Teoría de Colas aplicada a la Computación

Pablo Real Gómez

Tutora: María Victoria Ibáñez Gual

Máster en Matemática Computacional

Índice general

1 Introducción **1**

2 Procesos estocásticos **5**
- 2.1 Definiciones . . . 5
- 2.2 Procesos de Poisson . . . 7
- 2.3 Procesos de Markov . . . 11
  - 2.3.1 Procesos de Poisson como procesos de Markov . . . 16
- 2.4 Procesos de nacimiento y muerte . . . 17

3 Colas **21**
- 3.1 Descripción de una cola . . . 21
- 3.2 Notación . . . 24
  - 3.2.1 Nomenclatura de Kendall . . . 27
- 3.3 Propiedades . . . 28

4 Modelos Markovianos para colas **31**
- 4.1 Modelos Markovianos simples . . . 31
  - 4.1.1 La cola $M/M/1$ . . . 31
  - 4.1.2 La cola $M/M/s$ . . . 33
  - 4.1.3 La cola $M/M/\infty$ . . . 36
  - 4.1.4 La cola $M/M/1/K$ . . . 37
  - 4.1.5 La cola $M/M/s/K$ . . . 39
  - 4.1.6 La cola $M/M/1/\infty/H$ . . . 41
  - 4.1.7 La cola $M/M/s/\infty/H$ . . . 43
- 4.2 Modelos Markovianos extendidos . . . 44
  - 4.2.1 La cola $M^X/M/1$ . . . 44
  - 4.2.2 La cola $M/M^X/1$ . . . 46

5 Redes de colas **47**
- 5.1 Introducción . . . 47
  - 5.1.1 Notación . . . 48
- 5.2 Redes abiertas de Jackson . . . 49

5.3 Redes cerradas de Jackson 50
5.4 Colas en serie 51
5.5 Colas cíclicas 52
5.6 Red BCMP 53

6 Modelos de colas en computación **55**
6.1 Introducción 55
6.2 Modelo de población finita 56
6.3 Modelo de servidor central 58
6.4 Modelos BCMP 61
6.4.1 Modelo de una clase cerrado 61
6.4.2 Modelo multiclase abierto 64
6.4.3 Modelo multiclase cerrado 65
6.4.4 Modelo multiclase mixto 67

Índice analítico **69**

Bibliografía **71**

Índice de figuras

1.1 Costes totales en un sistema de cola 2

3.1 Representación de un sistema de cola 23

5.1 Ejemplo de una red de colas. 48
5.2 Red de colas en serie. 51
5.3 Red de colas cíclicas. 52

6.1 Modelo de población finita 57
6.2 Modelo de servidor central 59
6.3 Modelo BCMP cerrado 62

1 | Introducción

En la vida moderna los momentos de espera son comunes. Sea cuando vamos al supermercado o cuando esperamos para entrar a un concierto, es habitual que la gente espere en línea a la espera de ser atendida.

Algunos ejemplos de estos eventos son:

- Clientes en un supermercado esperando para pagar.
- Barcos esperando a ser descargados y cargados.
- Pacientes esperando en la sala de espera.
- Productos en una fábrica esperando a ser procesados.
- Mensajes esperando una línea libre para ser enviados.
- Coches esperando en un atasco.

Estas líneas de espera o sistemas de colas tienen algunas cualidades en común que permiten su modelado: clientes entran en el sistema y esperan en una cola a que un servidor quede libre y procese su servicio. Siendo estas colas un fenómeno social, sería beneficioso para todos el poder mejorar las condiciones de servicio, tanto para los que esperan el servicio como los que lo proporcionan.

Una de las preguntas fundamentales es por qué aparece este fenómeno de espera, ya que nadie quiere esperar. Aunque una cola tenga la capacidad suficiente para satisfacer la demanda, la variabilidad de las llegadas de clientes y los tiempos de servicio pueden producir a esperas.

La siguiente pregunta a hacer es hasta qué punto se pueden reducir estos tiempos de espera sin invertir de más en la cola: si en un supermercado contratamos a un centenar de dependientes, los clientes no tendrán que esperar, pero los sueldos serán un sobrecoste no factible. Esto se puede traducir en dos términos usados habitualmente: los costes de espera, es decir, los costes derivados del hecho de que los clientes tengan que esperar más (de esta manera, al incrementarse el nivel de

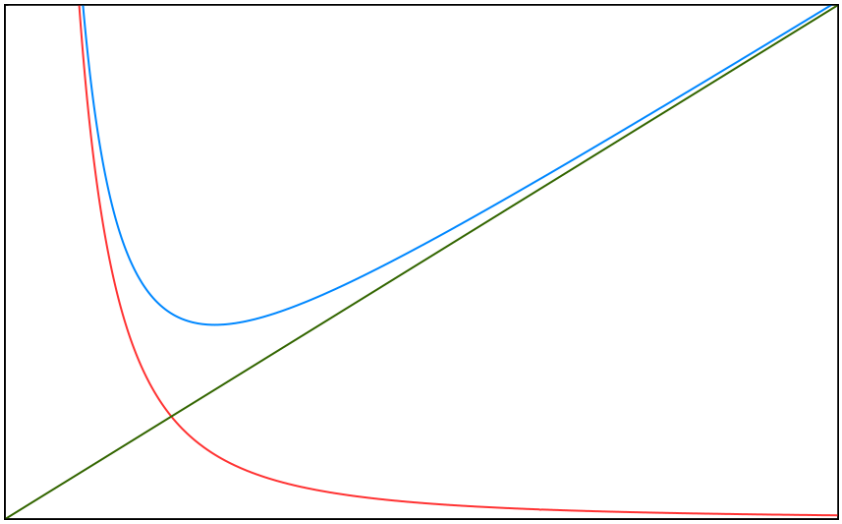

Figura 1.1: Costes totales (en azul) a partir de los costes de servicio (en verde) y los costes de espera (en rojo). El eje de abscisas representa el nivel de servicio, mientras que el eje de ordenadas representa el coste.

servicio, los costes de espera disminuyen), y los costes de servicio, que son los costes derivados de mantener el nivel de servicio deseado (de esta manera, al incrementarse el nivel de servicio, los costes de servicio aumentan). La idea básica en este ámbito es modelizar los costes de espera y los costes de servicio para efectuar un balance y minimizar los costes, como en la figura 1.

A raíz de esta idea, se empezaron a desarrollar los primeros trabajos de la teoría de colas, debidos al danés A. K. Erlang, que en 1909 publicó *The Theory of Probabilities and Telephone Conversations* ([Erl09]). Este trabajo fue una consecuencia natural de su trabajo en la Compañía Telefónica Danesa, aplicando técnicas de la teoría de probabilidad al problema de determinar el número óptimo de líneas telefónicas en una centralita.

Aunque las aportaciones fueron escasas en los inicios de la teoría, las contribuciones se fueron incrementando a partir de los años 50 con la aplicación de la teoría de colas a campos como la Informática, las Telecomunicaciones o la Computación.

A lo largo de todo este trabajo intentaremos responder a las preguntas planteadas anteriormente desde dos enfoques fundamentales: un enfoque matemático, dando fundamento a los modelos usuales usados para describir los sistemas de colas, y desde un enfoque más computacional, intentando analizar propiedades y detalles que hagan más fácil la implementación y simulación de este tipo de sistemas. Para

este propósito, tomaremos como referencias [Aba02, All90, Bha15, MOVG10] para explicar y detallar los distintos modelos de colas.

2 | Procesos estocásticos

Los procesos estocásticos son útiles para describir algunos eventos no deterministas que no pueden ser descritos a partir de elementos habituales de la probabilidad como las distribuciones.

Daremos las definiciones básicas y algunos resultados relevantes respecto a procesos estocásticos en la sección 2.1. Introduciremos los procesos de Poisson, un tipo específico de procesos de conteo que generaliza las distribuciones de Poisson, y que nos serán útiles para describir colas, en la sección 2.2. En la sección 2.3 detallaremos un tipo de proceso estocástico ampliamente estudiado proceso de Poisson, y obtendremos algunos resultados más para los procesos de Poisson mediante el uso de técnicas asociadas a los procesos de Poisson en la subsección 2.3.1. Por último, estudiaremos los procesos de nacimiento y muerte en la sección 2.4, otra generalización relevante en el estudio de colas.

2.1 | Definiciones

A veces los datos obtenidos a partir de una variable aleatoria son insuficientes, teniendo la necesidad de variar esta información con el tiempo o similares. A partir de esta necesidad aparece la definición de *proceso estocástico*.

Definición 2.1. Llamaremos *proceso estocástico* a una familia de variables aleatorias $\{X_t\}_{t\in T}$, donde cada X_t es una variable aleatoria $X_t\colon \Omega \to \mathbb{R}$, y todas ellas están definidas en el mismo espacio de probabilidad Ω.

El subconjunto $E \subseteq \mathbb{R}$ en el que toman valores las variables se denomina *espacio de estados*, mientras que el conjunto de índices T se denomina *espacio de tiempos*.

Según E y T sean finitos (o infinitos numerables) o contengan al menos un intervalo, diremos que son procesos estocásticos con *espacio de estados discreto o continuo* y en *tiempo discreto o continuo*.

Esta definición incluye las variables aleatorias clásicas usando un índice temporal

de un único elemento. También incluye la definición de vector aleatorio al usar un espacio temporal finito, o el de una sucesión de variables aleatorias cuando el espacio temporal es precisamente el conjunto de los naturales. Por tanto, la mayor novedad de este concepto se da cuando usamos un intervalo como espacio temporal.

Ejemplo 2.2. Algunos ejemplos de situaciones reales que se pueden modelizar mediante procesos estocásticos son:

- El tiempo de respuesta en un sistema informático multiusuario según la hora del día, con $E = (0, \infty)$ y $T = [0, 24)$.
- El número de terminales conectadas a un servidor según la hora del día, con $E = \{0, 1, 2, \dots\}$ y $T = [0, 24)$.
- El tiempo de CPU de un servidor dedicado a usuarios según el día del año, con $E = [0, \infty)$ y $T = \{1, \dots, 365\}$.

Recordamos que para cada variable aleatoria se pueden definir su media y su varianza, además de la covarianza de dos variables aleatorias. Este concepto también se puede extender para procesos estocásticos.

Definición 2.3. Se denomina *función de medias* de un proceso estocástico $\{X_t\}_{t \in T}$ a la aplicación definida por:

$$m(t) = E[X_t] \qquad \forall t \in T$$

Se denomina *función de autocovarianzas* de un proceso estocástico $\{X_t\}_{t \in T}$ a la aplicación definida por:

$$c(s, t) = E[(X_s - m(s))(X_t - m(t))] \qquad \forall s, t \in T$$

Nota 2.4. La función de autocovarianzas es equivalente a la covarianza de las variables asociadas, i.e., $c(s, t) = Cov(X_s, X_t)$. Además, de manera automática hereda la simetría de la covarianza, y por tanto la función de autocovarianzas es simétrica, i.e., $c(s, t) = c(t, s)$ para todos $s, t \in T$.

Es muy cómodo tratar con procesos estocásticos que siguen determinados patrones, como por ejemplo, que su comportamiento sea similar aunque cambie el índice temporal.

Definición 2.5. Diremos que un proceso estocástico $\{X_t\}_{t\in T}$ es *débilmente estacionario* si:

1. Su función de medias es constante, es decir, $m(t) = m(s)$ para todos $t, s \in T$.
2. La función de autocovarianzas depende únicamente de la diferencia de tiempos, es decir, $c(s,t) = c(t-s)$ para todos $s \leq t \in T$.

Definición 2.6. Diremos que un proceso estocástico $\{X_t\}_{t\in T}$ es *fuertemente estacionario* si para cualquier $n \in \mathbb{N}$, cualesquiera instantes $t_1 < \ldots < t_n$, y cualquier h tal que $t_i + h \in T$ con $i = 1, \ldots, n$, se verifica que el vector aleatorio $(X_{t_1}, \ldots, X_{t_n})$ tiene la misma distribución que $(X_{t_1+h}, \ldots, X_{t_n+h})$.

