la versión en tiempo continuo del teorema ergódico para cadenas en tiempo discreto. No vamos a ver la demostración, pero puede encontrarse en textos clásicos de Procesos Estocásticos.

Observación 1.4.2. *Se puede demostrar que, en una cadena de Markov en tiempo continuo recurrente nula o transitoria, la distribución límite no existe. Un ejemplo de esto se da en los procesos de nacimiento puros, como en el proceso de Poisson, ya que en este caso todos los estados son transitorios (jamás se vuelve al mismo estado).*

Como consecuencia de esto, tenemos que no hace falta verificar la recurrencia de las cadenas de Markov. Vamos a resolver el sistema dado por $\vec{\pi}Q = 0$ y, en caso de que tenga solución positiva, se trata de la distribución límite, mientras que si no la tiene, no existe la distribución límite.

Observación 1.4.3. *El sistema de ecuaciones $\vec{\pi}Q = 0$ se llama las ecuaciones de balance. Vamos a verlas: $\forall j \in S$,*

$$\pi_j \sum_{k\neq j} q_{j,k} = \sum_{k\neq j} \pi_k q_{k,j},$$

donde la parte de la izquierda de igualdad en la expresión anterior indica la tasa de salida del estado j y la parte de la derecha la tasa de llegada al estado j.

1.4.2. Ejemplo: proceso de nacimiento y muerte

Consideremos un proceso de nacimiento y muerte con espacio de estados $S = \{0, 1, 2, \dots\}$ y tasas de nacimiento λ_n y muerte μ_n. El generador tiene la forma:

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \end{pmatrix}.$$

Si el proceso es irreducible y recurrente positivo, entonces existe una única distribución estacionaria $\vec{\pi} = (\pi_n)_{n\in S}$ que satisface:

$$\vec{\pi}Q = 0, \qquad \sum_{n=0}^{\infty} \pi_n = 1.$$

De las ecuaciones de balance se obtiene:

$$\pi_{n+1} = \pi_n \frac{\lambda_n}{\mu_{n+1}}, \qquad n \geq 0.$$

Por tanto, para todo $n \geq 1$

$$\pi_n = \pi_0 \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}}.$$

Observamos que, para todo $n \geq 1$,

$$\pi_n = \prod_{i=0}^{n} \frac{\lambda_k}{\mu_{k+1}} \pi_0.$$

Ahora, para calcular π_0, usamos que $\sum_{n=0}^{\infty} \pi_n = 1$ y así

$$1 = \sum_{n=0}^{\infty} \pi_n = \pi_0 \left(1 + \sum_{n=1}^{\infty} \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}} \right) \iff \pi_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}}}$$

Cuando esta suma es finita, la cadena es positiva recurrente y el teorema ergódico se aplica: para todo estado inicial i,

$$\lim_{t\to\infty} \mathbb{P}(X_t = n \mid X_0 = i) = \pi_n,$$

con

$$\pi_n = \frac{\prod_{i=0}^{n} \frac{\lambda_k}{\mu_{k+1}}}{1 + \sum_{n=1}^{\infty} \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}}}.$$

Así, la distribución límite coincide con la distribución estacionaria.

Aplicaremos este resultado al modelo de crecimiento lineal con inmigración visto anteriormente. Para este caso, tenemos que

$$\lambda_n = n\lambda + \theta, \quad \mu_n = n\mu,$$

para todo $n \geq 1$ y $\lambda_0 = \theta$. Tenemos que

$$\pi_n = \frac{\frac{\prod_{i=0}^{n}(i\lambda+\theta)}{k!\mu^k}}{1 + \sum_{j=1}^{\infty} \frac{\prod_{k=0}^{j-1}(k\lambda+\theta)}{k!\mu^k}}.$$

1.5. Cadenas Encajadas y Uniformizadas

En esta sección, vamos a ver que podemos transformar las cadenas de Markov en tiempo continuo a cadenas de Markov en tiempo discreto, pero que tenemos que tener cuidado sobre cómo lo hacemos.

1.5.1. Cadenas Encajadas

La cadena encajada de una cadena de Markov en tiempo continuo es una cadena de Markov en tiempo discreto en donde sólamente se tiene en cuenta los cambios de estado (y no el tiempo de permanencia en cada estado).

Como ejemplo, consideremos la cadena de Markov en tiempo contínuo cuya matriz de tasas de transición es:

$$Q = \begin{pmatrix} -2 & 2 & 0 \\ 0 & -2 & 2 \\ 1 & 1 & -2 \end{pmatrix}.$$

Resolviendo $\vec{\pi}Q = 0$, obtenemos que la distribución estacionaria de esta cadena de Markov es $(\frac{2}{5}, \frac{2}{5}, \frac{1}{5})$.

Para esta cadena de Markov, se tiene que los tiempos de permanencia son: $\nu_i = 2$, para $i = 1, 2, 3$. Por tanto, la cadena encajada tiene la siguiente matriz de transición (reemplazando ν_i por uno):

$$P = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$

Calculamos la distribución estacionaria de esta cadena de Markov en tiempo discreto, haciendo $\vec{\pi}P = \vec{\pi}$ y obtenemos $(\frac{2}{5}, \frac{2}{5}, \frac{1}{5})$. Es decir, la cadena encajada y la original tienen la misma distribución estacionaria en este caso. Por tanto, parece que la distribución estacionaria es siempre la misma y en la cadena de Markov en tiempo continuo y en su encajada. Vamos a ver otro ejemplo que demuestra que esto no es siempre cierto.

Consideremos la cadena de Markov en tiempo contínuo cuya matriz de tasas de transición es:

$$Q = \begin{pmatrix} -2 & 2 & 0 \\ 0 & -1 & 1 \\ 3/2 & 3/2 & -3 \end{pmatrix}.$$

Resolviendo $\vec{\pi}Q = 0$, obtenemos que la distribución estacionaria de esta cadena de Markov es $(\frac{3}{19}, \frac{12}{19}, \frac{4}{19})$.

En esta cadena de Markov, se tiene que los tiempos de permanencia son: $\nu_1 = 2$, $\nu_2 = 1$ y $\nu_3 = 3$. Por tanto, su cadena encajada tiene la siguiente matriz de transición

(reemplazando ν_i por uno en todo i):

$$\begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1/2 & 1/2 & 0 \end{pmatrix},$$

que es la misma cadena encajada que en el caso anterior, por tanto, su distribución estacionaria es $(\frac{2}{5}, \frac{2}{5}, \frac{1}{5})$, que es distinta a la de la cadena de Markov original. En conclusión, una cadena de Markov en tiempo continuo y su cadena encajada pueden no tener la misma distribución estacionaria.

Observación 1.5.1. *Estos dos ejemplos muestran que una misma cadena encajada puede provenir de dos cadenas de Markov en tiempo continuo cuyas distribuciones estacionarias son las mismas. Esto no ocurre con las cadenas uniformizadas.*

1.5.2. Cadenas uniformizadas

Sea $\{X_t\}_{t\geq 0}$ una cadena de Markov en tiempo continuo con matriz de tasa de transición $Q = (q_{i,j} \in S)$. Sea C una constante fija tal que $C \geq \text{máx}_{i\in S}\{|q_{i,i}|\}_{i\in S}$. Entonces, la cadena uniformizada de esta cadena de Markov en tiempo continuo es una cadena de Markov en tiempo discreto con la siguiente matriz de probabilidades

$$P = I + \frac{Q}{C},$$

donde I es la matriz identidad del tamaño $|S|$.

Consideremos, de nuevo, la cadena de Markov en tiempo contínuo cuya matriz de tasas de transición es:

$$Q = \begin{pmatrix} -2 & 2 & 0 \\ 0 & -1 & 1 \\ 3/2 & 3/2 & -3 \end{pmatrix}.$$

Definimos $C = 3 \geq \text{máx}\{1, 2, 3\}$. En este caso, la matriz de transición de la cadena uniformizada es

$$P = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 0 & 2/3 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$

Vamos a calcular la distribución estacionaria resolviendo $\vec{\pi}P = \vec{\pi}$ y obtenemos $(\frac{3}{19}, \frac{12}{19}, \frac{4}{19})$, que, según lo que hemos calculado anteriormente, coincide con la distribución estacionaria de la cadena de Markov en tiempo continuo.

De hecho, se puede demostrar que para cualquier cadena de Markov en tiempo contínuo y cualquier constante de uniformización C (que cumpla que sea mayor o igual que $|q_{i,i}|$), su distribución estacionaria y la de su cadena uniformizada coinciden.

2

Análisis de una Única Cola

La teoría de colas es una rama de la probabilidad y los procesos estocásticos que estudia sistemas en los que unidades (clientes, trabajos, tareas, paquetes, etc.) llegan para recibir un servicio, pudiendo formarse colas cuando los servidores están ocupados. Estos modelos aparecen en ámbitos tan diversos como telecomunicaciones, informática, transporte, ingeniería, logística o sanidad.

En en primer lugar, comenzaremos por introducir los elementos básicos de un sistema de colas y, después, presentamos la notación estándar utilizada para clasificarlos: la notación de Kendall.

2.1. Componentes fundamentales de un sistema de colas

Un *sistema de colas* está formado por los siguientes elementos:

2.1.1. Llegadas

Los clientes llegan al sistema según algún mecanismo aleatorio. La secuencia de tiempos de llegada puede modelarse mediante:

- un **proceso de Poisson** (caso más habitual),
- una distribución general para los tiempos entre llegadas,
- flujos deterministas o controlados.

Denotaremos por λ la tasa media de llegadas.

2.1.2. Servicio

Cada cliente requiere un tiempo aleatorio de servicio. El servidor (o los servidores) atiende a un cliente y, una vez terminado, pasa al siguiente según alguna disciplina preestablecida. Los tiempos de servicio pueden seguir:

- una distribución exponencial,
- una distribución general,
- un tiempo determinista.

Denotaremos por μ la tasa media de servicio.

2.1.3. Sala de espera o cola

Cuando un cliente llega y todos los servidores están ocupados, el cliente puede esperar en una *cola*. Las características de la sala de espera pueden incluir:

- **capacidad ilimitada**: la cola puede crecer sin restricción,
- **capacidad finita**: sólo pueden esperar un número máximo de clientes.

La disciplina de servicio más habitual es:

- **FIFO** (First-In First-Out): el primero que llega es el primero que se sirve (también conocido como FCFS: First-Come-First-Served),

aunque existen otras: LIFO, prioridades, servicio aleatorio, Processor sharing, etc.

2.1.4. Servidores

El sistema puede tener uno o varios servidores. En general, se puede modelar:

- un único servidor ($s = 1$),
- varios servidores en paralelo ($s > 1$),

2.2. Notación de Kendall

Para clasificar los modelos de colas se utiliza la *notación de Kendall*, de la forma:

$$A/B/s/K/N/D.$$

Los significados son los siguientes:

- A: distribución del **tiempo entre llegadas**.

- B: distribución del **tiempo de servicio**.
- s: número de **servidores**.
- K: capacidad total del sistema.
- D: disciplina de servicio.

Las letras más comunes para las distribuciones son:

- M: *Markovian*, es decir, distribución exponencial,
- D: determinista,
- G: distribución general.

Obsérvese que, en el caso de llegadas, M significa que la distribución del tiempo entre llegadas de dos clientes consecutivos sigue una distribución exponencial; es decir, las llegadas siguen un Proceso de Poisson.

En la mayor parte de los textos sólo se usan los tres primeros componentes: $A/B/s$ y, en ese caso, se supone que $K = \infty$ y D es FIFO.

Por ejemplo:

- $M/M/1$: proceso de llegadas de Poisson, tiempos de servicio exponenciales, un único servidor.
- $M/G/1$: llegadas Poisson, servicio general, un único servidor.
- $G/M/s$: tiempos entre llegadas generales, servicio exponencial, s servidores.

2.3. Ley De Little

La Ley de Little es uno de los resultados más importantes en teoría de colas y aparece en una gran cantidad de contextos, tanto estocásticos como deterministas.