Proposición 2.7. *Todo proceso estocástico fuertemente estacionario también es un proceso estocástico débilmente estacionario.*

Demostración.

Supongamos que $\{X_t\}_{t\in T}$ es un proceso estocástico fuertemente estacionario.

Obsérvese que tomando $n = 1$ con $t_1 = t$ y $h = s - t$ en la condición de fuertemente estacionario, tenemos que X_t y X_s tienen idéntica distribución para todos $t, s \in T$. De esta manera tenemos que $m(s) = E[X_s] = E[X_t] = m(t)$ para todos $t, s \in T$.

Ahora, aplicando de nuevo en la condición de fuertemente estacionario para $n = 2$, con $t_1 = s$ y $t_2 = t$, y $h = -s$, tenemos que (X_s, X_t) tiene misma distribución que (X_0, X_{t-s}). De esta manera, tenemos que $c(s,t) = c(0, t-s) = \gamma(t-s)$.

Como se cumplen las dos condiciones requeridas, tenemos que $\{X_t\}_{t\in T}$ es un proceso estocástico débilmente estacionario.

□

2.2 | Procesos de Poisson

Definición 2.8. Sea un proceso estocástico en tiempo continuo y con espacio de estados discreto $\{N_t\}_{t\geq 0}$. Diremos que es un *proceso de conteo* si satisface:

1. $N_0 = 0$.
2. La variable aleatoria N_t toma valores enteros no negativos.
3. Si $s < t$, entonces $N_s(\omega) \leq N_t(\omega)$ para todo $\omega \in \Omega$.

El nombre de este tipo de procesos es bastante explícito: sirven para modelizar el número de ocurrencias de un cierto fenómeno desde el inicio de tiempo (fijado en $t = 0$) hasta el instante t.

Las condiciones de la definición tienen también una interpretación clara:

1. $N_0 = 0$ porque en el inicio de tiempos aún no puede haber ninguna ocurrencia.
2. N_t debe tomar valores enteros no negativos porque el número de ocurrencias también es un valor entero no negativo.
3. Si $s < t$ entonces $N_s(\omega) \leq N_t(\omega)$ para cualquier $\omega \in \Omega$ porque el número de ocurrencias no puede decrecer con el tiempo.

Ejemplo 2.9. Algunos ejemplos de situaciones reales que se pueden modelizar mediante procesos de conteo son:

- Número de procesos enviados a un servidor hasta un tiempo t.
- Número de paquetes entregados por un servicio de mensajería hasta un tiempo t.
- El número de corredores de una maratón que han llegado a la meta antes del instante t.

Para definir correctamente los procesos de Poisson usaremos el concepto siguiente:

Definición 2.10. Sea $f(x)$ una función. Diremos que $f \in o(g(x))$ si se tiene que:

$$\frac{f(x)}{g(x)} \to 0$$

Es decir, $f(x)$ crece mucho más lentamente que $g(x)$.

Es habitual abusar de esta notación en expresiones, y usar $o(g(x))$ en lugar de $f(x)$ para expresar una función cuyo valor exacto no es relevante, siendo únicamente importante notar que su crecimiento es lento respecto a $g(x)$.

Nosotros también aprovecharemos este abuso de notación para definir el proceso de Poisson.

Definición 2.11. Sea un proceso de contar $\{N_t\}_{t\geq 0}$. Diremos que es un *proceso de Poisson* de parámetro $\lambda > 0$ si verifica las siguientes propiedades:

1. El proceso tiene incrementos independientes. Es decir, para instantes $0 \leq t_0 < t_1 < \ldots < t_n$, las variables aleatorias $N_{t_1} - N_{t_0}, N_{t_2} - N_{t_1}, \ldots, N_{t_n} - N_{t_{n-1}}$ son independientes.

2. El proceso tiene incrementos estacionarios. Es decir, para todos $s < t$ y h arbitrario, se tiene que $N_{t+h} - N_{s+h}$ tiene la misma distribución que $N_t - N_s$.

3. La probabilidad de que se dé exactamente una ocurrencia del fenómeno en el intervalo $[0, h]$ es $P(N_h = 1) = \lambda h + o(h)$.

4. La probabilidad de que en el mismo intervalo $[0, h]$ haya dos o más ocurrencias en ese mismo intervalo es $P(N_h \geq 2) = o(h)$.

Este proceso es un caso concreto de proceso de conteo, en el que las ocurrencias siguen unos patrones concretos, reflejados por el parámetro λ asociado al proceso.

Como antes, las condiciones que plantea la definición tienen las siguientes interpretaciones:

1. Notamos que las variables aleatorias $N_{t_{i+1}} - N_{t_i}$ representan los números de ocurrencias en los intervalos $(t_i, t_{i+1}]$. Por tanto, la independencia de los incrementos significa que el número de ocurrencias en un intervalo es no depende de las ocurrencias anteriores.

2. Notamos que para todos $s < t$ se tiene que $N_t - N_s$ tiene la misma distribución que N_{t-s}. Esto quiere decir que el conteo de ocurrencias en el intervalo $(s, t]$ solo depende de la longitud del intervalo $(t - s)$.

3. La probabilidad de exactamente una ocurrencia en un intervalo de longitud h es $\lambda h + o(h)$, lo que quiere decir que la probabilidad es aproximadamente proporcional al parámetro λ.

4. De igual manera al anterior, la probabilidad de más de una ocurrencia es $o(h)$, lo que quiere decir que la probabilidad es mucho menor que el tamaño del intervalo, haciendo este suceso poco probable para valores pequeños de h.

Ejemplo 2.12. Algunos ejemplos de situaciones reales que se pueden modelizar mediante procesos de Poisson son:

- El número de goles en un partido de fútbol.
- El número de defectos en una línea de producción.

A continuación se presentan algunos resultados teóricos relevantes para el proceso de Poisson.

Proposición 2.13. *El número medio de ocurrencias por unidad de tiempo de un fenómeno descrito por un proceso de Poisson de parámetro λ es precisamente λ. Consecuentemente, la media de tiempo entre ocurrencias es λ^{-1}.*

Demostración. Primero notamos que $P(N_h = 0) = 1 - P(N_h = 1) - P(N_h \geq 2) = 1 - \lambda h + o(h)$. Recordamos que por ser los incrementos estacionarios, las ocurrencias del fenómeno en un intervalo de longitud h, dadas por $N_{t+h} - N_t$, tienen la misma distribución que N_h. Así, tenemos que:

$$E[N_h] = 0 \cdot (1 - \lambda h + o(h)) + 1 \cdot (\lambda h + o(h)) + o(h) = \lambda h + o(h)$$

De esta manera, el número medio de ocurrencias por unidad de tiempo es:

$$\frac{E[N_{t+h} - N_t]}{h} = \frac{E[N_h]}{h} = \frac{\lambda h + o(h)}{h} = \lambda + \frac{o(h)}{h} \to \lambda$$

Por último, la media de tiempo entre ocurrencias es simplemente el inverso de la media de ocurrencias por unidad de tiempo.

□

El siguiente resultado nos sirve para simplificar la simulación del proceso de Poisson o de calcular algunas probabilidades asociadas.

Teorema 2.14. *Sea $\{N_t\}_{t\geq 0}$ un proceso de Poisson de parámetro λ y sean $0 = t_0 < t_1 < t_2 < \dots$ los tiempos de ocurrencia del fenómeno en cuestión. Definimos los tiempos entre ocurrencias $\tau_1 = t_1 - t_0, \tau_2 = t_2 - t_1, \dots$, entonces las variables aleatorias τ son i.i.d. con distribución exponencial de parámetro λ. El recíproco también es cierto: si $\{N_t\}_{t\geq 0}$ es un proceso de contar y los tiempos τ_i son variables i.i.d. con distribución exponencial de parámetro λ, entonces $\{N_t\}_{t\geq 0}$ es un proceso de Poisson de parámetro λ.*

Demostración. Denotamos por X_t la variable aleatoria referente al tiempo que tarda en haber una ocurrencia suponiendo que ha habido una ocurrencia en el instante t. Con esta notación, tenemos que $P(X_t > x) = P(N_t = N_{t+x})$. De esta manera:

$$P(X_t \leq x) = 1 - P(X_t > x) = 1 - P(N_{t+x} - N_t = 0) = 1 - P(N_x = 0)$$

Sabemos que la probabilidad de que no haya ninguna ocurrencia tiene una distribución de Poisson, así que:

$$P(X_t > x) = 1 - P(N_x = 0) = 1 - \frac{(\lambda x)^0}{0!} e^{-\lambda x} = 1 - e^{-\lambda x}$$

Esta distribución es claramente la de una exponencial. Que sean independientes se deduce de que los incrementos del proceso de Poisson son estacionarios.

La demostración en el otro sentido es fácil de comprobar: a partir de que los tiempos entre llegadas tengan distribución exponencial, se puede ver fácilmente que el proceso $\{N_t\}_{t\geq 0}$ satisface las propiedades de un proceso de Poisson.

□

2.3 | Procesos de Markov

Algunos de los modelos usados ampliamente en la teoría de colas se basan en procesos de Markov. Empezamos pues definiendo este tipo de procesos.

Definición 2.15. Sea un proceso estocástico $\{X_t\}_{t\in T}$. Diremos que X es un *proceso de Markov* o *proceso Markoviano* si:

$$P(X_t \leq x | X_{t_1} = x_1, \ldots, X_{t_n} = x_n) = P(X_t \leq x | X_{t_n} = x_n)$$

En caso de que los espacios de tiempos y de estados sean discretos, es usual decir que $\{X_n\}$ es una *cadena de Markov* o *cadena Markoviana*. También es usual denotar:

$$P(X_t = j | X_s = i) = P_{i\to j}^{s\to t}$$

La dependencia que se exhibe en estos procesos es una dependencia de un estado, donde el estado del proceso únicamente depende del último tiempo en el que conocemos la información del proceso y del tiempo transcurrido entre entonces, de manera que la información anterior al último estado conocido no afecta al estado del proceso. Esta dependencia simplifica sustancialmente el análisis de este tipo de procesos a la vez que mantiene la información esencial de los sistemas que modeliza.

Debido a la definición de los procesos Markovianos, es usual estudiarlos en base a las probabilidades de transición de un estado a otro. Como en los procesos de Poisson, se puede pasar de estas probabilidades condicionadas a probabilidades incondicionadas al considerar el estado anterior como el inicio de tiempos.

Nota 2.16. Es usual usar en espacio de estados discreto, sobre todo en el ámbito que nos ocupa. Por tanto, a partir de ahora supondremos que los procesos Markovianos estudiados tienen espacio de estados discreto. Notamos que el análisis para procesos Markovianos con espacio de estados continuo es similar, usando conceptos análogos cuando sea necesario (i.e., cambiando sumatorios por integrales, límites de sucesiones por límites de funciones, etc.).

NOTA 2.17. En el ámbito de los procesos de Markov, los incrementos son fuertemente estacionarios si $P_{i\to j}^{m\to n}$ es igual que $P_{i\to j}^{m+h\to n+h}$.

Una de las propiedades fundamentales de los procesos de Markov es la siguiente proposición.

PROPOSICIÓN 2.18 (Chapman–Kolmogorov). *Sea $\{X_t\}$ un proceso Markoviano con incrementos fuertemente estacionarios. Entonces para todos $s < t$, y para u con $s < u < t$, se tiene que:*

$$P_{i\to j}^{s\to t} = \sum_k P_{i\to k}^{s\to u} P_{k\to j}^{u\to t} \tag{2.3.1}$$

Demostración.

$$\begin{aligned}
P_{i\to j}^{s\to t} &= P(X_t = j | X_s = i) \\
&= \frac{P(X_s = i, X_t = j)}{P(X_s = i)} \\
&= \sum_k \frac{P(X_s = i, X_u = k, X_t = j)}{P(X_s = i)} \\
&= \sum_k \frac{P(X_s = i, X_u = k)}{P(X_s = i)} \frac{P(X_s = i, X_u = k, X_t = j)}{P(X_s = i, X_u = k)} \\
&= \sum_k P(X_u = k | X_s = i) P(X_t = j | X_s = i, X_u = k) \\
&= \sum_k P(X_u = k | X_s = i) P(X_t = j | X_u = k) \\
&= \sum_k P_{i\to k}^{s\to u} P_{k\to j}^{u\to t}
\end{aligned}$$

En la última igualdad hemos usado la condición de que un proceso Markoviano depende únicamente de su último estado conocido.

□

Esta igualdad es fácil de entender al considerar su sentido: la probabilidad de pasar de un estado i a un estado j es la suma de todos los caminos posibles del estado i al estado j, suponiendo que pasamos por el estado k. La probabilidad de cada uno de estos caminos es la probabilidad de ir del estado i al estado k multiplicada por la probabilidad de ir del estado k al estado j.

Con esta propiedad y algunas suposiciones, podemos hallar fácilmente el comportamiento del sistema y las probabilidades de cada uno de los estados.

NOTA 2.19. Podemos simplificar aún más la notación. Usaremos $P_{ij}^{(t)} = P_{i\to j}^{0\to t}$.

DEFINICIÓN 2.20. Sea $\{X_n\}$ una cadena de Markov con incrementos fuertemente estacionarios. Entonces podemos definir la *matriz de transición al tiempo n* (o simplemente *matriz de transición*) como:

$$P^{(n)} = \begin{pmatrix} P_{00}^{(n)} & P_{01}^{(n)} & P_{02}^{(n)} & \cdots \\ P_{10}^{(n)} & P_{11}^{(n)} & P_{12}^{(n)} & \cdots \\ P_{20}^{(n)} & P_{21}^{(n)} & P_{22}^{(n)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Es decir, la matriz de transición es la matriz con valor $P_{ij}^{(n)}$ en la posición (i, j).

NOTA 2.21. Es usual denotar $P^{(1)} = P$ a la matriz de transición de un paso.

Observemos que esta matriz tiene unas propiedades (heredadas de las propiedades de los procesos Markovianos) que la hacen muy útil para obtener probabilidades.

PROPOSICIÓN 2.22. *Sea $P^{(n)}$ la matriz de transición de una cadena Markoviana. Entonces se tiene que:*

1. *Para todos los estados i y j, se tiene que $0 \leq P_{ij}^{(n)} \leq 1$.*
2. *Para todos los estados i, se tiene que $\sum_j P_{ij}^{(n)} = 1$.*
3. *Para todos los estados i y j, y para todos los tiempos $0 < r < n$, se tiene que:*
$$P_{ij}^{(n)} = \sum_k P_{ik}^{(r)} P_{kj}^{(n-r)}$$
4. *En general, se tiene que:*
$$P^{(n)} = P^{(r)} P^{(n-r)}$$
5. *En general, se tiene que:*
$$P^{(n)} = P^n$$

Ahora pasamos al caso en el que el espacio de tiempos es continuo, ya profundizando en casos más apropiados para el trabajo que nos ocupa. En este caso, y de manera análoga a cuando considerábamos tiempo discreto, podemos considerar $P_{ij}(t) = P(X(t) = j | X(0) = i)$.

Para efectuar el siguiente análisis necesitamos ciertas hipótesis, que tomamos como ciertas en base a su sentido e interpretación reales. La primera es que $P_{ij}(t)$

es continua, y la segunda es que $P_{ij}(t)$ se acerca a 1 cuando $t \to 0$ y $i = j$, pero se acerca a 0 cuando $t \to 0$ y $i \neq j$.

Desarrollamos en serie de Taylor, en un entorno h para un tiempo cualquiera t:

$$P_{ij}(t, t+h) = P_{ij}(t) + P'_{ij}(t)h + P''_{ij}(t)\frac{h^2}{2} + \dots$$

Podemos centrar este desarrollo en serie de Taylor alrededor de $t = 0$, de manera similar a lo razonado anteriormente, debido a la estacionareidad, obteniendo así:

$$P_{ij}(h) = P_{ij}(0) + P'_{ij}(0)h + P''_{ij}(0)\frac{h^2}{2} + \dots$$

A partir de aquí y de las condiciones anteriores cuando $i = j$ y $i \neq j$ podemos obtener:

$$\lim_{h\to 0} \frac{P_{ij}(h)}{h} = P'_{ij}(0) = \lambda_{ij}$$

$$\lim_{h\to 0} \frac{P_{ii}(h) - 1}{h} = P'_{ii}(0) = -\lambda_{ii}$$

Definición 2.23. Definimos los λ_{ij} anteriores como las *tasas de transición*. Estas tasas de transición también se conocen como *generadores*, y se representan en una matriz como:

$$\Lambda = \begin{pmatrix} -\lambda_{00} & \lambda_{01} & \lambda_{02} & \cdots \\ \lambda_{10} & -\lambda_{11} & \lambda_{12} & \cdots \\ \lambda_{20} & \lambda_{21} & -\lambda_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

En los procesos Markovianos con espacio de tiempos continuo, la matriz Λ tiene el mismo papel que la matriz de transición P para procesos Markovianos con espacio de tiempos discreto.

Las tasas de transición cumplen una propiedad importante, presentada en la siguiente proposición.

Proposición 2.24. *Las tasas de transición de un proceso Markoviano cumplen que:*

$$\lambda_{ii} = \sum_{j\neq i} \lambda_{ij}$$

Demostración. Recordamos que $\sum_j P_{ij}(t) = 1$. Derivando esta expresión y aplicándola en $t = 0$ obtenemos la igualdad deseada.

□

Estas tasas de transición nos permiten caracterizar el comportamiento de las probabilidades de transición.

TEOREMA 2.25. *Sea $\{X_t\}$ un proceso Markoviano con incrementos fuertemente estacionarios. Entonces se tiene que:*

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

Demostración. Recordamos que:

$$P_{ij}(t+h) = \sum_k P_{ik}(t)P_{kj}(h)$$

Podemos restar $P_{ij}(t)$ a ambos lados y dividir por h, obteniendo:

$$\frac{P_{ij}(t+h) - P_{ij}(t)}{h} = P_{ij}(t)\frac{P_{jj}(h) - 1}{h} + \sum_{k \neq j} \frac{P_{ik}(t)P_{kj}(h)}{h}$$

Tomando ahora el límite cuando $h \to 0$, tenemos:

$$P'_{ij}(t) = -\lambda_{jj}P_{ij}(t) + \sum_{k \neq j} \lambda_{kj}P_{ik}(t) \tag{2.3.2}$$

En notación matricial, podemos reescribir como:

$$P'(t) = P(t)\Lambda \tag{2.3.3}$$

□

COROLARIO 2.26. *Sea $\{X_t\}$ un proceso Markoviano con incrementos fuertemente estacionarios. Entonces se tiene que:*

$$P'(t) = \Lambda P(t) \tag{2.3.4}$$

Demostración. Es suficiente repetir el proceso anterior tomando:

$$P_{ij}(h+t) = \sum_k P_{ik}(h)P_{kj}(t)$$

□

Estas ecuaciones diferenciales se pueden resolver con una condición inicial. En nuestro caso, simplemente tomamos $P(0) = I$, ya que en el inicio de tiempos, el sistema no ha tenido tiempo de cambiar de estado.

Corolario 2.27. *La solución de las ecuaciones diferenciales anteriores son iguales y son:*

$$P(t) = e^{\Lambda t} = \sum_{n=0}^{\infty} \Lambda^n \frac{t^n}{n!} \tag{2.3.5}$$

Con este resultado podemos analizar fácilmente cualquier proceso de Markov con incrementos fuertemente estacionarios. En concreto, pasaremos a analizar los procesos de Poisson como procesos de Markov.

2.3.1 | Procesos de Poisson como procesos de Markov

Vamos a ver el proceso de Poisson como un proceso de Markov, para así usar las técnicas discutidas en 2.3 en el análisis de los procesos de Poisson.

En este caso $P_{ij}(t)$ representa la probabilidad de pasar del estado en el que han habido i ocurrencias a el estado en el que han habido j ocurrencias en un tiempo t.

El primer paso para el análisis es obtener la matriz Λ. Observamos que como estamos tratando con un proceso de conteo, el número de ocurrencias no puede decrecer en el tiempo, de manera que los elementos λ_{ij} con $i > j$ son nulos. Por otra parte, la probabilidad de que haya dos o más ocurrencias es de $o(h)$, por lo que tenemos:

$$P'_{ij}(0) = \lim_{h\to 0} \frac{P_{ij}(h)}{h} = \lim_{h\to 0} \frac{o(h)}{h} = 0 \text{ si } j > i+1$$

Por otra parte, la probabilidad de que se dé exactamente una ocurrencia es de $\lambda h + o(h)$, de manera que:

$$P'_{i,i+1}(0) = \lim_{h\to 0} \frac{P_{i,i+1}(h)}{h} = \lim_{h\to 0} \frac{\lambda h + o(h)}{h} = \lambda$$

Por último, la probabilidad de que no se dé ninguna ocurrencia es $1 - \lambda h + o(h)$, y así:

$$P'_{ii}(0) = \lim_{h\to 0} \frac{P_{ii}(h) - 1}{h} = \lim_{h\to 0} \frac{1 - \lambda h + o(h) - 1}{h} = -\lambda$$

En resumen, tenemos que:

$$P'_{i,i+1}(0) = \lambda \qquad P'_{ii}(0) = -\lambda$$

De este modo, obtenemos la matriz:

$$\Lambda = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Como además el proceso de Poisson es un proceso de conteo con valor inicial nulo (es decir, $X(0) = 0$), podemos escribir simplemente $P_n(t)$ en vez de escribir $P_{0\to n}(t)$. De esta manera, $P_n(t)$ es la probabilidad de que hasta el instante t haya exactamente n ocurrencias del fenómeno estudiado. Con esta notación y las ecuaciones obtenidas a partir de la matriz Λ, obtenemos las ecuaciones:

$$\begin{aligned} P_0'(t) &= -\lambda P_0(t) \\ P_n'(t) &= -\lambda P_n(t) + \lambda P_{n-1}(t) \end{aligned} \tag{2.3.6}$$

Teorema 2.28. *Las marginales de un proceso de Poisson de parámetro λ son distribuciones de Poisson de parámetro λt. Es decir, $X_t \sim Poi(\lambda t)$.*

Demostración. Resolviendo 2.3.6 obtenemos:

$$P_n(t) = P(X(t) = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!} \tag{2.3.7}$$

Estas soluciones se corresponden claramente con las distribuciones deseadas.

□

Las soluciones a estos procesos no suelen ser tan fáciles de obtener. En este caso, la estructura bidiagonal de Λ y el hecho de que las tasas de transición sean constantes simplifican enormemente el proceso, pero en otras ocasiones la solución no es tan fácil. En caso de que Λ sea finita y diagonalizable, se pueden usar los valores propios de Λ para determinar la solución usando 2.3.5. También se pueden usar métodos computacionales para obtener aproximaciones de la solución.

Repetiremos análisis análogos al que acabamos de realizar al definir procesos que modelizan distintos tipos de colas.

2.4 | Procesos de nacimiento y muerte

En 2.2 hemos visto los procesos de Poisson, ejemplos de procesos de conteo de especial utilidad en la teoría de colas. Aún más generales que este tipo de procesos

son los procesos de nacimiento y muerte, que también admiten la posibilidad de que dicho conteo descienda. Además de generalizar en este sentido, también permiten que las tasas de nacimiento y de muerte puedan depender del número de individuos en la población.

Definición 2.29. Sea un proceso estocástico $\{N_t\}_{t\geq 0}$ con espacio de estados $E = \{0, 1, 2, \dots\}$. Diremos que es un *proceso estocástico de nacimiento y muerte* si existen λ_n y μ_n llamadas *tasa de nacimiento* y *tasa de muerte* que verifican las siguientes propiedades:

- Se puede cambiar del estado $n = 0$ a $n = 1$, y en general, de n a $n-1$ y $n+1$ cuando $n \geq 1$.
- Si el sistema está en estado n, la probabilidad de que dentro de un tiempo h haya pasado a $n + 1$ es de $\lambda_n h + o(h)$.
- Si el sistema está en estado $n \geq 1$, la probabilidad de que dentro de un tiempo h haya pasado a $n - 1$ es de $\mu_n h + o(h)$.
- La probabilidad de que ocurra más de un cambio dentro de un tiempo h es $o(h)$.