Theorem 2.3.1 (Ley de Little)**.** *Sea un sistema de colas estable, con tasa media de llegadas λ, tiempo medio que un cliente pasa en el sistema T (espera + servicio) y número medio de clientes en el sistema L. Entonces,*

$$L = \lambda T.$$

El resultado es sorprendentemente general: no depende de la distribución de llegadas, de la disciplina de servicio, del número de servidores ni de la distribución de los tiempos de servicio. Lo único necesario es que el sistema sea estable (alcance un régimen estacionario) y que el tiempo medio de permanencia y de llegada existan.

A partir de la Ley de Little también obtenemos:

$$L_q = \lambda T_q,$$

donde L_q es el número medio de clientes en la cola (excluyendo servicio) y T_q es el tiempo medio de espera.

2.4. Propiedad PASTA

La propiedad PASTA (*Poisson Arrivals See Time Averages*) establece que cuando las llegadas siguen un proceso de Poisson, los clientes "ven" el sistema en el mismo estado que reflejan sus promedios temporales.

Theorem 2.4.1 (Propiedad PASTA)**.** *En un sistema de colas en régimen estacionario, si las llegadas forman un proceso de Poisson con tasa λ, entonces para cualquier evento A sobre el estado del sistema,*

$$\mathbb{P}(\textit{el cliente que llega observa } A) = \mathbb{P}(\textit{el sistema está en } A \textit{ en un instante arbitrario}).$$

Es decir, la probabilidad de que un cliente que llega vea, por ejemplo, "exactamente n clientes en el sistema", coincide con la probabilidad estacionaria de que el sistema tenga exactamente n clientes.

Una consecuencia inmediata es que, si X es el número de clientes en el sistema en régimen estacionario,

$$\mathbb{P}(\text{el cliente que llega ve } X = n) = \pi_n,$$

donde (π_n) es la distribución estacionaria.

La propiedad PASTA no se cumple en general si las llegadas no son Poisson.

A continuación, comenzaremos el estudio riguroso del modelo más básico e importante, la cola $M/M/1$.

2.5. El modelo $M/M/1$

El modelo $M/M/1$ es el sistema de colas más básico y uno de los más importantes por su papel fundamental en teoría de colas. Consta de:

- llegadas según un proceso de Poisson con tasa λ;
- tiempos de servicio independientes y exponenciales con media $1/\mu$;
- un único servidor ($s = 1$);
- disciplina FIFO y capacidad ilimitada.

Sea X_t el número de clientes en el sistema (cola + servicio) en el instante t. Entonces $\{X_t\}_{t\geq 0}$ es un *proceso de nacimiento y muerte* con:

$$\lambda_n = \lambda, \qquad \mu_n = \mu \quad (n \geq 1), \qquad \mu_0 = 0.$$

Es decir, las tasas de nacimiento y muerte no dependen del estado n (excepto que en 0 no hay muertes posibles).

2.5.1. Ecuaciones de balance y distribución estacionaria

En régimen estacionario, la distribución $\vec{\pi} = (\pi_n)$ debe satisfacer:

$$\vec{\pi} Q = 0, \qquad \sum_{n=0}^{\infty} \pi_n = 1,$$

donde Q es la matriz de tasas de transición:

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ 0 & \mu & -(\lambda+\mu) & \lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix}.$$

De las ecuaciones de balance y de $\vec{\pi} Q = 0$, obtenemos:

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

Por tanto,

$$\pi_{n+1} = \frac{\lambda}{\mu}\,\pi_n = \rho\,\pi_n, \qquad \text{donde} \quad \rho := \frac{\lambda}{\mu}.$$

Iterando:

$$\pi_n = \rho^n \pi_0, \qquad n \geq 0.$$

Para normalizar usamos:

$$1 = \sum_{n=0}^{\infty} \pi_n = \pi_0 \sum_{n=0}^{\infty} \rho^n.$$

La serie converge si y sólo si $\rho < 1$ (condición de estabilidad), y obtenemos:

$$\sum_{n=0}^{\infty} \rho^n = \frac{1}{1-\rho}.$$

Entonces, si $\rho < 1$:

$$\pi_0(1-\rho)^{-1} = 1 \quad \Longrightarrow \quad \pi_0 = 1-\rho.$$

Por tanto, $\rho < 1$:

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

Esto muestra que la distribución estacionaria del número de clientes en el sistema es una *distribución geométrica* de parámetro $1-\rho$. Es decir, si N es el número de clientes en la cola (en servicio y esperando) en régimen estacionario, $N \sim Geo(1-\rho)$. En consecuencia, el número medio de clientes (que antes lo hemos denotado por L) es $\mathbb{E}[N]$ y, como N es geométrica, tenemos que, si $\rho < 1$,

$$L := \sum_{n=0}^{\infty} n\,\pi_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = (1-\rho)\frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho}.$$

Es decir, si $\rho < 1$, $L = \frac{\rho}{1-\rho}$.

Por la Ley de Little, $L = \lambda T$, donde T es el tiempo medio total que un cliente pasa en el sistema (espera + servicio). Por tanto:

$$T = \frac{L}{\lambda} = \frac{\rho}{(1-\rho)\lambda} = \frac{1}{\mu-\lambda}.$$

Observación 2.5.1. *El tiempo medio de espera en la cola, T_q, se obtiene restando el tiempo medio de servicio:*

$$T_q = T - \frac{1}{\mu} = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu-\lambda}.$$

2.5.2. Condición de estabilidad

Recordemos que el modelo $M/M/1$ tiene tasa de llegadas λ y tasa de servicio μ. El parámetro fundamental del sistema es

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

que se conoce como la carga del sistema o el factor de utilización.

Decimos que el sistema $M/M/1$ es *estable* si su distribución estacionaria existe, es decir, si la cadena de Markov subyacente es recurrente positiva. Esto ocurre si y sólo si

$$\rho < 1.$$

Observación 2.5.2. *La condición $\rho < 1$ garantiza que el servidor, en promedio, es capaz de procesar más trabajo del que llega al sistema. En ese caso la cadena de Markov*

es recurrente positiva y existe una distribución estacionaria

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

Cuando $\rho \geq 1$, la serie $\sum_{n=0}^{\infty} \rho^n$ diverge y por tanto no es posible normalizarla. Esto refleja que, con más trabajo llegando que siendo servido, la cola crecerá indefinidamente y no alcanzará equilibrio.

Observación 2.5.3. *La cantidad ρ representa la fracción de tiempo que el servidor está ocupado. En el régimen estacionario del $M/M/1$:*

$$\mathbb{P}(\textit{servidor ocupado}) = \rho.$$

Si $\rho \geq 1$, el servidor necesitaría estar ocupado más del 100 % del tiempo, lo cual es imposible; consecuentemente, el sistema se congestiona y la cola tiende a infinito casi seguro.

Observación 2.5.4. *Como*

$$L = \mathbb{E}[X] = \frac{\rho}{1-\rho},$$

se observa que L crece sin límite cuando $\rho \to 1^-$. Esto tiene una interpretación clara:

- *Si el servidor está muy ocupado ($\rho \approx 1$), el sistema funciona cerca de la saturación.*
- *Un pequeño incremento en la carga produce un aumento desproporcionado en el número esperado de clientes.*

Aplicando la Ley de Little:

$$W = \frac{L}{\lambda} = \frac{1}{\mu - \lambda}.$$

Como $W_q = W - 1/\mu$, cuando $\rho \to 1^-$, tanto W como W_q explotan, reflejando la mala calidad de servicio cuando un sistema se encuentra saturado.

2.5.3. Comparación de sistemas

A continuación, vamos a comparar el rendimiento de varios sistemas basados en la cola M/M/1. En todos los casos vamos a suponer que $\lambda < \mu$ para garantizar estabilidad de la cola.

En primer lugar, vamos a escalar los parámetros, es decir, vamos a multiplicar tanto la tasa de llegada como de servicio por el mismo factor k y vamos a ver qué ocurre con el rendimiento del sistema resultante. Por un lado, se ve que ρ, que es la carga del sistema, no varía cuando escalamos los parámetros. Lo mismo ocurre con el tiempo medio de clientes L, ya que depende únicamente de ρ en la cola M/M/1. Para acabar, vemos que para el tiempo medio de permanencia se divide por k. Sea W el tiempo medio de

permanencia del sistema original y W' es del sistema con parámetros escalados. Se tiene que

$$W' = \frac{1}{k\mu - k\lambda} = \frac{1}{k}\frac{1}{\mu - \lambda} = \frac{W}{k}.$$

En segundo lugar, vamos a comparar el tiempo medio de permanencia de una cola M/M/1 con parámetros λ y μ con m colas M/M/1 con tasa de llegada λ/m y tasa de servicio μ/m. Sea W el tiempo medio de permanencia del sistema original y W' es del sistema con m colas en paralelo. Se tiene que

$$W' = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\mu/m - \lambda/m} = \frac{m}{\mu - \lambda} = mW.$$

Es decir, en el sistema de colas en paralelo, el tiempo medio de permanencia es m veces superior al sistema original.

2.6. El modelo $M/M/1/K$

El modelo $M/M/1/K$ es una variante de la cola $M/M/1$ en la que la capacidad total del sistema es finita. En concreto, el sistema consta de:

- llegadas según un proceso de Poisson con tasa λ;
- tiempos de servicio independientes y exponenciales con media $1/\mu$;
- un único servidor;
- disciplina FIFO;
- capacidad máxima de K clientes en el sistema, incluyendo tanto al cliente que está siendo atendido como a los que están esperando.

Por tanto, como máximo puede haber un cliente en servicio y $K-1$ clientes esperando en la cola. Si una llegada encuentra K clientes en el sistema, no puede entrar y se pierde. Por este motivo, el modelo $M/M/1/K$ también se denomina *sistema con pérdidas* o *cola con capacidad finita*.

Sea X_t el número de clientes presentes en el sistema en el instante t. Entonces $\{X_t\}_{t\geq 0}$ es un proceso de nacimiento y muerte con espacio de estados finito

$$S = \{0, 1, \ldots, K\}.$$

Las tasas de nacimiento y muerte son

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

La matriz generadora es

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ \mu & -(\lambda+\mu) & \lambda & \ddots & \vdots \\ 0 & \mu & -(\lambda+\mu) & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \lambda \\ 0 & \cdots & 0 & \mu & -\mu \end{pmatrix}.$$

Definimos, como en la cola $M/M/1$, $\rho = \frac{\lambda}{\mu}$.

2.6.1. Distribución estacionaria

Como el espacio de estados es finito y la cadena es irreducible, el sistema posee una única distribución estacionaria para cualesquiera $\lambda > 0$ y $\mu > 0$.

Cuando $\rho \neq 1$, la distribución estacionaria viene dada por

$$\pi_n = \frac{1-\rho}{1-\rho^{K+1}}\rho^n, \qquad 0 \leq n \leq K.$$

Cuando $\rho = 1$, la distribución estacionaria es uniforme:

$$\pi_n = \frac{1}{K+1}, \qquad 0 \leq n \leq K.$$

A diferencia de lo que ocurre en la cola $M/M/1$, no es necesario imponer la condición $\rho < 1$ para que exista una distribución estacionaria. Incluso si $\lambda \geq \mu$, el número de clientes no puede superar K, por lo que la cola no puede crecer indefinidamente.

No obstante, cuando ρ es grande, el sistema pasa una proporción elevada del tiempo completamente lleno y, por tanto, se pierde una fracción importante de las llegadas.

2.6.2. Probabilidad de bloqueo

Una llegada es rechazada si encuentra K clientes en el sistema. Como las llegadas forman un proceso de Poisson, la propiedad PASTA permite identificar la probabilidad de bloqueo con la probabilidad estacionaria de que el sistema esté lleno. Por tanto,

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

Para $\rho \neq 1$,

$$P_{\text{bloqueo}} = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}},$$

mientras que, si $\rho = 1$,

$$P_{\text{bloqueo}} = \frac{1}{K+1}.$$

Sea N el número de clientes presentes en el sistema en régimen estacionario. El número medio de clientes es

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

Usando la Ley de Little, así como $T_q = T - \frac{1}{\mu}$, se pueden obtener el resto de medidas de rendimiento vistas anteriormente.