Observemos que el proceso de Poisson es un caso concreto de la definición anterior al considerar tasas de muerte nulas y tasas de nacimiento constantes e iguales a λ. También podemos ver este proceso como un proceso de Markov de manera análoga.

Denotamos por $P_n(t)$ la probabilidad de que el sistema esté en el estado n en el tiempo t. Podemos estudiar fácilmente estas probabilidades. Si $n \geq 1$, tenemos:

$$P_n(t+h) = P_{n-1}(t)\lambda_{n-1}h + P_n(t)(1 - \lambda_n h - \mu_n h) + P_{n+1}(t)\mu_{n+1}h + o(h)$$

De esta manera, se tiene que:

$$P_n'(t) = \lim_{h\to 0} \frac{P_n(t+h) - P_n(t)}{h} = \lambda_{n-1}P_{n-1}(t) - (\lambda_n + \mu_n)P_n(t) + \mu_{n+1}P_{n+1}(t)$$

De esta manera, y trabajando de forma análoga para $n = 0$, tenemos las *ecuaciones diferenciales de balance*:

$$\begin{cases} P_n'(t) = \lambda_{n-1}(t) - (\lambda_n + \mu_n)P_n(t) + \mu_{n+1}P_{n+1}(t) & n \geq 1 \\ P_0'(t) = -\lambda_0 P_0(t) + \mu_1 P_1(t) \end{cases}$$

Podemos suponer que el proceso de nacimiento y muerte es estacionario, de manera que las funciones $P_n(t)$ son constantes p_n, de manera que el sistema de ecuaciones diferenciales se transforma en el sistema:

$$\begin{cases} (\lambda_n + \mu_n)p_n = \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1} & n \geq 1 \\ \lambda_0 p_0 = \mu_1 p_1 \end{cases}$$

Este sistema es fácil de resolver con la siguiente propiedad.

PROPOSICIÓN 2.30. *Para cualquier $n \geq 1$, se tiene que:*

$$\mu_n p_n = \lambda_{n-1}p_{n-1}$$

Demostración.

Procedemos por inducción. Para $n = 1$, el resultado se obtiene directamente de la ecuación correspondiente al estado $n = 0$ del sistema. Ahora, para $n > 1$, tenemos:

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

Por la hipótesis de inducción, tenemos:

$$(\lambda_n + \mu_n)p_n = \mu_n p_n + \mu_{n+1}p_{n+1}$$

Por tanto, tenemos que:

$$\lambda_n p_n = \mu_{n+1}p_{n+1}$$

□

De esta manera tenemos que en general:

$$p_n = p_0 \prod_{k=1}^{n} \frac{\lambda_{k-1}}{\mu_k}$$

Definimos:

$$c_n = \prod_{k=1}^{n} \frac{\lambda_{k-1}}{\mu_k}$$

Además, podemos definir $c_0 = 1$. De esta manera, $p_n = c_n p_0$. Como las p_n representan probabilidades, su suma debe ser exactamente 1, y así:

$$1 = \sum_{n=0}^{\infty} p_n = p_0 \sum_{n=0}^{\infty} c_n$$

De esta manera:

$$p_0 = \left(\sum_{n=0}^{\infty} c_n \right)^{-1}$$

En este caso, tenemos que suponer que esta serie no diverge. Esta última es la llamada *condición de estado estacionario*.

TEOREMA 2.31 (Condición de estado estacionario). *Existe un proceso de nacimiento y muerte estacionario con tasas de nacimiento λ_n y tasas de muerte μ_n si y solo si la serie $\sum_{n=0}^{\infty} c_n$ es convergente.*

3 | Colas

En este capítulo introduciremos el concepto de cola, propiamente dicho, y todos sus parámetros asociados, en la sección 3.1. Estableceremos la terminología básica del campo en la sección 3.2, y por último daremos algunas propiedades básicas de las colas y algunos conceptos de interés (como las fórmulas de Little) en la sección 3.3.

3.1 | Descripción de una cola

Ya hemos visto en el capítulo 1 algunos ejemplos de fenómenos reales en los que se forman colas o líneas de espera. Estas suelen ocurrir cuando la demanda de un servicio es superior a la capacidad para dar dicho servicio.

Estos fenómenos de colas son frecuentes en el contexto de la informática y las telecomunicaciones. Por ejemplo, los procesos enviados a un procesador para su ejecución forman una cola mientras no son atendidos, o podemos recibir la señal de líneas ocupadas si la central de telefonía de la que dependemos está colapsada.

Todos estos fenómenos tienen algunas características comunes que dan lugar al modelo de sistema de una cola. En dicho modelo distinguimos los siguientes elementos:

- *Población:* Es el conjunto de individuos (no hablamos necesariamente de seres vivos) que pueden llegar a solicitar el servicio. Podemos considerar este conjunto finito o infinito. Claramente, este segundo caso no es realista, pero permite resolver de manera más sencilla los casos en los que la población es muy grande. Es usual considerar que la población es infinita cuando la población es grande (aunque finita) y el número de individuos que ya están solicitando el servicio no afecta a la frecuencia con la que la población genera nuevas peticiones de servicio.

- *Llegadas:* El proceso de llegada indica los tiempos de llegada de cada uno de

los clientes (también pueden llegar varios clientes en grupo, como veremos en la sección 4.2). Como hemos visto en la sección 2.2, es usual usar los tiempos entre llegadas para describir este proceso.

- *Cliente:* Es el individuo de la población que solicita el servicio. Más que el cliente individual, es importante conocer el patrón según el que la población genera clientes. Es habitual tomar como referencia los tiempos entre llegada de los clientes. También es habitual suponer que para una población infinita estos tiempos de llegada no dependen del número de clientes en espera para completar el servicio.

- *Capacidad:* Es el máximo número de clientes que pueden estar haciendo cola (no en el sistema de la cola, es decir, no contamos los que ya están siendo servidos). Esta también se puede considerar finita o infinita, siendo más sencillo considerar la segunda opción, como antes. Esto no es una gran restricción en caso de que la capacidad sea suficientemente grande, es decir, si es difícil llegar a la situación en la que no puedan añadirse más clientes a la cola.

- *Disciplina:* Es el modo en que se escoge el siguiente cliente en ser servido. Hay varias disciplinas habituales:

  - *FIFO (First In First Out):* El primer cliente en ser atendido es el que antes haya llegado.
  - *LIFO (Last In First Out):* El primer cliente en ser atendido es el último que haya llegado.
  - *SIRO (Service In Random Order):* El primer cliente en ser atendido se decide aleatoriamente.
  - *PS (Processor Sharing):* Se otorga a cada cliente un pequeño tiempo de servicio de forma secuencial, hasta completar el tiempo de servicio. Este tipo de disciplina solo tiene sentido en algunas circunstancias, como el ámbito de la informática.
  - *PR (Priority):* Cada cliente tiene una prioridad de servicio, y los que tienen mayor prioridad son los primeros en ser servidos. Este tipo de disciplina se usa por ejemplo en las salas de urgencias de los hospitales.
  - *GD (General Discipline):* Otra disciplina a especificar.

- *Servicio:* Procedimiento mediante el cual se da servicio a los clientes que lo han solicitado. Para especificarlo se indican el número de servidores (es decir, el número de servicios que se pueden atender al mismo tiempo), o

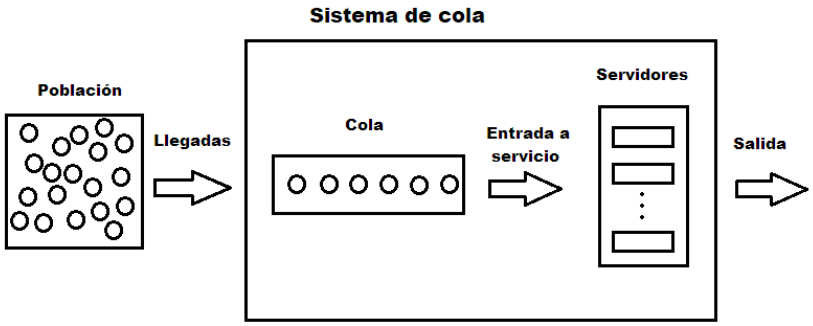

Figura 3.1: Representación de un sistema de cola

en su defecto, si este número es aleatorio, la distribución de probabilidad del mismo. También depende de la distribución de probabilidad del tiempo necesario para dar un servicio por cada servidor (notemos que dos servidores no necesariamente deben ofrecer en el mismo tiempo).

- *Cola:* La cola, propiamente dicha, es el conjunto de clientes que están esperando y ya han solicitado el servicio pero aún no están siendo atendidos por ningún servidor.

- *Sistema:* Es el conjunto formado por la cola, su disciplina, y el mecanismo de servicio.

Todos estos elementos forman el modelo de la cola. Es habitual usar indistintamente los términos cola, sistema de cola, y modelo de cola, siempre que se entienda por el contexto que no nos referimos a la componente llamada cola del sistema de cola (es decir, a los clientes que están esperando). Podemos observar una representación esquemática de un sistema de cola en la figura 3.1.

Además, podemos resumir el funcionamiento de la cola en el siguiente proceso: un cliente llega al sistema de cola. Si hay algún servidor libre su servicio empieza inmediatamente, en caso contrario se une a la cola. En este caso, cuando un servidor quede libre y el cliente sea seleccionado para ser servido, empezará su servicio. Cuando haya acabado su servicio, el cliente sale del sistema.

Notemos que en la descripción anterior hemos considerado una cola aislada, pero más adelante hablaremos de redes de colas en el capítulo 3.

3.2 | Notación

El primer paso hacia un análisis más exhaustivo de las colas es el de establecer la notación que usaremos para el análisis de las colas. Usaremos una notación parecida a la que hemos visto en los procesos de Markov.

- $N(t)$: número de clientes en el sistema en el instante t.
- $N_q(t)$: número de clientes en la cola en el instante t.
- $N_s(t)$: número de clientes siendo atendidos en el instante t.
- $p_n(t)$: probabilidad de que en el instante t se encuentren n clientes en el sistema.
- s: número de servidores en el mecanismo de servicio.
- λ_n: número medio de llegadas de clientes al sistema por unidad de tiempo cuando ya hay n clientes en el sistema. También llamada tasa de llegada. Se denota simplemente λ si no depende de n. En caso de que no sea constante, podemos usar $\bar{\lambda} = E[\lambda]$.
- μ_n: número medio de clientes que salen del sistema con el servicio completado por unidad de tiempo cuando ya hay n clientes en el sistema. También llamada tasa de servicio. Denotamos por μ el número medio de clientes que se pueden atender por unidad de tiempo si los tiempos de servicio de distintos servidores tienen la misma distribución.
- ρ: constante de utilización del sistema. También llamada intensidad de tráfico. Se define como:

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

- $\bar{\rho}$: constante de utilización real del sistema. También llamada intensidad real de tráfico. Se usa cuando $\lambda \neq \bar{\lambda}$, y se define como:

$$\bar{\rho} = \frac{\bar{\lambda}}{s\mu}$$

Es usual tomar $p_0(0) = 1$, es decir, asumir que el sistema no tiene clientes en el inicio de tiempos. También es útil analizar $p_0(t)$, ya que esta variable cuantifica la probabilidad de que el sistema esté parado en un tiempo t, lo que permite cuantificar el rendimiento del sistema (además de ρ).

Además, suponiendo que los tiempos de servicio tienen distribuciones idénticas, podemos escribir simplemente $\mu_n = n\mu$ cuando $n < s$ y $\mu_n = s\mu$ cuando $n \geq s$.

Por último, ρ representa los recursos del sistema consumidos por los clientes. De esta manera, tiene sentido pedir que $\rho < 1$. Cuanto más bajo sea ρ, menos habrán de esperar los clientes en general, y más tiempo libre tendrán los servidores.