Para terminar, vamos a ver que cuando $K \to \infty$ y $\rho < 1$, las expresiones anteriores convergen a las correspondientes al modelo $M/M/1$:

$$\pi_n \longrightarrow (1-\rho)\rho^n, \qquad L \longrightarrow \frac{\rho}{1-\rho},$$

$$L_q \longrightarrow \frac{\rho^2}{1-\rho}, \qquad T \longrightarrow \frac{1}{\mu - \lambda}.$$

2.7. Los modelos $M/M/K$ y $M/M/K/K$

En esta sección estudiamos dos sistemas con K servidores en paralelo. En ambos modelos, las llegadas siguen un proceso de Poisson con tasa λ, los tiempos de servicio son independientes y exponenciales con media $1/\mu$, y cada servidor atiende como máximo a un cliente.

La diferencia fundamental entre ambos modelos es la capacidad de espera:

- en el modelo $M/M/K$, los clientes que encuentran todos los servidores ocupados esperan en una cola de capacidad ilimitada;
- en el modelo $M/M/K/K$, no existe sala de espera y los clientes que encuentran todos los servidores ocupados son rechazados.

En ambos casos, introducimos la cantidad $a := \frac{\lambda}{\mu}$.

2.7.1. El modelo $M/M/K$

El modelo $M/M/K$ consta de:

- llegadas según un proceso de Poisson con tasa λ;
- tiempos de servicio independientes y exponenciales con media $1/\mu$;
- K servidores idénticos trabajando en paralelo;
- disciplina FIFO;
- capacidad ilimitada del sistema.

Si un cliente llega y hay algún servidor libre, comienza inmediatamente su servicio. Si todos los servidores están ocupados, el cliente se incorpora a la cola y espera hasta que alguno quede disponible.

Sea X_t el número de clientes presentes en el sistema en el instante t. Entonces $\{X_t\}_{t\geq 0}$ es un proceso de nacimiento y muerte con espacio de estados

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

Las tasas de nacimiento y muerte son

$$\lambda_n = \lambda, \qquad n \geq 0,$$

y

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

Cuando hay $n < K$ clientes en el sistema, todos ellos están recibiendo servicio y la tasa total de finalización es $n\mu$. Cuando hay al menos K clientes, los K servidores están ocupados y la tasa total de servicio es $K\mu$.

La matriz generadora tiene la forma

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & \cdots \\ 0 & 0 & 3\mu & -(\lambda+3\mu) & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix},$$

donde, a partir del estado K, las tasas de muerte son iguales a $K\mu$.

Condición de estabilidad

Definimos la intensidad de tráfico por servidor como

$$\rho := \frac{\lambda}{K\mu}.$$

El modelo $M/M/K$ posee una distribución estacionaria si y sólo si $\rho < 1$. Esta condición indica que la tasa total máxima de servicio debe ser mayor que la tasa de llegadas.

Distribución estacionaria

Si $\rho < 1$, la distribución estacionaria viene dada por

$$\pi_n = \begin{cases} \pi_0 \dfrac{a^n}{n!}, & 0 \leq n < K, \\ \pi_0 \dfrac{a^n}{K!\, K^{n-K}}, & n \geq K. \end{cases}$$

Equivalentemente, para $n \geq K$,

$$\pi_n = \pi_0 \frac{a^K}{K!} \rho^{n-K}.$$

La probabilidad de que el sistema esté vacío es

$$\pi_0 = \left[\sum_{n=0}^{K-1} \frac{a^n}{n!} + \frac{a^K}{K!} \frac{1}{1-\rho} \right]^{-1}.$$

Probabilidad de espera: fórmula de Erlang C

Un cliente que llega debe esperar si encuentra ocupados los K servidores, es decir, si encuentra al menos K clientes en el sistema. Debido a la propiedad PASTA, la probabilidad de espera viene dada por

$$P_{\text{espera}} = \mathbb{P}(X \geq K) = \sum_{n=K}^{\infty} \pi_n.$$

Por tanto,

$$P_{\text{espera}} = \frac{\dfrac{a^K}{K!} \dfrac{1}{1-\rho}}{\displaystyle\sum_{n=0}^{K-1} \frac{a^n}{n!} + \frac{a^K}{K!} \frac{1}{1-\rho}}.$$

Esta expresión se conoce como *fórmula de Erlang C*.

2.7.2. El modelo $M/M/K/K$

El modelo $M/M/K/K$, también denominado *sistema de pérdidas de Erlang*, consta de:

- llegadas según un proceso de Poisson con tasa λ;
- tiempos de servicio independientes y exponenciales con media $1/\mu$;
- K servidores idénticos trabajando en paralelo;
- capacidad total igual a K;
- ausencia de sala de espera.

Si una llegada encuentra algún servidor libre, comienza inmediatamente su servicio. Si encuentra los K servidores ocupados, no puede entrar en el sistema y se pierde.

Sea X_t el número de clientes presentes en el sistema en el instante t. Entonces $\{X_t\}_{t\geq 0}$ es un proceso de nacimiento y muerte con espacio de estados finito

$$S = \{0, 1, \ldots, K\}.$$

Las tasas de nacimiento son

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

mientras que las tasas de muerte son

$$\mu_n = n\mu, \qquad 1 \leq n \leq K.$$

La matriz generadora es

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ \mu & -(\lambda+\mu) & \lambda & \ddots & \vdots \\ 0 & 2\mu & -(\lambda+2\mu) & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \lambda \\ 0 & \cdots & 0 & K\mu & -K\mu \end{pmatrix}.$$

Distribución estacionaria

Como el espacio de estados es finito y la cadena es irreducible, existe una única distribución estacionaria para cualesquiera $\lambda > 0$ y $\mu > 0$. No es necesario imponer ninguna condición de estabilidad adicional.

Utilizando la carga ofrecida

$$a = \frac{\lambda}{\mu},$$

la distribución estacionaria viene dada por

$$\pi_n = \frac{\dfrac{a^n}{n!}}{\displaystyle\sum_{j=0}^{K} \frac{a^j}{j!}}, \qquad 0 \leq n \leq K.$$

Por tanto,

$$\pi_0 = \left(\sum_{j=0}^{K} \frac{a^j}{j!} \right)^{-1}.$$

La distribución estacionaria puede interpretarse como una distribución de Poisson de parámetro a, condicionada a tomar valores en $\{0, 1, \ldots, K\}$.

Probabilidad de bloqueo: fórmula de Erlang B

Una llegada es rechazada si encuentra los K servidores ocupados. Por la propiedad PASTA, la probabilidad de bloqueo coincide con la probabilidad estacionaria de que el sistema esté lleno:

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

En consecuencia,

$$P_{\text{bloqueo}} = B(K, a) := \frac{\dfrac{a^K}{K!}}{\displaystyle\sum_{j=0}^{K} \frac{a^j}{j!}}.$$

Esta expresión se conoce como *fórmula de Erlang B*.

2.7.3. Aplicación al diseño de un call center

Consideremos un call center con K agentes. Supondremos que:

- las llamadas llegan según un proceso de Poisson con tasa λ;
- la duración de cada llamada es exponencial con media $1/\mu$;
- los agentes trabajan de manera independiente;
- cada agente puede atender una única llamada;
- si todos los agentes están ocupados, la llamada entrante se bloquea y abandona el sistema;
- las llamadas bloqueadas no vuelven a intentarlo.

Bajo estas hipótesis, el call center se modeliza mediante una cola $M/M/K/K$. La probabilidad de que una llamada sea bloqueada es

$$P_{\text{bloqueo}} = B(K, a), \qquad a = \frac{\lambda}{\mu}.$$

El objetivo es encontrar el menor número de agentes K tal que

$$B(K, a) < 0{,}02.$$

Es decir, se desea que menos del 2 % de las llamadas sean rechazadas.

Primer caso: $\lambda = 5$ **y** $\mu = 1$

En este caso, la carga ofrecida es

$$a = \frac{\lambda}{\mu} = 5.$$

La siguiente tabla muestra cómo varía la probabilidad de bloqueo al aumentar el número de agentes.

Número de agentes K	$B(K, 5)$	Porcentaje de bloqueo
1	0.833333	83,3333 %
2	0.675676	67,5676 %
3	0.529661	52,9661 %
4	0.398343	39,8343 %
5	0.284868	28,4868 %
6	0.191847	19,1847 %
7	0.120519	12,0519 %
8	0.070048	7,0048 %
9	0.037458	3,7458 %
10	0.018385	1,8385 %

Con 9 agentes, la probabilidad de bloqueo es

$$B(9, 5) \approx 0{,}037458 > 0{,}02.$$

Con 10 agentes, se obtiene

$$B(10, 5) \approx 0{,}018385 < 0{,}02.$$

Por tanto, el número mínimo de agentes necesario es

$$\boxed{K_{\text{mín}} = 10}.$$

Segundo caso: $\lambda = 20$ **y** $\mu = 1$

En este caso, la carga ofrecida es

$$a = \frac{\lambda}{\mu} = 20.$$

La evolución de la probabilidad de bloqueo para diferentes números de agentes es la siguiente:

Número de agentes K	$B(K, 20)$	Porcentaje de bloqueo
15	0.329997	32,9997 %
16	0.292033	29,2033 %
17	0.255714	25,5714 %
18	0.221260	22,1260 %
19	0.188908	18,8908 %
20	0.158892	15,8892 %
21	0.131436	13,1436 %
22	0.106734	10,6734 %
23	0.084930	8,4930 %
24	0.066097	6,6097 %
25	0.050222	5,0222 %
26	0.037195	3,7195 %
27	0.026813	2,6813 %
28	0.018792	1,8792 %

Con 27 agentes, la probabilidad de bloqueo es

$$B(27, 20) \approx 0{,}026813 > 0{,}02.$$

Con 28 agentes, obtenemos

$$B(28, 20) \approx 0{,}018792 < 0{,}02.$$

Por tanto, el número mínimo de agentes necesario es

$$\boxed{K_{\text{mín}} = 28}.$$

2.8. El modelo $M/M/\infty$

El modelo $M/M/\infty$ representa un sistema con un número ilimitado de servidores. Consta de:

- llegadas según un proceso de Poisson con tasa λ;
- tiempos de servicio independientes y exponenciales con media $1/\mu$;
- infinitos servidores disponibles;
- ausencia de cola de espera.

Como siempre hay un servidor disponible, todo cliente que llega comienza inmediatamente su servicio. Por tanto, en este modelo no se producen esperas ni rechazos.

Sea X_t el número de clientes presentes en el sistema en el instante t. Entonces $\{X_t\}_{t\geq 0}$ es un proceso de nacimiento y muerte con espacio de estados

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

Las tasas de nacimiento son

$$\lambda_n = \lambda, \qquad n \geq 0,$$

mientras que las tasas de muerte son

$$\mu_n = n\mu, \qquad n \geq 1.$$

En efecto, cuando hay n clientes en el sistema, cada uno está siendo atendido por un servidor distinto. Como los tiempos de servicio son exponenciales de parámetro μ, la tasa total de finalización de servicios es $n\mu$.

La matriz generadora tiene la forma

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & \cdots \\ 0 & 0 & 3\mu & -(\lambda+3\mu) & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}.$$

2.8.1. Distribución estacionaria

Introducimos, como en los modelos anteriores, la carga ofrecida

$$a := \frac{\lambda}{\mu}.$$

El proceso $M/M/\infty$ posee una única distribución estacionaria para cualesquiera $\lambda > 0$ y $\mu > 0$. No es necesario imponer una condición de estabilidad como $\lambda < \mu$, puesto que el número de servidores disponibles es ilimitado y la capacidad total de servicio aumenta con el número de clientes presentes.

La distribución estacionaria viene dada por

$$\pi_n = e^{-a}\frac{a^n}{n!}, \qquad n \geq 0.$$

Por tanto, el número de clientes presentes en el sistema en régimen estacionario sigue una distribución de Poisson de parámetro a:

$$X \sim \text{Pois}(a) = \text{Pois}\left(\frac{\lambda}{\mu}\right).$$

En particular, la probabilidad de que el sistema esté vacío es

$$\pi_0 = e^{-a} = e^{-\lambda/\mu}.$$

La probabilidad de que haya exactamente n clientes recibiendo servicio es

$$\mathbb{P}(X = n) = e^{-\lambda/\mu}\frac{(\lambda/\mu)^n}{n!}.$$

Además,

$$\mathbb{E}[X] = \frac{\lambda}{\mu}$$

y

$$\mathrm{Var}(X) = \frac{\lambda}{\mu}.$$

Así, se ha visto que tanto la media como la varianza del número de clientes presentes en el sistema coinciden y son iguales a a.

3

Análisis de Redes de Colas

En el capítulo anterior hemos estudiado sistemas formados por una única cola. Sin embargo, en muchas aplicaciones un cliente debe recibir varios servicios sucesivos antes de abandonar el sistema. En estos casos, resulta natural considerar una *red de colas*, es decir, un conjunto de estaciones de servicio conectadas entre sí.

Comenzaremos analizando la estructura más sencilla: una red de colas en *tándem*. En ella, todos los clientes visitan las estaciones en un orden fijo: tras recibir servicio en la primera cola pasan a la segunda, después a la tercera, y así sucesivamente, hasta abandonar el sistema.

3.1. Colas en tándem

Consideremos una red formada por n colas $M/M/1$ conectadas en serie. Los clientes llegan desde el exterior a la primera estación según un proceso de Poisson con tasa λ. Después de ser atendidos en la estación k, pasan inmediatamente a la estación $k+1$. Tras completar su servicio en la estación n, abandonan la red.

La estructura del sistema puede representarse mediante

$$\text{Llegadas} \xrightarrow{\lambda} \boxed{\text{Cola 1}} \longrightarrow \boxed{\text{Cola 2}} \longrightarrow \cdots \longrightarrow \boxed{\text{Cola } n} \longrightarrow \text{Salida.}$$

Supondremos que:

- las llegadas externas forman un proceso de Poisson con tasa λ;
- la estación k tiene un único servidor;

- los tiempos de servicio en la estación k son independientes y exponenciales con parámetro μ_k;
- los tiempos de servicio son independientes entre sí y del proceso de llegadas;
- cada estación utiliza una disciplina FIFO;
- todas las colas tienen capacidad ilimitada;
- todos los clientes visitan las estaciones en el orden $1, 2, \ldots, n$.

Sea

$$X_k(t) = \text{número de clientes en la estación } k \text{ en el instante } t,$$

incluyendo tanto al cliente que está siendo atendido como a los que están esperando. El estado de la red en el instante t es el vector

$$\boldsymbol{X}(t) = \big(X_1(t), X_2(t), \ldots, X_n(t)\big),$$

cuyo espacio de estados es

$$S = \mathbb{N}_0^n.$$

Debido a la hipótesis de tiempos de servicio exponenciales, el proceso

$$\{\boldsymbol{X}(t)\}_{t\geq 0}$$

es una cadena de Markov en tiempo continuo.

3.1.1. Transiciones de la cadena

Denotemos por $\boldsymbol{e}_k$ el k-ésimo vector de la base canónica de $\mathbb{R}^n$. Desde un estado

$$\boldsymbol{x} = (x_1, \ldots, x_n),$$

pueden producirse las siguientes transiciones:

- una llegada externa a la primera estación:

$$\boldsymbol{x} \longrightarrow \boldsymbol{x} + \boldsymbol{e}_1 \qquad \text{con tasa } \lambda;$$

- una finalización de servicio en la estación k, para $1 \leq k < n$:

$$\boldsymbol{x} \longrightarrow \boldsymbol{x} - \boldsymbol{e}_k + \boldsymbol{e}_{k+1} \qquad \text{con tasa } \mu_k,$$

siempre que $x_k > 0$;

- una finalización de servicio en la última estación:

$$\boldsymbol{x} \longrightarrow \boldsymbol{x} - \boldsymbol{e}_n \qquad \text{con tasa } \mu_n,$$

siempre que $x_n > 0$.

Como todos los clientes recorren todas las estaciones, la tasa de tráfico que llega a cada una de ellas es λ. Por tanto, definimos la intensidad de tráfico de la estación k mediante

$$\rho_k := \frac{\lambda}{\mu_k}, \qquad k = 1, \ldots, n.$$

3.1.2. Condición de estabilidad

La red de colas en tándem posee una distribución estacionaria si y sólo si todas las estaciones son estables, es decir, si

$$\rho_k < 1, \qquad k = 1, \ldots, n.$$

Equivalentemente,

$$\lambda < \mu_k, \qquad k = 1, \ldots, n,$$

o bien

$$\lambda < \min_{1 \le k \le n} \mu_k.$$

La estación con menor tasa de servicio constituye el *cuello de botella* de la red. Si alguna estación satisface $\lambda \ge \mu_k$, dicha estación no puede procesar, en promedio, todo el tráfico que recibe y el número de clientes en ella no admite una distribución estacionaria.

3.1.3. Distribución estacionaria

Si

$$\rho_k = \frac{\lambda}{\mu_k} < 1, \qquad k = 1, \ldots, n,$$

la distribución estacionaria de la red tiene forma producto:

$$\boxed{\pi(x_1, \ldots, x_n) = \prod_{k=1}^{n} (1 - \rho_k)\rho_k^{x_k}, \qquad (x_1, \ldots, x_n) \in \mathbb{N}_0^n.}$$

Por tanto, en régimen estacionario, las variables

$$X_1, X_2, \ldots, X_n$$

son independientes y cada X_k sigue una distribución geométrica:

$$\mathbb{P}(X_k = m) = (1-\rho_k)\rho_k^m, \qquad m \in \mathbb{N}_0.$$

Es decir,

$$X_k \sim \text{Geo}(1-\rho_k), \qquad k = 1, \dots, n,$$

donde utilizamos la convención de que la distribución geométrica tiene soporte $\mathbb{N}_0$. La normalización de la distribución se obtiene de

$$\begin{aligned}\sum_{(x_1,\dots,x_n)\in\mathbb{N}_0^n} \pi(x_1,\dots,x_n) &= \prod_{k=1}^{n}\left(\sum_{x_k=0}^{\infty}(1-\rho_k)\rho_k^{x_k}\right)\\ &= 1.\end{aligned}$$

La forma producto es un resultado notable: aunque los movimientos de los clientes conectan las distintas estaciones y los procesos $\{X_k(t)\}_{t\geq 0}$ no son independientes, los números de clientes en las diferentes estaciones sí son independientes en un instante estacionario.

3.1.4. Medidas de rendimiento

El número medio de clientes en la estación k es

$$L_k = \mathbb{E}[X_k] = \frac{\rho_k}{1-\rho_k} = \frac{\lambda}{\mu_k - \lambda}.$$

El número medio total de clientes en la red es

$$L = \mathbb{E}\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n}\frac{\rho_k}{1-\rho_k}.$$

Como todos los clientes visitan todas las estaciones y la tasa de entrada a la red es λ, la Ley de Little proporciona

$$L = \lambda T,$$

donde T es el tiempo medio total que un cliente permanece en la red. En consecuencia,

$$T = \frac{L}{\lambda} = \sum_{k=1}^{n}\frac{1}{\mu_k - \lambda}.$$

El tiempo medio que un cliente pasa en la estación k, incluyendo espera y servicio, es

$$T_k = \frac{1}{\mu_k - \lambda},$$

por lo que

$$T = \sum_{k=1}^{n} T_k.$$

3.1.5. Ejemplo

Consideremos ahora el caso particular de dos colas en tándem. Los clientes llegan a la primera estación con tasa λ, reciben servicio con tasa μ_1, pasan a la segunda estación y, después de recibir servicio con tasa μ_2, abandonan el sistema.

El estado de la red es

$$(i, j) \in \mathbb{N}_0^2,$$

donde:

- i es el número de clientes en la primera estación;
- j es el número de clientes en la segunda estación.

Las posibles transiciones son

$$(i, j) \longrightarrow (i + 1, j) \qquad \text{con tasa } \lambda,$$

$$(i, j) \longrightarrow (i - 1, j + 1) \qquad \text{con tasa } \mu_1, \quad \text{si } i > 0,$$

y

$$(i, j) \longrightarrow (i, j - 1) \qquad \text{con tasa } \mu_2, \quad \text{si } j > 0.$$

Supondremos que

$$\rho_1 = \frac{\lambda}{\mu_1} < 1, \qquad \rho_2 = \frac{\lambda}{\mu_2} < 1.$$

La distribución estacionaria propuesta es

$$\pi_{i,j} = (1 - \rho_1)(1 - \rho_2)\rho_1^i \rho_2^j, \qquad i, j \geq 0.$$

Comprobaremos que esta distribución satisface las ecuaciones de balance global.

Estados interiores: $i > 0$**,** $j > 0$

Desde el estado (i, j) se puede salir por:

- una llegada externa, con tasa λ;
- una finalización de servicio en la primera estación, con tasa μ_1;
- una finalización de servicio en la segunda estación, con tasa μ_2.

El flujo total de salida es, por tanto,

$$\pi_{i,j}(\lambda + \mu_1 + \mu_2).$$

Se puede entrar en el estado (i, j) desde:

- el estado $(i-1, j)$, mediante una llegada externa;
- el estado $(i+1, j-1)$, mediante una finalización de servicio en la primera estación;
- el estado $(i, j+1)$, mediante una finalización de servicio en la segunda estación.

La ecuación de balance global es

$$\pi_{i,j}(\lambda + \mu_1 + \mu_2) = \lambda\pi_{i-1,j} + \mu_1\pi_{i+1,j-1} + \mu_2\pi_{i,j+1}.$$

Dividiendo entre $\pi_{i,j} > 0$, obtenemos

$$\lambda + \mu_1 + \mu_2 = \lambda\frac{\pi_{i-1,j}}{\pi_{i,j}} + \mu_1\frac{\pi_{i+1,j-1}}{\pi_{i,j}} + \mu_2\frac{\pi_{i,j+1}}{\pi_{i,j}}.$$

De la forma producto,

$$\frac{\pi_{i-1,j}}{\pi_{i,j}} = \frac{1}{\rho_1},$$

$$\frac{\pi_{i+1,j-1}}{\pi_{i,j}} = \frac{\rho_1}{\rho_2},$$

y

$$\frac{\pi_{i,j+1}}{\pi_{i,j}} = \rho_2.$$

Así, el miembro derecho es

$$\frac{\lambda}{\rho_1} + \mu_1\frac{\rho_1}{\rho_2} + \mu_2\rho_2.$$

Como

$$\rho_1 = \frac{\lambda}{\mu_1}, \qquad \rho_2 = \frac{\lambda}{\mu_2},$$

se tiene

$$\frac{\lambda}{\rho_1} = \mu_1,$$

$$\mu_1 \frac{\rho_1}{\rho_2} = \mu_1 \frac{\lambda/\mu_1}{\lambda/\mu_2} = \mu_2,$$

y

$$\mu_2 \rho_2 = \lambda.$$

Por tanto,

$$\lambda \frac{\pi_{i-1,j}}{\pi_{i,j}} + \mu_1 \frac{\pi_{i+1,j-1}}{\pi_{i,j}} + \mu_2 \frac{\pi_{i,j+1}}{\pi_{i,j}} = \mu_1 + \mu_2 + \lambda,$$

y queda verificada la ecuación de balance para $i > 0$ y $j > 0$.

Frontera $i = 0$**,** $j > 0$

Cuando la primera estación está vacía, desde $(0, j)$ solamente pueden producirse:

- una llegada externa, con tasa λ;
- una finalización de servicio en la segunda estación, con tasa μ_2.

La ecuación de balance es

$$\pi_{0,j}(\lambda + \mu_2) = \mu_1 \pi_{1,j-1} + \mu_2 \pi_{0,j+1}.$$

Dividiendo entre $\pi_{0,j}$, el miembro derecho se convierte en

$$\mu_1 \frac{\rho_1}{\rho_2} + \mu_2 \rho_2 = \mu_2 + \lambda.$$

Por tanto,

$$\pi_{0,j}(\lambda + \mu_2) = \mu_1 \pi_{1,j-1} + \mu_2 \pi_{0,j+1}.$$

Frontera $i > 0$**,** $j = 0$

Cuando la segunda estación está vacía, desde $(i, 0)$ pueden producirse:

- una llegada externa, con tasa λ;
- una finalización de servicio en la primera estación, con tasa μ_1.