Estudiaremos modelos estacionarios, de manera que las distribuciones de $N(t), N_q(t)$ y $N_s(t)$ no cambian con t. De esta manera, podemos definir:

- N: número de clientes en el sistema.
- N_q: número de clientes en la cola.
- N_s: número de clientes siendo servidos.
- p_n: probabilidad de que se encuentren n clientes en el sistema.
- L: número medio de clientes en el sistema.
- L_q: número medio de clientes en la cola.
- L_s: número medio de clientes siendo servidos.
- $\mathcal{W}$: tiempo que un cliente pasa en el sistema.
- $\mathcal{W}_q$: tiempo que un cliente espera en la cola.
- $\mathcal{W}_s$: tiempo que un cliente está siendo servido.
- W: tiempo medio que un cliente está en el sistema.
- W_q: tiempo medio que un cliente espera en la cola.
- W_s: tiempo miedo que un cliente está siendo servido.

De igual manera que antes, es muy útil obtener p_0 para analizar el rendimiento del sistema, puesto que p_0 es la probabilidad de que el sistema esté parado, y es un indicador claro del uso del sistema..

Claramente, las medias de clientes en el sistema en cada uno de sus estados esta relacionado con las variables que definen los números de clientes en cada uno de sus estados, de manera que:

$$L = E[N] \qquad L_q = E[N_q] \qquad L_s = E[N_s]$$

De igual manera, los tiempos medios en cada uno de los estados están relacionados con los tiempos de los clientes, con:

$$W = E[\mathcal{W}] \qquad W_q = E[\mathcal{W}_q] \qquad W_s = E[\mathcal{W}_s]$$

Además, claramente todas estas variables están relacionadas entre ellas de la manera siguiente:

$$\begin{cases} N = N_q + N_s \\ L = L_q + L_s \\ \mathcal{W} = \mathcal{W}_q + \mathcal{W}_s \\ W = W_q + W_s \end{cases}$$

Por último, el tiempo medio que un cliente está siendo servido es la inversa de μ, ya que μ era el número medio de servicios por unidad de tiempo, es decir:

$$W_s = E[\mathcal{W}_s] = \frac{1}{\mu}$$

Notemos que cuando no ha transcurrido mucho tiempo en el sistema, el número de clientes y los servicios dependen fuertemente de las condiciones iniciales. Sin embargo, dejando pasar el tiempo, alcanzamos lo que se llama *estado estacionario*, cuando la influencia de las condiciones iniciales es casi inexistente. Tomaremos todas estas variables cuando el sistema haya alcanzado su estado estacionario.

Estos últimos parámetros definidos son relevantes para describir las colas, pero su importancia depende del caso en el que se apliquen. Por ejemplo, supongamos una compañía de transportes, y la cola del taller en el que se efectúa el mantenimiento de los camiones de la compañía. En este contexto, L indica el número medio de camiones que no se pueden usar para el transporte, y W el tiempo medio en el que un camión está inoperativo. Por otra parte, en una oficina de correos, nos interesará conocer la satisfacción del cliente en base al tiempo medio de espera en la cola descrito por W_q, y también podemos aprovechar el número medio de clientes en espera L_q para diseñar una sala de espera más o menos grande.

En general, los estimadores más usados para analizar el rendimiento de un sistema son los siguientes:

- W: tiempo medio que pasa un cliente en el sistema.

- W_q: tiempo medio que pasa un cliente en la cola.
- L: número medio de clientes en el sistema.
- L_q: número medio de clientes en la cola.
- p_n: probabilidad de que hayan n clientes en el sistema.
- ρ: intensidad de tráfico del sistema.
- $\bar{\rho}$: intensidad real de tráfico del sistema.

3.2.1 | Nomenclatura de Kendall

Una vez hemos definido la notación a usar en las colas, clasificamos los posibles sistemas de colas con las características que determinan los elementos que componen el sistema.

La nomenclatura usada en la actualidad es una extensión de la introducida en 1953 por Kendall ([Ken53]), consistente en designar el sistema de una cola con la nomenclatura $A/B/s/K/H/Z$, donde:

- A: es la distribución del tiempo entre llegadas. Se usan abreviaturas para las distribuciones de estos tiempos entre llegadas. Las más comunes son:
  - D: determinista.
  - E_k: Erlang de parámetro k (es decir, suma de k exponenciales).
  - G: distribución genérica.
  - Γ: gamma.
  - M: exponencial (es decir, proviene de un proceso de Poisson).
  - M^X: exponencial, con X variable aleatoria que describe el número de llegadas por grupo.
  - U: uniforme.
- B: distribución de los tiempos de servicio. Se usan las mismas abreviaturas que hemos presentado en A.
- s: número de servidores en el sistema. Puede ser un número entero positivo o bien infinito.

- K: cantidad máxima de clientes en el sistema de la cola. También es un número entero positivo o infinito. Este valor se puede omitir, considerando así que $K = \infty$.
- H: tamaño de la población. Como hemos dicho anteriormente, también puede ser finito o infinito, siendo este último el valor que se toma por defecto.
- Z: disciplina de la cola. Es común usar las abreviaturas explicadas anteriormente, siendo su valor por defecto *FIFO*.

El sistema que Kendall introdujo únicamente usaba la notación $A/B/s$, usando los valores por defecto para K, H y Z. Este tipo de abreviatura solo se puede realizar si todos los parámetros son iguales a los valores por defecto a partir de uno dado. Por tanto, el modelo $M/M/5/\infty/3/FIFO$ se puede abreviar como $M/M/5/\infty/3$, pero no como $M/M/5/3/FIFO$.

3.3 | Propiedades

En los modelos de cola con distribuciones de tiempo exponencial tanto para las llegadas como para los servicios se verifican algunas propiedades importantes, como las fórmulas de Little ([Lit61]).

Teorema 3.1 (Primera fórmula de Little). *Supongamos que las tasas de llegadas son constantes. Entonces se tiene que:*

$$L = \lambda W \tag{3.3.1}$$

Es sencillo obtener esta fórmula. Supongamos que un cliente entra al sistema y sale un tiempo después, que de media es W. Como llegan una media de λ clientes por unidad de tiempo, en total habrán llegado λW clientes mientras nuestro cliente entra y sale del sistema. Por otra parte, el número medio de clientes en este periodo es precisamente L.

Este razonamiento se puede aplicar también a el intervalo en el que el cliente está esperando a ser servido, obteniendo más fórmulas de Little.

Teorema 3.2 (Fórmulas de Little). *Supongamos que las tasas de llegadas son constantes. Entonces se tiene que:*

$$L_q = \lambda W_q \tag{3.3.2}$$

$$L_s = \lambda W_s \tag{3.3.3}$$

Notamos que a partir de 3.3.2 y 3.3.3 podemos obtener de nuevo 3.3.1.

Estos resultados nos permiten obtener fácilmente unos estadísticos a partir de otros.

COROLARIO 3.3. *Si λ y W_s son conocidos para cualquier sistema de colas, podemos calcular L, L_q, W y W_q, suponiendo que alguno de ellos sea conocido.*

Demostración. Tenemos los distintos casos:

- Si conocemos L tenemos:

$$W = \frac{L}{\lambda} \qquad W_q = W - W_s \qquad L_q = \lambda W_q$$

- Si conocemos L_q tenemos:

$$L = L_q + \lambda W_s \qquad W = \frac{L}{\lambda} \qquad W_q = W - W_s$$

- Si conocemos W tenemos:

$$L = \lambda W \qquad W_q = W - W_s \qquad L_q = \lambda W_q$$

- Si conocemos W_q tenemos:

$$L_q = \lambda W_q \qquad W = W_q + W_s \qquad L = \lambda W$$

□

En el caso de que las tasas no sean constantes, existen resultados análogos.

TEOREMA 3.4 (Fórmulas de Little generalizadas). *Para cualquier sistema de cola, se tiene que:*

$$L = \bar{\lambda} W \tag{3.3.4}$$

$$L_q = \bar{\lambda} W_q \tag{3.3.5}$$

$$L_s = \bar{\lambda} W_s \tag{3.3.6}$$

En este caso, hemos considerado:

$$\bar{\lambda} = \sum_n \lambda_n p_n$$

Aparte de las fórmulas de Little, existen otras propiedades.

Proposición 3.5. *Para cualquier sistema de cola, se tiene que:*

$$W = W_q + \frac{1}{\mu} \tag{3.3.7}$$

La interpretación de este resultado es clara: el tiempo medio de un cliente en un sistema es precisamente la suma del tiempo medio en la cola y del tiempo en ser servido. Este último término es la inversa de μ, que representa el número medio de clientes atendidos por unidad de tiempo.

Por último, el siguiente resultado nos permite estimar el tiempo que pasará un cliente en el sistema a partir de los clientes que ya hay en el sistema a su llegada.

Proposición 3.6. *Supongamos que un cliente llega a un sistema $M/M/s$ cuando ya hay n clientes dentro del sistema, con $n > s$. Entonces el tiempo que pasará el cliente en la cola tiene distribución $\Gamma(s\mu, n - s + 1)$.*

Este resultado también se puede ver fácilmente ya que el tiempo que tardará en salir servido el próximo cliente tiene distribución $\exp(s\mu)$, y para que nuestro cliente entre en el proceso de servicio han de salir $n - s + 1$ clientes. La distribución gamma aparece como suma de exponenciales.

4 | Modelos Markovianos para colas

En el capítulo 2 hemos visto algunos tipos de procesos estocásticos que serán relevantes a la hora de modelizar los sistemas de colas definidos en el capítulo 3, así como técnicas para su estudio.

En este capítulo aprovecharemos las herramientas arriba mencionadas para analizar algunos modelos de colas. En la sección 4.1 estudiaremos los casos más simples, que se pueden expresar como procesos de Markov, analizando sus parámetros y obtendremos algunos indicadores de uso. Por otro lado, en la sección 4.2 analizaremos algunos modelos más complejos pero en los que podemos llegar a obtener un proceso de Markov que describa el sistema.

4.1 | Modelos Markovianos simples

4.1.1 | La cola $M/M/1$

En este modelo el tiempo entre llegadas es exponencial de parámetro λ, mientras que el tiempo de servicio es exponencial de parámetro μ, con un único servidor. Por defecto consideramos que la población potencial y la capacidad del sistema son infinitas, con disciplina *FIFO*. Es fácil ver que $\lambda_n = \lambda$ y $\mu_n = \mu$. Podemos ver este sistema de cola como un proceso de vida y muerte, y así tenemos que:

$$c_n = \prod_{k=1}^{n} \frac{\lambda_{k-1}}{\mu_k} = \prod_{k=1}^{n} \frac{\lambda}{\mu} = \left(\frac{\lambda}{\mu}\right)^n = \rho^n$$

Como $c_n = \rho^n$, la serie asociada es una geométrica, que será convergente si $\rho < 1$, de manera que el modelo de cola será estacionario si y solo si $\rho < 1$. Observemos que esto tiene sentido: si $\rho < 1$, entonces $\lambda < \mu$, es decir, el número de clientes que entran al sistema por unidad de tiempo debe ser menor que el número medio

de clientes atendidos por el servidor por unidad de tiempo para que el sistema funcione correctamente.