La ecuación de balance es

$$\pi_{i,0}(\lambda + \mu_1) = \lambda \pi_{i-1,0} + \mu_2 \pi_{i,1}.$$

En este caso, no existe un término procedente de una finalización de servicio en la primera estación, pues para llegar a $(i,0)$ de esta forma sería necesario partir de un estado $(i+1,-1)$, que no pertenece al espacio de estados.

Dividiendo entre $\pi_{i,0}$, el miembro derecho es

$$\frac{\lambda}{\rho_1} + \mu_2\rho_2 = \mu_1 + \lambda.$$

Por tanto, también se satisface la ecuación de balance:

$$\pi_{i,0}(\lambda+\mu_1) = \lambda\pi_{i-1,0} + \mu_2\pi_{i,1}.$$

Estado $(0,0)$

Finalmente, consideramos el estado en el que ambas estaciones están vacías. Desde $(0,0)$ solamente puede producirse una llegada externa, con tasa λ. Por otra parte, únicamente se puede entrar en $(0,0)$ desde $(0,1)$, mediante una finalización de servicio en la segunda estación.

La ecuación de balance es

$$\lambda\pi_{0,0} = \mu_2\pi_{0,1}.$$

Como

$$\pi_{0,1} = \rho_2\pi_{0,0},$$

se obtiene

$$\mu_2\pi_{0,1} = \mu_2\rho_2\pi_{0,0} = \lambda\pi_{0,0}.$$

Por tanto, la ecuación de balance también se verifica en el estado $(0,0)$.

Hemos comprobado que

$$\pi_{i,j} = (1-\rho_1)(1-\rho_2)\rho_1^i\rho_2^j$$

satisface todas las ecuaciones de balance global. Además,

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\pi_{i,j} = (1-\rho_1)(1-\rho_2)\left(\sum_{i=0}^{\infty}\rho_1^i\right)\left(\sum_{j=0}^{\infty}\rho_2^j\right) = 1.$$

Por consiguiente, es la distribución estacionaria de la red de dos colas en tándem.

3.2. Redes acíclicas de colas

Las colas en tándem constituyen un caso particular de una clase más amplia de redes abiertas: las *redes acíclicas de colas*. En estas redes, después de completar el servicio

en una estación, un cliente puede dirigirse a una de varias estaciones o abandonar el sistema, pero no puede regresar a una estación que haya visitado anteriormente.

La ausencia de ciclos permite ordenar las estaciones de forma que los clientes se desplacen siempre desde estaciones con índice menor hacia estaciones con índice mayor.

3.2.1. Descripción general del modelo

Consideremos una red formada por n estaciones de servicio. Supondremos que:

- a la estación i llegan clientes desde el exterior según un proceso de Poisson con tasa α_i;
- cada estación tiene un único servidor;
- los tiempos de servicio en la estación i son independientes y exponenciales con parámetro μ_i;
- los tiempos de servicio son independientes entre sí y de los procesos de llegadas externas;
- las colas tienen capacidad ilimitada y la disciplina de servicio es FIFO;
- después de completar el servicio en la estación i, un cliente se dirige a la estación j con probabilidad $p_{i,j}$, o abandona la red con probabilidad

$$p_{i,0} = 1 - \sum_{j=1}^{n} p_{i,j}.$$

Las decisiones de encaminamiento de los clientes se realizan de manera independiente.

Como la red es acíclica, podemos numerar las estaciones de modo que

$$p_{i,j} = 0 \qquad \text{si } j \leq i.$$

Es decir, un cliente solamente puede desplazarse desde una estación i hacia una estación j con $j > i$, o abandonar la red.

Sea

$$X_i(t) = \text{número de clientes en la estación } i \text{ en el instante } t,$$

incluyendo tanto al cliente que está siendo atendido como a los que están esperando. El estado de la red es

$$\boldsymbol{X}(t) = \big(X_1(t), \ldots, X_n(t)\big) \in \mathbb{N}_0^n.$$

Debido a la hipótesis de tiempos de servicio exponenciales, el proceso

$$\{\boldsymbol{X}(t)\}_{t\geq 0}$$

es una cadena de Markov en tiempo continuo.

3.2.2. Ecuaciones de tráfico

La tasa total de llegadas a una estación incluye tanto las llegadas externas como las procedentes de otras estaciones de la red.

Denotemos por λ_i la *tasa total de llegadas* a la estación i, las cuales satisfacen las ecuaciones de tráfico

$$\boxed{\lambda_i = \alpha_i + \sum_{j=1}^{n} \lambda_j p_{j,i}, \qquad i = 1, \dots, n.}$$

El término α_i representa las llegadas externas a la estación i, mientras que $\lambda_j p_{j,i}$ representa la tasa de clientes que, después de completar su servicio en la estación j, se dirigen a la estación i.

Como la red es acíclica y las estaciones están ordenadas de modo que $p_{j,i} = 0$ cuando $j \geq i$.

Definimos la intensidad de tráfico de la estación i mediante

$$\rho_i := \frac{\lambda_i}{\mu_i}.$$

La red posee una distribución estacionaria si todas las estaciones son estables, es decir, si

$$\boxed{\rho_i < 1, \qquad i = 1, \dots, n.}$$

3.2.3. Distribución estacionaria

Si

$$\rho_i = \frac{\lambda_i}{\mu_i} < 1, \qquad i = 1, \dots, n,$$

la distribución estacionaria de la red tiene forma producto:

$$\boxed{\pi(x_1, \dots, x_n) = \prod_{i=1}^{n} (1 - \rho_i)\rho_i^{x_i}, \qquad (x_1, \dots, x_n) \in \mathbb{N}_0^n.}$$

La normalización se obtiene inmediatamente:

$$\sum_{(x_1,\ldots,x_n)\in\mathbb{N}_0^n} \pi(x_1,\ldots,x_n) = \prod_{i=1}^{n}\left(\sum_{x_i=0}^{\infty}(1-\rho_i)\rho_i^{x_i}\right)$$
$$= 1.$$

En consecuencia, en régimen estacionario, las variables

$$X_1,\ldots,X_n$$

son independientes y

$$X_i \sim \text{Geo}(1-\rho_i), \qquad i=1,\ldots,n,$$

donde utilizamos la convención

$$\mathbb{P}(X_i = k) = (1-\rho_i)\rho_i^k, \qquad k\in\mathbb{N}_0.$$

3.2.4. Medidas de rendimiento

El número medio de clientes en la estación i es

$$L_i = \mathbb{E}[X_i] = \frac{\rho_i}{1-\rho_i} = \frac{\lambda_i}{\mu_i-\lambda_i}.$$

El número medio total de clientes en la red es

$$L = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n}\frac{\rho_i}{1-\rho_i}.$$

El resto de las medidas de rendimiento se pueden obtener usando la Ley de Little.

3.2.5. Ejemplo: una red acíclica de dos estaciones

Consideremos una red formada por dos estaciones. A la primera estación llegan clientes desde el exterior con tasa α_1, y a la segunda llegan clientes desde el exterior con tasa α_2.

Después de completar su servicio en la primera estación:

- el cliente pasa a la segunda estación con probabilidad p;
- el cliente abandona la red con probabilidad $1-p$.

Después de completar su servicio en la segunda estación, el cliente abandona siempre la red.

Los tiempos de servicio en las estaciones 1 y 2 son exponenciales con parámetros μ_1 y μ_2, respectivamente.

Las ecuaciones de tráfico son

$$\lambda_1 = \alpha_1$$

y

$$\lambda_2 = \alpha_2 + p\lambda_1 = \alpha_2 + p\alpha_1.$$

Definimos

$$\rho_1 = \frac{\alpha_1}{\mu_1}, \qquad \rho_2 = \frac{\alpha_2 + p\alpha_1}{\mu_2}.$$

La condición de estabilidad es

$$\rho_1 < 1, \qquad \rho_2 < 1.$$

Bajo estas condiciones, la distribución estacionaria propuesta es

$$\boxed{\pi_{i,j} = (1-\rho_1)(1-\rho_2)\rho_1^i \rho_2^j, \qquad i,j \geq 0.}$$

A continuación, comprobaremos que esta expresión satisface las ecuaciones de balance global.

Transiciones de la cadena

El estado del sistema es

$$(i,j) \in \mathbb{N}_0^2,$$

donde i y j son los números de clientes presentes en las estaciones 1 y 2, respectivamente.

Las posibles transiciones son:

$$(i,j) \longrightarrow (i+1,j) \qquad \text{con tasa } \alpha_1,$$

debida a una llegada externa a la primera estación;

$$(i,j) \longrightarrow (i,j+1) \qquad \text{con tasa } \alpha_2,$$

debida a una llegada externa a la segunda estación;

$$(i,j) \longrightarrow (i-1,j+1) \qquad \text{con tasa } p\mu_1, \quad i > 0,$$

cuando un cliente completa el servicio en la primera estación y se dirige a la segunda;

$$(i,j) \longrightarrow (i-1,j) \qquad \text{con tasa } (1-p)\mu_1, \quad i > 0,$$

cuando un cliente completa el servicio en la primera estación y abandona la red; y

$$(i,j) \longrightarrow (i,j-1) \qquad \text{con tasa } \mu_2, \quad j>0,$$

cuando un cliente completa el servicio en la segunda estación y abandona la red.

Estados interiores: $i>0$, $j>0$

Desde un estado interior (i,j) pueden producirse dos llegadas externas o una finalización de servicio en cualquiera de las dos estaciones. Por tanto, la tasa total de salida del estado es

$$\alpha_1+\alpha_2+\mu_1+\mu_2.$$

La ecuación de balance global es

$$\begin{aligned}\pi_{i,j}(\alpha_1+\alpha_2+\mu_1+\mu_2) = {}& \alpha_1\pi_{i-1,j}+\alpha_2\pi_{i,j-1}\\ &+p\mu_1\pi_{i+1,j-1}+(1-p)\mu_1\pi_{i+1,j}\\ &+\mu_2\pi_{i,j+1}.\end{aligned}$$

De la forma producto se deduce que

$$\frac{\pi_{i-1,j}}{\pi_{i,j}}=\frac{1}{\rho_1}, \qquad \frac{\pi_{i,j-1}}{\pi_{i,j}}=\frac{1}{\rho_2},$$

$$\frac{\pi_{i+1,j-1}}{\pi_{i,j}}=\frac{\rho_1}{\rho_2}, \qquad \frac{\pi_{i+1,j}}{\pi_{i,j}}=\rho_1, \qquad \frac{\pi_{i,j+1}}{\pi_{i,j}}=\rho_2.$$

Dividiendo la ecuación de balance entre $\pi_{i,j}$, el miembro derecho queda

$$\frac{\alpha_1}{\rho_1}+\frac{\alpha_2}{\rho_2}+p\mu_1\frac{\rho_1}{\rho_2}+(1-p)\mu_1\rho_1+\mu_2\rho_2.$$

Como

$$\mu_1\rho_1=\alpha_1$$

y

$$\mu_2\rho_2=\alpha_2+p\alpha_1,$$

los dos términos que contienen $1/\rho_2$ pueden agruparse:

$$\frac{\alpha_2}{\rho_2}+p\mu_1\frac{\rho_1}{\rho_2}=\frac{\alpha_2+p\alpha_1}{\rho_2}=\mu_2.$$

Además,

$$\frac{\alpha_1}{\rho_1}=\mu_1,$$

$$(1-p)\mu_1\rho_1 = (1-p)\alpha_1$$

y

$$\mu_2\rho_2 = \alpha_2 + p\alpha_1.$$

Por tanto, el miembro derecho es

$$\mu_1 + \mu_2 + (1-p)\alpha_1 + \alpha_2 + p\alpha_1 = \mu_1 + \mu_2 + \alpha_1 + \alpha_2,$$

que coincide con la tasa total de salida. La ecuación de balance queda verificada para $i > 0$ y $j > 0$.