Para analizar este sistema, podemos suponer que $\rho < 1$ y por tanto el sistema es estacionario. Así, podemos obtener la suma de la serie de las c_n, que es:

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

De aquí podemos obtener la probabilidad de que el sistema esté en reposo:

$$p_0 = \left(\sum_{n=0}^{\infty} c_n \right)^{-1} = 1 - \rho$$

Y así podemos obtener la probabilidad de que haya n clientes en el sistema:

$$P(N = n) = p_n = p_0 c_n = (1-\rho)\rho^n$$

Por tanto, N tiene distribución geométrica de parámetro $1 - \rho$. Esto nos permite obtener su esperanza fácilmente, ya que la distribución geométrica es conocida:

$$L = E[N] = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$$

Ahora podemos usar el corolario 3.3 para obtener los parámetros restantes. Así, tenemos:

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

$$W_q = W - W_s = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$L_q = \lambda W_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

Ya hemos encontrado los parámetros usuales de las colas. Para tener aún más información sobre el proceso de espera de los clientes podemos calcular $\mathcal{W}$ y $\mathcal{W}_q$, y para ello aprovecharemos la proposición 3.6, que nos dice que $\mathcal{W}$ cuando $N = n$ tiene distribución $\Gamma(\mu, n+1)$. Así, tenemos:

$$
\begin{aligned}
P(\mathcal{W} \leq t) &= \sum_{n=0}^{\infty} P(\mathcal{W} \leq t|_{N=n}) P(N = n) \\
&= \sum_{n=0}^{\infty} \left(\int_0^t \frac{\mu^{n+1}}{n!} x^n e^{-\mu x} dx \right) \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n \\
&= \int_0^t \mu \left(1 - \frac{\lambda}{\mu}\right) e^{-\mu x} \left(\sum_{n=0}^{\infty} \frac{(\lambda x)^n}{n!} \right) dx \\
&= \int_0^t \mu \left(1 - \frac{\lambda}{\mu}\right) e^{-\mu x} e^{\lambda x} dx \\
&= \int_0^t (\mu - \lambda) e^{-(\mu-\lambda)x} dx \\
&= [-e^{-(\mu-\lambda)x}]_{x=0}^{x=t} \\
&= 1 - e^{-(\mu-\lambda)t}
\end{aligned}
$$

Esta función de distribución es la de una exponencial de parámetro $\mu - \lambda$, de manera que $\mathcal{W} \sim \exp(\mu - \lambda)$. Por último, obtenemos la distribución del tiempo que un cliente está en la cola. Para ello usamos la siguiente descomposición:

$$P(\mathcal{W}_q \leq t) = P(\mathcal{W}_q = 0) + P(0 < \mathcal{W}_q \leq t)$$

Aplicando a la segunda probabilidad el mismo procedimiento que hemos usado con $\mathcal{W}$ tenemos:

$$P(0 < \mathcal{W}_q \leq t) = 1 - \frac{\lambda}{\mu} e^{-(\mu-\lambda)t}$$

Por tanto, $\mathcal{W}$ es una mezcla de continua y discreta que toma el valor cero con probabilidad p_0 y con probabilidad $1 - p_0$ una exponencial de parámetro $\mu - \lambda$.

4.1.2 | La cola $M/M/s$

Ahora pasamos a una generalización del modelo anterior. En este caso tenemos s servidores, manteniendo las distribuciones entre llegadas y de servicio, así como los valores por defecto para el resto de parámetros. Las tasas de llegada siguen siendo constantes con $\lambda_n = \lambda$. Por otra parte, las tasas de servicio vienen dadas por:

$$\mu_n = \begin{cases} n\mu & n = 1, \ldots, s \\ s\mu & n = s+1, s+2, \ldots \end{cases}$$

De esta manera, tenemos:

$$c_n = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} & n = 1, \ldots, s \\ \dfrac{\lambda^n}{s!s^{n-s}\mu^n} & n = s+1, s+2, \ldots \end{cases}$$

Ahora podemos analizar cuándo el sistema será estacionario, es decir, la convergencia de la serie:

$$\sum_{n=0}^{\infty} c_n = \sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \sum_{n=s}^{\infty} \frac{\lambda^n}{s!s^{n-s}\mu^n} = \sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s}{s!\mu^s} \sum_{n=s}^{\infty} \rho^{n-s}$$

Esta serie es geométrica de razón ρ a partir del término s-ésimo, y por tanto será convergente siempre que $\rho < 1$, o lo que es lo mismo, cuando $\lambda < s\mu$. En este caso podemos sumar la serie y obtenemos:

$$\sum_{n=0}^{\infty} c_n = \sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s}{(s-1)!\mu^{s-1}(s\mu - \lambda)}$$

Por tanto, p_0 viene dado por:

$$p_0 = \left(\sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s}{(s-1)!\mu^{s-1}(s\mu - \lambda)} \right)^{-1}$$

De esta manera, podemos obtener las probabilidades:

$$p_n = c_n p_0 = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} p_0 & n = 1, \ldots, s \\ \dfrac{\lambda^s}{s!\mu^s} \rho^{n-s} p_0 & n = s+1, s+2, \ldots \end{cases}$$

Conocidas estas expresiones, podemos encontrar el valor de L_q:

$$
\begin{aligned}
L_q &= E[N_q] \\
&= \sum_{n=0}^{s} 0p_n + \sum_{n=s+1}^{\infty} (n-s)p_n \\
&= \sum_{n=s+1}^{\infty} (n-s)p_n \\
&= \sum_{n=s}^{\infty} (n-s)\frac{\lambda^n}{s!s^{n-s}\mu^n}p_0 \\
&= \frac{\lambda^s}{s!\mu^s}p_0 \sum_{n=s}^{\infty} (n-s)\rho^{n-s} \\
&= \frac{\lambda^s}{s!\mu^s}p_0 \sum_{k=0}^{\infty} k\rho^k \\
&= \frac{\lambda^s p_0}{s!\mu^s}\frac{\rho}{(1-\rho)^2} \\
&= \frac{\lambda^{s+1}p_0}{(s-1)!\mu^{s-1}(s\mu-\lambda)^2}
\end{aligned}
$$

Ahora podemos obtener W_q, W y L mediante el corolario 3.3. Así, tenemos:

$$
\begin{aligned}
W_q &= \frac{\lambda^s p_0}{(s-1)!\mu^{s-1}(s\mu-\lambda)^2} \\
W &= \frac{\lambda^s p_0}{(s-1)!\mu^{s-1}(s\mu-\lambda)^2} + \frac{1}{\mu} \\
L &= \frac{\lambda^{s+1} p_0}{(s-1)!\mu^{s-1}(s\mu-\lambda)^2} + \frac{\lambda}{\mu}
\end{aligned}
$$

Una vez obtenidos los parámetros principales, de forma análoga al caso estudiado en la subsección 4.1.1, podemos obtener las distribuciones de los tiempos que un cliente está tanto en el sistema como en la cola. Obtenemos:

$$
P(\mathcal{W}_q \leq t) = 1 - \frac{\lambda^s p_0}{(s-1)!\mu^{s-1}(s\mu-\lambda)} e^{-(s\mu-\lambda)t}
$$

Para obtener la distribución de $\mathcal{W}$ necesitamos distinguir dos casos:

$$
P(\mathcal{W} \leq t) = \begin{cases} 1 + \frac{\lambda - s\mu + \mu P(\mathcal{W}_q = 0)}{s\mu - \lambda - \mu} e^{-\mu t} + \frac{\lambda^s p_0}{(s-1)!\mu^{s-2}(s\mu-\lambda)(s\mu-\lambda-\mu)} e^{-(s\mu-\lambda)t} & \frac{\lambda}{\mu} \neq s-1 \\ 1 - \left(1 + \frac{\lambda^s p_0 t}{(s-1)!\mu^{s-2}(s\mu-\lambda)}\right) e^{-\mu t} & \frac{\lambda}{\mu} = s-1 \end{cases}
$$

Por último, observamos un resultado intuitivo:

$$P(\mathcal{W}_q = 0) = P(\mathcal{W}_q \leq 0) = 1 - \sum_{n=s}^{\infty} p_n = \sum_{n=0}^{s-1} p_n = P(N \leq s-1)$$

Este resultado es fácil de entender ya que la probabilidad de que un cliente no tenga que esperar en la cola y pase a ser servido directamente es precisamente la probabilidad de que haya algún servidor libre a la llegada del cliente.

4.1.3 | La cola $M/M/\infty$

Este modelo considera infinitos servidores y distribuciones exponenciales de tiempos entre llegadas y de servicio exponenciales. Como hay infinitos servidores, los parámetros de disciplina y de capacidad son irrelevantes. Este tipo de modelos tienen sentido cuando se considera que cada cliente puede darse servicio automáticamente, es decir, es un modelo de autoservicio.

En esta cola podemos obtener algunos parámetros sin ningún cálculo. Es de esperar que N_q y $\mathcal{W}_q$ tomen valor nulo con probabilidad 1, y como consecuencia $L_q = W_q = 0$. Por tanto, tenemos que $\mathcal{W} = \mathcal{W}_s$ y así $\mathcal{W}$ es una exponencial de parámetro μ. De aquí se deduce que $W = \frac{1}{\mu}$ y $L = \frac{\lambda}{\mu}$.

Analizamos este modelo un poco más detalladamente. Las tasas de llegada son $\lambda_n = \lambda$ y las de servicio son $\mu_n = n\mu$, por lo que tenemos:

$$c_m = \frac{\lambda^n}{n!\mu^n}$$

Por tanto podemos analizar la serie asociada:

$$\sum_{n=0}^{\infty} c_n = \sum_{n=0}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} = e^{\frac{\lambda}{\mu}}$$

Esta serie es siempre convergente, así que el modelo alcanza siempre su estado estacionario. Esto tiene sentido ya que la intensidad de tráfico es nula: la capacidad de servicio es infinitamente mayor que la llegada de clientes al sistema. Además, podemos obtener fácilmente las probabilidades asociadas:

$$p_0 = \left(\sum_{n=0}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}\right)^{-1} = \frac{1}{e^{\frac{\lambda}{\mu}}} = e^{-\frac{\lambda}{\mu}}$$

$$p_n = c_n p_0 = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} e^{-\frac{\lambda}{\mu}}$$

De aquí se deduce que N tiene por distribución una Poisson de parámetro $\frac{\lambda}{\mu}$.

4.1.4 | La cola $M/M/1/K$

Este modelo es similar al estudiado en la Subsección 4.1.1, con el cambio de que puede haber un máximo de K clientes en todo el sistema. Esto implica que un cliente puede no conseguir entrar en el sistema si este está lleno.

Podemos deducir fácilmente las tasas de llegada:

$$\lambda_n = \begin{cases} \lambda & n = 0, \dots, K \\ 0 & n = K+1, K+2, \dots \end{cases}$$

Por otra parte, $\mu_n = \mu$ en general, con lo que podemos obtener inmediatamente los c_n:

$$c_n = \begin{cases} \rho^n & n = 0, \dots, K+1 \\ 0 & n = K+2, K+3, \dots \end{cases}$$

Como estos términos son nulos a partir de cierto término, la serie asociada es convergente sin ningún tipo de condición sobre ρ como en los casos anteriores. De hecho, esto se puede interpretar como que la propia limitación del tamaño de la cola fuerza a la estacionareidad, incluso en el caso en el que el sistema está saturado prácticamente todo el tiempo. La suma de la serie es una suma finita y vale:

$$\sum_{n=0}^{\infty} c_n = \sum_{n=0}^{K+1} \rho^n = \begin{cases} \dfrac{\rho^{K+2} - 1}{\rho - 1} & \rho \neq 1 \\ K+2 & \rho = 1 \end{cases}$$

Es necesario hacer esta misma distinción en todos los cálculos.

Empezamos suponiendo que $\rho \neq 1$. Podemos obtener el valor de p_0 a partir de la suma de la serie:

$$p_0 = \left(\sum_{n=0}^{\infty} c_n \right)^{-1} = \frac{\rho - 1}{\rho^{K+2} - 1}$$

De aquí podemos obtener fácilmente las p_n:

$$p_n = \begin{cases} \dfrac{\rho - 1}{\rho^{K+2} - 1} \rho^n & n = 0, \dots, K+1 \\ 0 & n = K+2, K+3, \dots \end{cases}$$

A partir de aquí podemos calcular el número medio de clientes en el sistema:

$$L = \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{K+1} n \frac{\rho - 1}{\rho^{K+2} - 1} \rho^n = \frac{\rho - 1}{\rho^{K+2} - 1} \sum_{n=0}^{K+1} n \rho^{n-1} =$$

$$= \frac{\rho - 1}{\rho^{K+2} - 1} \frac{(K+1)\rho^{K+2} - (K+2)\rho^{K+1} + 1}{(\rho - 1)^2} = \frac{\rho}{1-\rho} - \frac{(K+2)\rho^{K+2}}{1 - \rho^{K+2}}$$

Observemos que el primer sumando es precisamente el valor de L obtenido en la Subsección 4.1.1 para un sistema $M/M/1$, mientras que el segundo valor tiende a 0 cuando K crece, de manera que cuando la cola tiene una capacidad muy grande en comparación con las tasas de llegada, el comportamiento es prácticamente el de una $M/M/1$. Ahora aplicaremos las fórmulas de Little en su caso más general, por lo que calcularemos $\bar{\lambda}$:

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n p_n = \lambda \sum_{n=0}^{K} p_n = \lambda(1 - p_{K+1}) = \lambda \left(1 - \frac{\rho - 1}{\rho^{K+2} - 1} \rho^{K+1}\right) = \frac{\lambda(\rho^{K+1} - 1)}{\rho^{K+2} - 1}$$

Ya podemos aplicar las fórmulas de Little, y obtenemos:

$$W = \frac{L}{\bar{\lambda}} = \frac{1}{\mu - \lambda} - \frac{(K+1)\rho^{K+2}}{\lambda(1 - \rho^{K+1})}$$

Ahora teniendo en cuenta que $W_q = W - \frac{1}{\mu}$, tenemos:

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} - \frac{(K+1)\rho^{K+2}}{\lambda(1 - \rho^{K+1})}$$

Por último:

$$L_q = \bar{\lambda} W_q = \frac{\rho^2}{1-\rho} - \frac{(K+1+\rho)\rho^{K+2}}{1 - \rho^{K+2}}$$

Ahora consideramos el caso más fácil en el que ρ es precisamente 1. En este caso p_0 viene dado por:

$$p_0 = \left(\sum_{n=0}^{\infty} c_n\right)^{-1} = \frac{1}{K+2}$$

Además $c_n = \rho^n = 1$, de manera que:

$$p_n = \frac{1}{K+2} \qquad n = 0, \ldots, K+1$$

Por tanto, N tiene una distribución uniforme discreta en los valores posibles. Así podemos obtener fácilmente los demás parámetros de la cola:

$$L = \sum_{n=0}^{K+1} n p_n = \frac{1}{K+2} \sum_{n=0}^{K+1} n = \frac{(K+1)(K+2)}{2(K+2)} = \frac{K+1}{2}$$

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n p_n = \lambda \sum_{n=0}^{K} p_n = \lambda(1 - p_{K+1}) = \lambda \left(1 - \frac{1}{K+2}\right) = \frac{\lambda(K+1)}{K+2}$$

$$W = \frac{L}{\bar{\lambda}} = \frac{\frac{K+1}{2}}{\frac{\lambda(K+1)}{K+2}} = \frac{K+2}{2\lambda}$$

$$W_q = W - \frac{1}{\mu} = \frac{K+2}{2\lambda} - \frac{1}{\lambda} = \frac{K}{2\lambda}$$

$$L_q = \bar{\lambda} W_q = \frac{\lambda(K+1)}{K+2} \frac{K}{2\lambda} = \frac{K(K+1)}{2(K+2)}$$

Por último, para cualquier ρ, podemos obtener también las funciones de distribución de las variables $\mathcal{W}$ y $\mathcal{W}_q$, obteniendo:

$$P(\mathcal{W} \leq t) = 1 - e^{-\mu t} \sum_{n=0}^{K} \frac{p_n}{1 - p_{K+1}} \sum_{m=0}^{n} \frac{(\mu t)^m}{m!}$$

$$P(\mathcal{W}_q \leq t) = 1 - e^{-\mu t} \sum_{n=1}^{K} \frac{p_n}{1 - p_{K+1}} \sum_{m=0}^{n-1} \frac{(\mu t)^m}{m!}$$

Es necesario notar que en este modelo ρ no refleja el sentido habitual de la intensidad de tráfico, ya que el numerador λ representa el número medio de *intentos* de llegada. Por tanto, podemos aprovechar $\bar{\lambda}$ para representar el número medio de entradas del sistema, y así obtenemos la intensidad real de tráfico:

$$\bar{\rho} = \frac{\bar{\lambda}}{\mu} = \begin{cases} \frac{\frac{\lambda(\rho^{K+1}-1)}{\rho^{K+2}-1}}{\mu} = \frac{\rho^{K+2} - \rho}{\rho^{K+2} - 1} & \rho \neq 1 \\ \frac{\frac{\lambda(K+1)}{K+2}}{\mu} = \frac{K+1}{K+2} & \rho = 1 \end{cases}$$

Notamos que $\bar{\rho}$ sí es menor que 1 y refleja la verdadera intensidad del tráfico.

4.1.5 | La cola $M/M/s/K$

Ahora consideramos una extensión del modelo anterior. Las tasas de llegada y de servicio son:

$$\lambda_n = \begin{cases} \lambda & n = 0, \ldots, K+s-1 \\ 0 & n = K+s, K+s+1, \ldots \end{cases}$$

$$\mu_n = \begin{cases} n\mu & n = 1, \ldots, s \\ s\mu & n = s+1, s+2, \ldots \end{cases}$$

A partir de estas tasas de llegada y servicio podemos obtener:

$$c_n = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} & n = 1, \dots, s \\ \dfrac{\lambda^n}{s!s^{n-s}\mu^n} & n = s+1, \dots, K+s \\ 0 & n = K+s+1, K+s+2, \dots \end{cases}$$

La serie asociada es de nuevo convergente al tener un número finito de términos no nulos. Por tanto, el sistema es estacionario independientemente del valor de ρ. Esta suma es:

$$\sum_{n=0}^{\infty} c_n = \sum_{n=0}^{K+s} c_n = \sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \sum_{n=s}^{K+s} \frac{\lambda^n}{s!s^{n-s}\mu^n} = \sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s}{s!\mu^s} \sum_{n=s}^{K+s} \rho^{n-s}$$

$$= \begin{cases} \displaystyle\sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s}{s!\mu^s} \frac{1-\rho^{K+1}}{1-\rho} & \rho \neq 1 \\ \displaystyle\sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s(K+1)}{s!\mu^s} & \rho = 1 \end{cases}$$

De aquí podemos obtener p_0:

$$p_0 = \begin{cases} \dfrac{1}{\displaystyle\sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s}{s!\mu^s} \frac{1-\rho^{K+1}}{1-\rho}} & \rho \neq 1 \\ \dfrac{1}{\displaystyle\sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s(K+1)}{s!\mu^s}} & \rho = 1 \end{cases}$$

Así podemos obtener las probabilidades del sistema:

$$p_n = c_n p_0 = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} p_0 & n = 1, \dots, s \\ \dfrac{\lambda^s}{s!\mu^s} \rho^{n-s} p_0 & n = s+1, \dots, K+s \end{cases}$$

Es fácil ver que si $\rho = 1$, las probabilidades de los estados en los que hay de $s-1$ a $K+s$ clientes son iguales, y por tanto tenemos una distribución uniforme discreta en este subconjunto de estados.

Ahora obtenemos los parámetros principales de la cola, empezando por L_q:

$$L_q = \begin{cases} \dfrac{\lambda^s p_0 \rho \left(1 + K\rho^{K+1} - (K+1)\rho^K\right)}{s!\mu^s(1-\rho)^2} & \rho \neq 1 \\ \dfrac{\lambda^{s-1} K(K+1) p_0}{(s-1)!2\mu^{s-1}} & \rho = 1 \end{cases}$$

Para poder usar las fórmulas de Little debemos obtener $\bar{\lambda}$:

$$\bar{\lambda} = \lambda \left(1 - \frac{p_0 \rho^K \lambda^s}{s! \mu^s}\right)$$

De nuevo, podemos obtener la intensidad real de tráfico:

$$\bar{\rho} = \frac{\bar{\lambda}}{s\mu} = \rho - \frac{\lambda^s}{s! \mu^s} \rho^{K+1} p_0$$

A partir de los valores de L_q y $\bar{\lambda}$ podemos obtener los otros parámetros:

$$W_q = \frac{L_q}{\bar{\lambda}}$$

$$W = W_q + \frac{1}{\mu}$$

$$L = \bar{\lambda} W$$

Por último podemos obtener la distribución del tiempo que un cliente pasa en cola:

$$P(\mathcal{W}_q \leq t) = 1 - e^{-s\mu t} \sum_{n=s}^{K+s-1} \frac{p_n}{1 - p_{K+s}} \sum_{m=0}^{n-s} \frac{(s\mu t)^m}{m!}$$

Es particularmente relevante un caso concreto de este modelo, el de $K = 0$. Este modelo es uno de los primeros que estudió Erlang, representando una central telefónica de s líneas. La capacidad es $K = 0$ porque cuando las líneas están ocupadas la llamada no se queda en espera, sino que recibe la señal que indica que la red está saturada. En este caso se tiene que:

$$p_n = \frac{\frac{\lambda^n}{n! \mu^n}}{\sum_{j=0}^{s} \frac{\lambda^j}{j! \mu^j}} \qquad n = 0, \ldots, s$$

Esta fórmula se debe a Erlang, que también probó que es válida para cualquier distribución del tiempo de servicio, entendiendo μ en ese caso como la inversa de la media de la distribución.

4.1.6 | La cola $M/M/1/\infty/H$

Añadimos una consideración más a los modelos estudiados, ahora con el caso de que la población es finita y formada por H individuos. Ahora el tiempo entre

llegadas depende de cuántos clientes estén en el sistema, siendo una exponencial de parámetro $\lambda(H-n)$ en el caso en el que hay n clientes en el sistema. Esta suposición de que la distribución es exponencial permite también modelizar este sistema como un proceso de vida y muerte.

Las tasas de servicio y de llegada son:

$$\lambda_n = \begin{cases} \lambda(H-n) & n = 0, \dots, H \\ 0 & n = H, H+1, \dots \end{cases}$$

$$\mu_n = \mu \qquad n = 1, 2, \dots$$

De aquí obtenemos c_n como:

$$c_n = \begin{cases} \frac{H!}{(H-n)!}\rho^n & n = 1, \dots, H \\ 0 & n = H+1, H+2, \dots \end{cases}$$

Como en los últimos casos estudiados, la serie es convergente por ser nula a partir de cierto término. Así, el modelo es siempre estacionario. Sin embargo, no podemos obtener fácilmente p_0 por la dificultad de sumar los términos c_n. Así, simplemente:

$$p_0 = \left(\sum_{n=0}^{H} c_n \right)^{-1}$$

De igual manera, no es fácil obtener una fórmula explícita para L ni L_q, debiendo calcular directamente:

$$L = \sum_{n=0}^{H} n p_n$$

Podemos obtener así $\bar{\lambda}$:

$$\bar{\lambda} = \lambda(H - L)$$

De aquí podemos obtener la intensidad real de tráfico:

$$\bar{\rho} = \rho(H - L)$$

Con estos parámetros, podemos obtener:

$$W = \frac{L}{\bar{\lambda}} = \frac{L}{\lambda(H-L)}$$

$$W_q = W - \frac{1}{\mu} = \frac{L\mu + \lambda(L-H)}{\lambda\mu(H-L)}$$

$$L_q = \bar{\lambda} W_q = \frac{\lambda(L-H) + L\mu}{\mu}$$

Por último podemos obtener las distribuciones de $\mathcal{W}$ y $\mathcal{W}_q$:

$$P(\mathcal{W} \leq t) = 1 - e^{-\mu t} \sum_{n=0}^{H-1} \frac{(H-n)p_n}{H-L} \sum_{m=0}^{n} \frac{(\mu t)^m}{m!}$$

$$P(\mathcal{W}_q \leq t) = 1 - e^{-\mu t} \sum_{n=0}^{H-1} \frac{(H-n)p_n}{H-L} \sum_{m=0}^{n-1} \frac{(\mu t)^m}{m!}$$

4.1.7 | La cola $M/M/s/\infty/H$

Este caso generaliza el caso anterior con s servidores. Consideraremos $s \leq H$, ya que el caso $s > H$ puede resolverse usando el caso $s = H$, además de ser el caso $s \leq H$ el más realista. Los valores de L, L_q, W, W_q y las p_n son las mismas para los modelos $s = H$ y $s > H$, siendo únicamente necesario reajustar algunos valores como el número medio de servidores ocupados.