Frontera $i = 0$, $j > 0$

Cuando la primera estación está vacía, no puede completarse ningún servicio en ella. La ecuación de balance es

$$\begin{aligned}\pi_{0,j}(\alpha_1 + \alpha_2 + \mu_2) = \alpha_2\pi_{0,j-1} + p\mu_1\pi_{1,j-1} \\ + (1-p)\mu_1\pi_{1,j} + \mu_2\pi_{0,j+1}.\end{aligned}$$

Dividiendo entre $\pi_{0,j}$, el miembro derecho es

$$\frac{\alpha_2}{\rho_2} + p\mu_1\frac{\rho_1}{\rho_2} + (1-p)\mu_1\rho_1 + \mu_2\rho_2.$$

Los dos primeros términos suman

$$\frac{\alpha_2 + p\alpha_1}{\rho_2} = \mu_2.$$

Los dos términos restantes suman

$$(1-p)\alpha_1 + \alpha_2 + p\alpha_1 = \alpha_1 + \alpha_2.$$

Por consiguiente,

$$\frac{\alpha_2}{\rho_2} + p\mu_1\frac{\rho_1}{\rho_2} + (1-p)\mu_1\rho_1 + \mu_2\rho_2 = \mu_2 + \alpha_1 + \alpha_2,$$

y se verifica la ecuación de balance en esta frontera.

Frontera $i > 0$, $j = 0$

Cuando la segunda estación está vacía, no puede producirse una salida desde ella. La ecuación de balance es

$$\pi_{i,0}(\alpha_1 + \alpha_2 + \mu_1) = \alpha_1\pi_{i-1,0} + (1-p)\mu_1\pi_{i+1,0} + \mu_2\pi_{i,1}.$$

Obsérvese que una finalización de servicio en la primera estación que envía al cliente a la segunda no puede conducir a un estado con segunda coordenada igual a cero.

Dividiendo entre $\pi_{i,0}$, el miembro derecho es

$$\frac{\alpha_1}{\rho_1} + (1-p)\mu_1\rho_1 + \mu_2\rho_2.$$

Utilizando las definiciones de ρ_1 y ρ_2,

$$\begin{aligned}\frac{\alpha_1}{\rho_1} + (1-p)\mu_1\rho_1 + \mu_2\rho_2 &= \mu_1 + (1-p)\alpha_1 + \alpha_2 + p\alpha_1 \\ &= \mu_1 + \alpha_1 + \alpha_2.\end{aligned}$$

Por tanto, la ecuación de balance también se satisface para $i > 0$ y $j = 0$.

Estado $(0,0)$

Desde el estado $(0,0)$ solamente pueden producirse llegadas externas. La tasa total de salida es

$$\alpha_1 + \alpha_2.$$

Se puede entrar en $(0,0)$:

- desde $(1,0)$, cuando un cliente finaliza su servicio en la primera estación y abandona la red;
- desde $(0,1)$, cuando un cliente finaliza su servicio en la segunda estación.

La ecuación de balance es

$$\pi_{0,0}(\alpha_1 + \alpha_2) = (1-p)\mu_1\pi_{1,0} + \mu_2\pi_{0,1}.$$

Como

$$\pi_{1,0} = \rho_1\pi_{0,0}, \qquad \pi_{0,1} = \rho_2\pi_{0,0},$$

el miembro derecho es

$$\begin{aligned}(1-p)\mu_1\pi_{1,0} + \mu_2\pi_{0,1} &= \left[(1-p)\mu_1\rho_1 + \mu_2\rho_2\right]\pi_{0,0} \\ &= \left[(1-p)\alpha_1 + \alpha_2 + p\alpha_1\right]\pi_{0,0} \\ &= (\alpha_1 + \alpha_2)\pi_{0,0}.\end{aligned}$$

Por tanto, la ecuación de balance se satisface también en el origen.

Finalmente, la distribución propuesta está normalizada, pues

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\pi_{i,j} = (1-\rho_1)(1-\rho_2)\left(\sum_{i=0}^{\infty}\rho_1^i\right)\left(\sum_{j=0}^{\infty}\rho_2^j\right)$$
$$= 1.$$

En consecuencia,

$$\pi_{i,j} = (1-\rho_1)(1-\rho_2)\rho_1^i\rho_2^j, \qquad i,j \geq 0,$$

es la distribución estacionaria de esta red acíclica.

Observación 3.2.1. *Este ejemplo no es una red en tándem, pues no todos los clientes siguen el mismo recorrido. Algunos clientes llegan directamente a la segunda estación y otros abandonan la red después de completar el servicio en la primera. No obstante, la distribución estacionaria sigue teniendo forma producto.*

3.3. Redes abiertas de Jackson

Las redes acíclicas estudiadas en la sección anterior pueden generalizarse permitiendo que los clientes regresen a estaciones que ya han visitado. De este modo, pueden aparecer ciclos en el recorrido de los clientes.

Una *red abierta de Jackson* está formada por varias estaciones de servicio con llegadas externas, encaminamiento probabilístico entre estaciones y salidas hacia el exterior. El término *abierta* indica que los clientes pueden entrar desde el exterior y, después de visitar una o varias estaciones, abandonar definitivamente la red.

3.3.1. Descripción general del modelo

Consideremos una red formada por n estaciones. Supondremos que:

- a la estación i llegan clientes desde el exterior según un proceso de Poisson con tasa α_i;
- los procesos de llegadas externas a las distintas estaciones son independientes;
- la estación i tiene un único servidor;
- los tiempos de servicio en la estación i son independientes y exponenciales con parámetro μ_i;
- todas las colas tienen capacidad ilimitada y disciplina de servicio FIFO;
- después de completar su servicio en la estación i, un cliente se dirige a la estación j con probabilidad $p_{i,j}$;

- después de completar su servicio en la estación i, un cliente abandona la red con probabilidad
$$p_{i,0} = 1 - \sum_{j=1}^{n} p_{i,j};$$
- las decisiones de encaminamiento son independientes entre sí, de los tiempos de servicio y de los procesos de llegadas externas.

Definimos la matriz

$$P = (p_{ij})_{1 \leq i,j \leq n}$$

se denomina *matriz de enrutamiento* de la red.
A diferencia de lo que sucede en una red acíclica, ahora puede ocurrir que

$$p_{ij} > 0 \qquad \text{y} \qquad p_{ji} > 0.$$

También es posible que un cliente regrese indirectamente a una estación que ya había visitado. Por ejemplo, puede seguir un recorrido de la forma

$$1 \longrightarrow 2 \longrightarrow 3 \longrightarrow 1.$$

Supondremos que la red es abierta, es decir, que todo cliente termina abandonando la red con probabilidad uno.
Sea

$$X_i(t) = \text{número de clientes en la estación } i \text{ en el instante } t,$$

incluyendo al cliente que está recibiendo servicio. El estado de la red en el instante t es

$$\boldsymbol{X}(t) = \big(X_1(t), \ldots, X_n(t)\big),$$

y el espacio de estados es

$$S = \mathbb{N}_0^n.$$

Debido a la hipótesis de tiempos de servicio exponenciales, el proceso

$$\{\boldsymbol{X}(t)\}_{t \geq 0}$$

es una cadena de Markov en tiempo continuo.

3.3.2. Transiciones de la cadena

Denotemos por $\boldsymbol{e}_i$ el i-ésimo vector de la base canónica de $\mathbb{R}^n$. Desde un estado

$$\boldsymbol{x} = (x_1, \ldots, x_n),$$

pueden producirse las siguientes transiciones:

- una llegada externa a la estación i:

$$\boldsymbol{x} \longrightarrow \boldsymbol{x} + \boldsymbol{e}_i \qquad \text{con tasa } \alpha_i;$$

- una finalización de servicio en la estación i, seguida de un traslado a la estación j:

$$\boldsymbol{x} \longrightarrow \boldsymbol{x} - \boldsymbol{e}_i + \boldsymbol{e}_j \qquad \text{con tasa } \mu_i p_{i,j},$$

  siempre que $x_i > 0$;

- una finalización de servicio en la estación i, seguida de la salida del cliente de la red:

$$\boldsymbol{x} \longrightarrow \boldsymbol{x} - \boldsymbol{e}_i \qquad \text{con tasa } \mu_i p_{i,0},$$

  siempre que $x_i > 0$.

Si se permite $p_{i,i} > 0$, una finalización de servicio seguida de un nuevo servicio en la misma estación no modifica el número de clientes y no produce un cambio de estado. Por simplicidad, supondremos en lo que sigue que

$$p_{i,i} = 0, \qquad i = 1, \ldots, n.$$

Denotemos por λ_i la tasa total de llegadas a la estación i. Esta tasa incluye tanto las llegadas externas como las procedentes de otras estaciones.

Las tasas totales de llegada satisfacen las *ecuaciones de tráfico*

$$\boxed{\lambda_i = \alpha_i + \sum_{j=1}^{n} \lambda_j p_{ji}, \qquad i = 1, \ldots, n.}$$

Definimos la intensidad de tráfico de la estación i mediante

$$\rho_i := \frac{\lambda_i}{\mu_i}.$$

La red de Jackson posee una distribución estacionaria si todas sus estaciones son estables, es decir, si

$$\boxed{\rho_i < 1, \qquad i = 1, \ldots, n.}$$

Equivalentemente,

$$\lambda_i < \mu_i, \qquad i = 1, \ldots, n.$$

3.3.3. Teorema de Jackson

Theorem 3.3.1 (Teorema de Jackson)**.** *Consideremos una red abierta de n estaciones $M/M/1$ que satisface las hipótesis anteriores. Sean $\lambda_1, \ldots, \lambda_n$ las soluciones de las ecuaciones de tráfico*

$$\lambda_i = \alpha_i + \sum_{j=1}^{n} \lambda_j p_{ji}, \qquad i = 1, \ldots, n.$$

Si

$$\rho_i = \frac{\lambda_i}{\mu_i} < 1, \qquad i = 1, \ldots, n,$$

entonces la red posee una única distribución estacionaria, dada por

$$\boxed{\pi(x_1, \ldots, x_n) = \prod_{i=1}^{n} (1 - \rho_i)\rho_i^{x_i}, \qquad (x_1, \ldots, x_n) \in \mathbb{N}_0^n.}$$

En consecuencia, en régimen estacionario, las variables aleatorias

$$X_1, \ldots, X_n$$

son independientes y

$$X_i \sim \text{Geo}(1 - \rho_i), \qquad i = 1, \ldots, n,$$

donde la distribución geométrica tiene soporte $\mathbb{N}_0$:

$$\mathbb{P}(X_i = k) = (1 - \rho_i)\rho_i^k, \qquad k \geq 0.$$

La independencia es una propiedad estacionaria. Los procesos

$$\{X_i(t)\}_{t \geq 0}, \qquad i = 1, \ldots, n,$$

no son independientes, pues una salida de una estación puede producir simultáneamente una llegada a otra estación.

3.3.4. Medidas de rendimiento

El número medio de clientes en la estación i es

$$L_i = \mathbb{E}[X_i] = \frac{\rho_i}{1 - \rho_i} = \frac{\lambda_i}{\mu_i - \lambda_i}.$$

El número medio total de clientes en la red es

$$L = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} \frac{\rho_i}{1-\rho_i}.$$

3.3.5. Ejemplo: una red de Jackson de dos estaciones

Consideremos una red formada por dos estaciones. Los clientes llegan desde el exterior únicamente a la primera estación, según un proceso de Poisson con tasa α.

Después de completar el servicio en la primera estación:

- el cliente pasa a la segunda estación con probabilidad p;
- abandona la red con probabilidad $1-p$.

Después de completar el servicio en la segunda estación:

- el cliente regresa a la primera estación con probabilidad q;
- abandona la red con probabilidad $1-q$.

Supondremos

$$0 \leq p, q < 1.$$

Esta red no es acíclica si $p > 0$ y $q > 0$, pues un cliente puede seguir el ciclo

$$1 \longrightarrow 2 \longrightarrow 1.$$

Ecuaciones de tráfico

Las tasas efectivas de llegadas satisfacen

$$\lambda_1 = \alpha + q\lambda_2$$

y

$$\lambda_2 = p\lambda_1.$$

Sustituyendo la segunda ecuación en la primera,

$$\lambda_1 = \alpha + pq\lambda_1.$$

Por tanto,

$$\boxed{\lambda_1 = \frac{\alpha}{1-pq}}$$

y

$$\boxed{\lambda_2 = \frac{p\alpha}{1-pq}.}$$

Las intensidades de tráfico son

$$\rho_1 = \frac{\alpha}{\mu_1(1-pq)}$$

y

$$\rho_2 = \frac{p\alpha}{\mu_2(1-pq)}.$$

La condición de estabilidad es

$$\rho_1 < 1, \qquad \rho_2 < 1.$$

Bajo estas condiciones, la distribución estacionaria propuesta por el teorema de Jackson es

$$\boxed{\pi_{i,j} = (1-\rho_1)(1-\rho_2)\rho_1^i \rho_2^j, \qquad i,j \geq 0.}$$

Comprobaremos que esta expresión satisface las ecuaciones de balance global.

Transiciones de la cadena

El estado del sistema es

$$(i,j) \in \mathbb{N}_0^2,$$

donde i y j son los números de clientes en las estaciones 1 y 2, respectivamente. Las posibles transiciones son:

$$(i,j) \longrightarrow (i+1,j) \qquad \text{con tasa } \alpha,$$

debida a una llegada externa;

$$(i,j) \longrightarrow (i-1,j+1) \qquad \text{con tasa } p\mu_1, \quad i > 0,$$

cuando un cliente pasa de la primera estación a la segunda;

$$(i,j) \longrightarrow (i-1,j) \qquad \text{con tasa } (1-p)\mu_1, \quad i > 0,$$

cuando un cliente abandona la red desde la primera estación;

$$(i,j) \longrightarrow (i+1,j-1) \qquad \text{con tasa } q\mu_2, \quad j > 0,$$

cuando un cliente regresa de la segunda estación a la primera; y

$$(i,j) \longrightarrow (i,j-1) \qquad \text{con tasa } (1-q)\mu_2, \quad j>0,$$

cuando un cliente abandona la red desde la segunda estación.
Las ecuaciones de tráfico proporcionan las identidades

$$\mu_1\rho_1 = \lambda_1 = \alpha + q\lambda_2$$

y

$$\mu_2\rho_2 = \lambda_2 = p\lambda_1.$$

También se verifica la identidad de conservación del flujo externo:

$$\boxed{\alpha = (1-p)\lambda_1 + (1-q)\lambda_2.}$$

En efecto,

$$\begin{aligned}(1-p)\lambda_1 + (1-q)\lambda_2 &= (1-p)\lambda_1 + (1-q)p\lambda_1 \\ &= (1-pq)\lambda_1 \\ &= \alpha.\end{aligned}$$

Estados interiores: $i>0$**,** $j>0$

Desde un estado interior (i,j) puede producirse una llegada externa o una finalización de servicio en cualquiera de las dos estaciones. La tasa total de salida es

$$\alpha + \mu_1 + \mu_2.$$

La ecuación de balance global es

$$\begin{aligned}\pi_{i,j}(\alpha+\mu_1+\mu_2) = {} & \alpha\pi_{i-1,j} + (1-p)\mu_1\pi_{i+1,j} \\ & + p\mu_1\pi_{i+1,j-1} + (1-q)\mu_2\pi_{i,j+1} \\ & + q\mu_2\pi_{i-1,j+1}.\end{aligned}$$

De la forma producto se deduce que

$$\frac{\pi_{i-1,j}}{\pi_{i,j}} = \frac{1}{\rho_1}, \qquad \frac{\pi_{i+1,j}}{\pi_{i,j}} = \rho_1,$$

$$\frac{\pi_{i+1,j-1}}{\pi_{i,j}} = \frac{\rho_1}{\rho_2}, \qquad \frac{\pi_{i,j+1}}{\pi_{i,j}} = \rho_2,$$

y

$$\frac{\pi_{i-1,j+1}}{\pi_{i,j}} = \frac{\rho_2}{\rho_1}.$$

Dividiendo entre $\pi_{i,j}$, el miembro derecho queda

$$\frac{\alpha}{\rho_1} + (1-p)\mu_1\rho_1 + p\mu_1\frac{\rho_1}{\rho_2} + (1-q)\mu_2\rho_2 + q\mu_2\frac{\rho_2}{\rho_1}.$$

Agrupando el primer y el último término,

$$\frac{\alpha}{\rho_1} + q\mu_2\frac{\rho_2}{\rho_1} = \frac{\alpha + q\lambda_2}{\rho_1} = \frac{\lambda_1}{\rho_1} = \mu_1.$$

Además,

$$p\mu_1\frac{\rho_1}{\rho_2} = \frac{p\lambda_1}{\rho_2} = \frac{\lambda_2}{\rho_2} = \mu_2.$$

Finalmente,

$$(1-p)\mu_1\rho_1 + (1-q)\mu_2\rho_2 = (1-p)\lambda_1 + (1-q)\lambda_2 = \alpha.$$

Por tanto, el miembro derecho es

$$\mu_1 + \mu_2 + \alpha,$$

que coincide con la tasa total de salida. La ecuación de balance se satisface para todos los estados interiores.

Frontera $i = 0$**,** $j > 0$

Cuando la primera estación está vacía, no puede completarse ningún servicio en ella. Desde el estado $(0, j)$ pueden producirse una llegada externa o una finalización de servicio en la segunda estación. La tasa total de salida es

$$\alpha + \mu_2.$$

La ecuación de balance global es

$$\begin{aligned}\pi_{0,j}(\alpha + \mu_2) = (1-p)\mu_1\pi_{1,j} + p\mu_1\pi_{1,j-1} \\ + (1-q)\mu_2\pi_{0,j+1}.\end{aligned}$$

No aparece un término asociado a un retorno desde la segunda estación, pues una transición de la segunda a la primera siempre incrementa la primera coordenada.

Dividiendo entre $\pi_{0,j}$, el miembro derecho es

$$(1-p)\mu_1\rho_1 + p\mu_1\frac{\rho_1}{\rho_2} + (1-q)\mu_2\rho_2.$$

Utilizando las ecuaciones de tráfico,

$$p\mu_1\frac{\rho_1}{\rho_2} = \mu_2,$$

mientras que

$$(1-p)\mu_1\rho_1 + (1-q)\mu_2\rho_2 = \alpha.$$

Por tanto,

$$(1-p)\mu_1\rho_1 + p\mu_1\frac{\rho_1}{\rho_2} + (1-q)\mu_2\rho_2 = \alpha + \mu_2.$$

La ecuación de balance queda verificada en esta frontera.

Frontera $i > 0$**,** $j = 0$

Cuando la segunda estación está vacía, desde $(i, 0)$ puede producirse una llegada externa o una finalización de servicio en la primera estación. La tasa total de salida es

$$\alpha + \mu_1.$$

La ecuación de balance es

$$\begin{aligned}\pi_{i,0}(\alpha+\mu_1) = \alpha\pi_{i-1,0} + (1-p)\mu_1\pi_{i+1,0} \\ + (1-q)\mu_2\pi_{i,1} + q\mu_2\pi_{i-1,1}.\end{aligned}$$

Dividiendo entre $\pi_{i,0}$, el miembro derecho es

$$\frac{\alpha}{\rho_1} + (1-p)\mu_1\rho_1 + (1-q)\mu_2\rho_2 + q\mu_2\frac{\rho_2}{\rho_1}.$$

Agrupando el primer y el último término,

$$\frac{\alpha}{\rho_1} + q\mu_2\frac{\rho_2}{\rho_1} = \frac{\alpha + q\lambda_2}{\rho_1} = \mu_1.$$

Los dos términos restantes suman

$$(1-p)\lambda_1 + (1-q)\lambda_2 = \alpha.$$

Por tanto, el miembro derecho es

$$\mu_1 + \alpha,$$

y se verifica la ecuación de balance para $i > 0$ y $j = 0$.

Estado $(0, 0)$

Desde el estado $(0, 0)$ solamente puede producirse una llegada externa a la primera estación. La tasa total de salida es, por tanto, α.

Se puede entrar en $(0, 0)$:

- desde $(1,0)$, si un cliente completa su servicio en la primera estación y abandona la red;

- desde $(0,1)$, si un cliente completa su servicio en la segunda estación y abandona la red.

La ecuación de balance es

$$\alpha\pi_{0,0} = (1-p)\mu_1\pi_{1,0} + (1-q)\mu_2\pi_{0,1}.$$

Como

$$\pi_{1,0} = \rho_1\pi_{0,0}$$

y

$$\pi_{0,1} = \rho_2\pi_{0,0},$$

el miembro derecho es

$$\begin{aligned}(1-p)\mu_1\pi_{1,0} + (1-q)\mu_2\pi_{0,1} &= \big[(1-p)\lambda_1 + (1-q)\lambda_2\big]\pi_{0,0} \\ &= \alpha\pi_{0,0}.\end{aligned}$$

Por tanto, la ecuación de balance también se satisface en el origen.
Finalmente,

$$\begin{aligned}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\pi_{i,j} &= (1-\rho_1)(1-\rho_2)\left(\sum_{i=0}^{\infty}\rho_1^i\right)\left(\sum_{j=0}^{\infty}\rho_2^j\right) \\ &= 1.\end{aligned}$$

En consecuencia,

$$\pi_{i,j} = (1-\rho_1)(1-\rho_2)\rho_1^i\rho_2^j, \qquad i,j \geq 0,$$

es la distribución estacionaria de esta red cíclica de dos estaciones.

Observación 3.3.1. *Cuando $q = 0$, los clientes no pueden regresar de la segunda estación a la primera. En ese caso, la red se vuelve acíclica y las ecuaciones de tráfico se reducen a*

$$\lambda_1 = \alpha, \qquad \lambda_2 = p\alpha.$$

Por tanto, las redes acíclicas son un caso particular de las redes abiertas de Jackson.

3.4. Redes cerradas de colas

En las redes estudiadas anteriormente, los clientes llegan desde el exterior y, después de visitar una o varias estaciones, abandonan el sistema. Este tipo de redes se denominan *redes abiertas*.

En una *red cerrada*, por el contrario, no existen llegadas ni salidas externas. Un número fijo de clientes circula indefinidamente entre las estaciones de la red. Cuando un cliente termina su servicio en una estación, se dirige inmediatamente a otra estación según un mecanismo probabilístico de enrutamiento.

Las redes cerradas aparecen, por ejemplo, en:

- sistemas informáticos con un número fijo de procesos o trabajos;
- sistemas de fabricación con un número constante de piezas;
- redes de transporte con una flota fija de vehículos;
- sistemas de reparación con un número fijo de máquinas;
- modelos de circulación de recursos entre diferentes estaciones.

3.4.1. Descripción general del modelo

Consideremos una red cerrada formada por n estaciones y un número fijo N de clientes. Supondremos que:

- no existen llegadas desde el exterior;
- ningún cliente abandona la red;
- la estación i tiene un único servidor;
- los tiempos de servicio en la estación i son independientes y exponenciales con parámetro μ_i;
- las colas tienen capacidad ilimitada y disciplina de servicio FIFO;
- después de completar su servicio en la estación i, un cliente se dirige a la estación j con probabilidad $p_{i,j}$;
- las decisiones de enrutamiento son independientes entre sí y de los tiempos de servicio.

La matriz

$$P = (p_{ij})_{1 \leq i,j \leq n}$$

es una matriz estocástica (lo cual es verdad para las redes de Jackson), es decir,

$$p_{ij} \geq 0, \qquad \sum_{j=1}^{n} p_{ij} = 1, \qquad i = 1, \ldots, n.$$

Sea

$$X_i(t) = \text{número de clientes en la estación } i \text{ en el instante } t,$$

incluyendo al cliente que está recibiendo servicio. El estado de la red es

$$\boldsymbol{X}(t) = \big(X_1(t), \ldots, X_n(t)\big).$$

Como el número total de clientes es constante, se verifica

$$\sum_{i=1}^{n} X_i(t) = N$$

para todo $t \geq 0$. Por tanto, el espacio de estados es

$$S = \left\{ \boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{N}_0^n : \sum_{i=1}^{n} x_i = N \right\}.$$

Debido a la hipótesis de tiempos de servicio exponenciales, el proceso

$$\{\boldsymbol{X}(t)\}_{t \geq 0}$$

es una cadena de Markov en tiempo continuo.

3.4.2. Transiciones de la cadena

Denotemos por $\boldsymbol{e}_i$ el i-ésimo vector de la base canónica. Desde un estado

$$\boldsymbol{x} = (x_1, \ldots, x_n) \in E_N,$$

si $x_i > 0$, puede completarse un servicio en la estación i. Si el cliente se dirige a la estación j, el nuevo estado es

$$\boldsymbol{x} - \boldsymbol{e}_i + \boldsymbol{e}_j.$$

Por tanto, las tasas de transición son

$$q\big(\boldsymbol{x}, \boldsymbol{x} - \boldsymbol{e}_i + \boldsymbol{e}_j\big) = \mu_i p_{ij}, \qquad x_i > 0.$$

Si $p_{ii} > 0$, un cliente puede completar un servicio y comenzar otro en la misma estación. Este suceso no modifica el vector que representa el número de clientes en cada estación. Por simplicidad, podemos suponer que

$$p_{ii} = 0, \qquad i = 1, \dots, n.$$

3.4.3. Razones de visita

En una red cerrada no existen tasas de llegadas externas. Por tanto, las tasas absolutas de tráfico no quedan determinadas de la misma forma que en una red abierta.

En su lugar, se introducen las *razones de visita* $v_1, \dots, v_n$, que satisfacen las ecuaciones

$$\boxed{v_i = \sum_{j=1}^{n} v_j p_{ji}, \qquad i = 1, \dots, n.}$$

En notación vectorial, considerando $\boldsymbol{v}$ como un vector fila,

$$\boxed{\boldsymbol{v} = \boldsymbol{v}P.}$$

Por tanto, $\boldsymbol{v}$ es un vector estacionario de la cadena de Markov discreta cuya matriz de transición es P.

Las razones de visita determinan las proporciones relativas de tráfico entre las estaciones. En particular, v_i/v_j representa la razón entre el número medio de visitas a la estación i y el número medio de visitas a la estación j.

Las ecuaciones de visita no determinan de forma única la escala del vector $\boldsymbol{v}$. Si $\boldsymbol{v}$ es una solución, entonces $c\boldsymbol{v}$ también lo es para cualquier $c > 0$. Se adopta la siguiente normalización:

$$v_1 = 1.$$

Supondremos que la matriz P es irreducible. De este modo, todas las estaciones se comunican entre sí y existe un vector $\boldsymbol{v}$ estrictamente positivo que satisface

$$\boldsymbol{v} = \boldsymbol{v}P.$$

3.4.4. Teorema de Gordon–Newell

Theorem 3.4.1 (Teorema de Gordon–Newell)**.** *Consideremos una red cerrada formada por n estaciones $M/M/1$ y N clientes. Supongamos que la matriz P es irreducible y que los tiempos de servicio en la estación i son exponenciales con parámetro μ_i.*

Sea $\boldsymbol{v} = (v_1, \dots, v_n)$ una solución estrictamente positiva de las ecuaciones de visita

$$\boldsymbol{v} = \boldsymbol{v}P.$$

Entonces la distribución estacionaria de la red es

$$\boxed{\pi(x_1,\ldots,x_n) = \frac{1}{C}\prod_{i=1}^{n}\left(\frac{v_i}{\mu_i}\right)^{x_i}, \qquad (x_1,\ldots,x_n)\in E_N,}$$

donde la constante de normalización es

$$\boxed{C = \sum_{\substack{x_1,\ldots,x_n\geq 0\\ x_1+\cdots+x_n=N}}\prod_{i=1}^{n}\left(\frac{v_i}{\mu_i}\right)^{x_i}.}$$

Esta distribución se denomina *distribución en forma producto*. Sin embargo, a diferencia de lo que sucede en una red abierta de Jackson, la forma producto no implica que los números de clientes en las distintas estaciones sean independientes.

En efecto, las variables satisfacen la restricción

$$X_1 + \cdots + X_n = N.$$

Por consiguiente, si una estación contiene un número elevado de clientes, necesariamente quedan menos clientes disponibles para las demás estaciones.

3.4.5. Existencia de la distribución estacionaria

A diferencia de las redes abiertas, en una red cerrada no es necesario imponer condiciones de la forma $\lambda_i < \mu_i$.

El número total de clientes es fijo y el espacio de estados S es finito. Si la cadena es irreducible, posee una única distribución estacionaria para cualesquiera tasas de servicio positivas:

$$\mu_i > 0, \qquad i = 1,\ldots,n.$$

Si una estación tiene una tasa de servicio pequeña, el sistema no se vuelve inestable. En su lugar, una proporción elevada de los N clientes tiende a acumularse en dicha estación, que actúa como cuello de botella.

3.4.6. Medidas de rendimiento

Una vez conocida la distribución estacionaria, el número medio de clientes en la estación i es

$$L_i = \mathbb{E}[X_i] = \sum_{\boldsymbol{x}\in E_N} x_i\pi(\boldsymbol{x}).$$

Como el número total de clientes es constante,

$$\sum_{i=1}^{n} L_i = N.$$

La probabilidad de que el servidor de la estación i esté ocupado es

$$\mathbb{P}(X_i > 0) = 1 - \mathbb{P}(X_i = 0).$$

3.4.7. Ejemplo: una red cerrada de dos estaciones

Consideremos una red cerrada formada por dos estaciones y N clientes.

Después de completar su servicio en la primera estación, cada cliente pasa a la segunda estación. Después de completar su servicio en la segunda estación, regresa a la primera. Por tanto,

$$p_{1,2} = 1, \qquad p_{2,1} = 1,$$

y

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Los tiempos de servicio en las estaciones 1 y 2 son exponenciales con parámetros μ_1 y μ_2, respectivamente.

No existen llegadas externas ni salidas del sistema. Los N clientes circulan continuamente entre las dos estaciones.

Las ecuaciones de visita son

$$v_1 = v_2 p_{21} = v_2$$

y

$$v_2 = v_1 p_{12} = v_1.$$

Por tanto,

$$v_1 = v_2.$$

Y usando la normalización $\nu_1 = 1$, obtenemos

$$v_1 = v_2 = 1.$$

En consecuencia, el teorema de Gordon–Newell propone la distribución

$$\pi(x_1, x_2) = \frac{1}{C} \left(\frac{1}{\mu_1}\right)^{x_1} \left(\frac{1}{\mu_2}\right)^{x_2},$$

para

$$x_1 + x_2 = N.$$

Como $x_2 = N - x_1$, basta con indicar el número de clientes en la primera estación. Escribimos

$$X(t) := X_1(t).$$

Entonces

$$X_2(t) = N - X(t)$$

y el espacio de estados reducido es

$$\{0, 1, \ldots, N\}.$$

La distribución propuesta puede escribirse como

$$\boxed{\pi_k = \frac{1}{C}\left(\frac{1}{\mu_1}\right)^k \left(\frac{1}{\mu_2}\right)^{N-k}, \qquad k = 0, \ldots, N,}$$

donde

$$C = \sum_{k=0}^{N} \left(\frac{1}{\mu_1}\right)^k \left(\frac{1}{\mu_2}\right)^{N-k}.$$

En el caso particular $N = 2$, se tiene que

$$C = \frac{1}{\mu_1^2} + \frac{1}{\mu_1}\frac{1}{\mu_2} + \frac{1}{\mu_2^2},$$

mientras que para $N = 3$

$$C = \frac{1}{\mu_1^3} + \frac{1}{\mu_1^2}\frac{1}{\mu_2} + \frac{1}{\mu_1^2}\frac{1}{\mu_2^2} + \frac{1}{\mu_2^3}.$$

Transiciones de la cadena

Si $k > 0$, la primera estación contiene algún cliente. Cuando se completa un servicio en ella, un cliente pasa a la segunda estación y el número de clientes en la primera disminuye en una unidad:

$$k \longrightarrow k - 1 \qquad \text{con tasa } \mu_1.$$

Si $k < N$, la segunda estación contiene algún cliente. Cuando se completa un servicio en ella, un cliente regresa a la primera estación y el número de clientes en esta aumenta en una unidad:

$$k \longrightarrow k + 1 \qquad \text{con tasa } \mu_2.$$

Por tanto, $\{X(t)\}_{t\geq 0}$ es un proceso de nacimiento y muerte finito con tasas

$$\lambda_k = \begin{cases} \mu_2, & 0 \leq k < N, \\ 0, & k = N, \end{cases}$$

y

$$\delta_k = \begin{cases} 0, & k = 0, \\ \mu_1, & 1 \leq k \leq N. \end{cases}$$

La matriz generadora es

$$Q = \begin{pmatrix} -\mu_2 & \mu_2 & 0 & \cdots & 0 \\ \mu_1 & -(\mu_1+\mu_2) & \mu_2 & \ddots & \vdots \\ 0 & \mu_1 & -(\mu_1+\mu_2) & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \mu_2 \\ 0 & \cdots & 0 & \mu_1 & -\mu_1 \end{pmatrix}.$$

Verificación mediante las ecuaciones de balance

Para los estados interiores $1 \leq k \leq N-1$, la ecuación de balance global es

$$\pi_k(\mu_1+\mu_2) = \mu_2\pi_{k-1} + \mu_1\pi_{k+1}.$$

La distribución propuesta satisface

$$\frac{\pi_{k+1}}{\pi_k} = \frac{\mu_2}{\mu_1}.$$

Por tanto,

$$\mu_1\pi_{k+1} = \mu_1\frac{\mu_2}{\mu_1}\pi_k = \mu_2\pi_k.$$

De forma análoga,

$$\frac{\pi_{k-1}}{\pi_k} = \frac{\mu_1}{\mu_2},$$

y, en consecuencia,

$$\mu_2\pi_{k-1} = \mu_2\frac{\mu_1}{\mu_2}\pi_k = \mu_1\pi_k.$$

Sumando ambas expresiones,

$$\mu_2\pi_{k-1} + \mu_1\pi_{k+1} = (\mu_1+\mu_2)\pi_k.$$

Por tanto, la ecuación de balance se satisface para todos los estados interiores.

En el estado $k = 0$, solo puede completarse un servicio en la segunda estación. La ecuación de balance es

$$\mu_2\pi_0 = \mu_1\pi_1.$$

Como

$$\frac{\pi_1}{\pi_0} = \frac{\mu_2}{\mu_1},$$

se tiene

$$\mu_1\pi_1 = \mu_1\frac{\mu_2}{\mu_1}\pi_0 = \mu_2\pi_0.$$

En el estado $k = N$, solo puede completarse un servicio en la primera estación. La ecuación de balance es

$$\mu_1\pi_N = \mu_2\pi_{N-1}.$$

Como

$$\frac{\pi_N}{\pi_{N-1}} = \frac{\mu_2}{\mu_1},$$

obtenemos

$$\mu_1\pi_N = \mu_1\frac{\mu_2}{\mu_1}\pi_{N-1} = \mu_2\pi_{N-1}.$$

Así, la distribución propuesta satisface las ecuaciones de balance en todos los estados. Además,

$$\sum_{k=0}^{N}\pi_k = \frac{\displaystyle\sum_{k=0}^{N}\left(\frac{1}{\mu_1}\right)^k\left(\frac{1}{\mu_2}\right)^{N-k}}{G_N} = 1.$$

Por tanto,

$$\pi_k = \frac{1}{G_N}\left(\frac{1}{\mu_1}\right)^k\left(\frac{1}{\mu_2}\right)^{N-k}, \qquad k = 0, \ldots, N,$$

es la distribución estacionaria de esta red cerrada de dos estaciones.