En este caso las tasas de llegada son las que hemos visto en el caso anterior:

$$\lambda_n = \begin{cases} \lambda(H-n) & n = 0, \ldots, H \\ 0 & n = H, H+1, \ldots \end{cases}$$

Por otra parte, las tasas de servicio son las que hemos visto en las subsecciones 4.1.2 y 4.1.5, que son:

$$\mu_n = \begin{cases} n\mu & n = 1, \ldots, s \\ s\mu & n = s, s+1, \ldots \end{cases}$$

La expresión de las c_n es:

$$c_n = \begin{cases} \frac{H!}{(H-n)!} \frac{\lambda^n}{n!\mu^n} & n = 1, \ldots, s \\ \frac{H!}{(H-n)!} \frac{\lambda^s}{s!\mu^s} \rho^{n-s} & n = s, \ldots, H \\ 0 & n = H+1, H+2, \ldots \end{cases}$$

Igual que en el caso anterior, la serie es convergente por ser nula a partir de cierto término, y el modelo es siempre estacionario. De nuevo no podemos obtener fácilmente p_0 por la dificultad de sumar los términos c_n. Así:

$$p_0 = \left(\sum_{n=0}^{H} c_n \right)^{-1}$$

De nuevo:

$$L = \sum_{n=0}^{H} np_n$$

$$\bar{\lambda} = \lambda(H - L)$$

$$\bar{\rho} = \rho(H - L)$$

$$W = \frac{L}{\bar{\lambda}} = \frac{L}{\lambda(H - L)}$$

$$W_q = W - \frac{1}{\mu} = \frac{L\mu + \lambda(L - H)}{\lambda\mu(H - L)}$$

$$L_q = \bar{\lambda}W_q = \frac{\lambda(L - H) + L\mu}{\mu}$$

Por último podemos obtener la distribución de $\mathcal{W}$:

$$P(\mathcal{W}_q \leq t) = 1 - e^{-\mu t} \sum_{n=0}^{H-1} \frac{(H - n)p_n}{H - L} \sum_{m=0}^{n-s} \frac{(\mu t)^m}{m!}$$

4.2 | Modelos Markovianos extendidos

En los sistemas de colas que hemos visto en la sección 4.1 no había características como llegadas en grupo, servicio en grupo, prioridades y demás, que proporcionan modelos más complicados. En esta sección introducimos algunos que aún se pueden modelizar transformándolos en procesos de Markov. Aunque existen más modelos de los que presentamos aquí que coinciden con esta descripción, estos dos son los más sencillos y más fáciles de tratar, siendo de los únicos de los que podemos obtener algunos parámetros explícitos. El lector interesado podrá encontrar más modelos así como algunas técnicas para analizarlos en [Bha15].

4.2.1 | La cola $M^X/M/1$

En algunas situaciones es habitual que los clientes lleguen en grupos. Consideramos que los clientes llegan en grupos de tamaño descrito por una variable aleatoria X que toma valores naturales. Los tiempos entre llegadas de grupos son independientes de los tamaños de los grupos, y se distribuyen según una exponencial de parámetro λ.

Denotemos $d_k = P(X = k)$, es decir, d_k la probabilidad de que lleguen k clientes en un mismo grupo. Por tanto, el proceso que modeliza la cola es un proceso de Markov similar a un proceso de vida o muerte, con la diferencia de que podemos saltar entre estados que no son contiguos. Este hecho es el que nos impide obtener directamente las probabilidades asociadas a la cola.

Sin embargo, podemos obtener algunos resultados usando funciones generatrices. Consideramos:

$$P(x) = \sum_{n=0}^{\infty} p_n x^n \qquad D(x) = \sum_{k=0}^{\infty} d_k x^k$$

A partir de estas dos funciones y usando las ecuaciones de equilibrio podemos obtener:

$$(\lambda + \mu)P(x) - \mu p_0 = \lambda D(x)P(x) + \frac{\mu}{x}\left(P(x) - p_0\right))$$

Esto se puede simplificar como:

$$P(x) = \frac{\mu p_0(1-x)}{\mu(1-x) - \lambda x(1 - D(x))}$$

Observemos que $P(1) = \sum_{n=0}^{\infty} p_n = 1$, y así podemos obtener que:

$$p_0 = 1 - \frac{\lambda D'(1)}{\mu}$$

Denotamos $D'(1) = E[X] = d$. Observamos que la intensidad de tráfico es precisamente:

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

De esta manera, tenemos que:

$$p_0 = 1 - \rho$$

Aunque la distribución de X sea sencilla, no es fácil obtener las p_n. Aún así, sí podemos obtener L con relativa sencillez:

$$L = P'(1) = \frac{2\rho + \frac{\lambda}{\mu} D''(1)}{2(1-\rho)}$$

Como $D''(1) = E[X^2] - E[X]$, podemos ver fácilmente que:

$$L = P'(1) = \frac{\rho + \frac{\lambda}{\mu} E[X^2]}{2(1-\rho)}$$

$$W = \frac{\mu + \lambda E[X^2]}{2\mu^2(1-\rho)}$$

$$W_q = W - \frac{1}{\mu} = \frac{\mu(1-2\rho) - \lambda E[X^2]}{2\mu^2(\rho-1)}$$

$$L_q = \lambda d W_q = \frac{\rho(\mu(1-2\rho) - \lambda E[X^2])}{2\mu(\rho-1)}$$

Por último, cuando los clientes llegan en grupos de tamaño constante d, el valor de L se simplifica:

$$L = \left(\frac{d+1}{2}\right)\frac{\rho}{1-\rho}$$

$$W = \frac{d+1}{2\mu(1-\rho)}$$

$$W_q = \frac{d+1-2(1-\rho)}{2\mu(1-\rho)}$$

$$L_q = \frac{\rho(d+1)}{2(1-\rho)} - \rho$$

4.2.2 | La cola $M/M^X/1$

En este modelo los clientes llegan de uno en uno como hemos estudiado anteriormente, pero pueden ser servidos en grupos de tamaño X. Para estudiar este tipo de colas suponemos que X es una constante d, y que el servidor espera a que haya d clientes para servirlos a todos. El lector interesado puede consultar [GH98] para estudiar el caso en el que el servicio empieza con menos clientes y los clientes se pueden ir añadiendo al servicio.

De manera análoga a la anterior, podemos obtener la función generatriz $P(x)$:

$$P(x) = \frac{(1-x^d)\sum_{n=0}^{d-1} p_n x^n}{\frac{\lambda}{\mu}x^{d+1} - \left(\frac{\lambda}{\mu}+1\right)x^d + 1}$$

Mediante argumentos analíticos, se puede ver que el denominador tiene exactamente una raíz x_0 de módulo mayor que uno. Esta raíz nos permite obtener las probabilidades p_n:

$$p_n = \begin{cases} \dfrac{x_0^{n+1}-1}{dx_0^{n+1}} & n < d \\ \dfrac{x_0^{d}-1}{dx_0^{n+1}} & n \geq d \end{cases}$$

Para obtener dicha raíz existen algunos algoritmos, que el lector interesado podrá encontrar en [CT83].

5 | Redes de colas

A lo largo del capítulo 3 y del análisis realizado en el capítulo 4 hemos usado modelos simples de colas, es decir, modelos en los que había una cola de espera formada por clientes esperando a ser servidos por uno o varios servidores englobados en un mismo tipo de servicio.

En este capítulo presentaremos algunos modelos en los que los clientes deben pasar por varios tipos de servicio de manera secuencial. Primero veremos en las secciones 5.2 y 5.3 un tipo concreto de redes llamadas redes de Jackson en sus versiones abierta y cerrada respectivamente. Aprovecharemos estos tipos de redes para después analizar más extensivamente algunos modelos comunes de rededs. En la sección 5.4 veremos el caso en el que simplemente los clientes pasan de manera secuencial de una cola a otra. En la sección 5.5 veremos el caso en el que los clientes pueden entrar en un ciclo, pudiendo así volver a una cola en la que ya han sido atendidos.

5.1 | Introducción

Una *red de colas* no es más que un conjunto de colas enlazadas entre ellas, de manera que un cliente pasará de una a otra en acabar el servicio correspondiente.

Normalmente representamos en un grafo las diferentes colas enlazadas en una red, donde los nodos representan las colas y las aristas representan las posibles rutas entre las diferentes colas. Suponemos además que la proporción de clientes que se mueven de una cola a otra viene dada por una distribución de probabilidad discreta.

En la figura 5.1 podemos observar un ejemplo de red de colas. En esta red, los clientes entran al sistema en las colas 1 y 4, y salen por las colas 3 y 5.

Las redes como la anterior se conocen como *redes abiertas*, donde se producen llegadas desde fuera de la red y/o salidas de clientes servidos hacia fuera de la red. Por otra parte, las *redes cerradas* son aquellas en las que no hay ni entradas

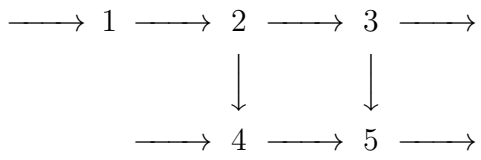

Figura 5.1: Ejemplo de una red de colas.

ni salidas fuera de la red, y por tanto el número de clientes es constante. Es relevante notar que cuando el número de entradas y el de salidas son iguales (o aproximadamente iguales) podemos modelizar las redes abiertas como redes cerradas sin perder mucha información en el modelo, y normalmente facilitando su análisis.

5.1.1 | Notación

Es habitual denotar por $1, \ldots, K$ los nodos que forman la red, o lo que es lo mismo, las etiquetas por las que denotamos cada una de las colas que forman la red. Aprovechamos esta notación para las transiciones de clientes: usamos p_{ij} para denotar la probabilidad de que un cliente que salga de la cola i pase a la cola j. En un abuso de notación, denotamos con el índice 0 el exterior de la red, denotando así los p_{0i} la probabilidad de que al entrar un cliente lo haga por la cola i y los p_{i0} la probabilidad de que un cliente abandone la cola al ser servido en el nodo i.

Por otra parte, denotamos por λ_i la tasa de llegada de clientes desde fuera del sistema a la cola i, y por μ_i la tasa de servicio de los servidores de la cola i. Denotaremos por $\lambda_T = \sum_{i=1}^{K}$ a la tasa de llegada del sistema. Además, Λ_i denotará la tasa de llegada de clientes a la cola i, contando aquí las llegadas tanto desde fuera del sistema como las de la propia red. Para evitar confusiones con λ_n y μ_n, que denotaban las tasas de llegadas y servicios cuando habían n clientes en el sistema, denotaremos estas por $\Lambda_i^{(n)}$ y $\mu_i^{(n)}$ cuando sea necesario.

Denotaremos por $p_i(n_i)$ la probabilidad de que en el nodo i haya n_i clientes, y por $p(n_1, \ldots, n_K)$ la probabilidad de que haya n_1 clientes en el nodo 1, n_2 clientes en el nodo 2, y así hasta el nodo K.

Por último, denotaremos por L_i, L_i^q, W_i, W_i^q los parámetros análogos a los vistos en la sección 3.2 cuando se refieren al nodo i, y por L_T, L_T^q, W_T, W_T^q los mismos parámetros cuando se refieren a toda la red.

5.2 | Redes abiertas de Jackson

Una *red de Jackson abierta* es una red de colas abierta que verifica las siguientes propiedades:

1. Cada nodo i tiene s_i servidores con los mismos tiempos exponenciales de servicio, con tasa de servicio μ_i.
2. Los clientes que llegan al nodo i desde fuera del sistema lo hacen según un proceso de Poisson de intensidad λ_i.
3. Cada cliente que sale servido del nodo i pasa instantáneamente al nodo j con probabilidad p_{ij} o bien sale del sistema con probabilidad p_{i0}.

Este tipo de redes es esencial en el estudio de redes de colas ya que abarcan una amplia cantidad de modelos comunes y/o relevantes. Además, son relativamente sencillos de analizar gracias al siguiente teorema debido a Jackson ([Jac04]), que permite obtener muy fácilmente las probabilidades de la red así como sus parámetros asociados.

Teorema 5.1 (Jackson). *Sea una red de Jackson abierta con K nodos. Se tiene que:*

$$\Lambda_j = \lambda_j + \sum_{i=1}^{K} \Lambda_i p_{ij} \qquad j = 1, \dots, K \tag{5.2.1}$$

Además, si $\rho_i < 1$ para todos los nodos, podemos obtener las probabilidades del sistema:

$$p(n_1, \dots, n_K) = \prod_{i=1}^{K} p_i(n_i)$$

Es decir, cada nodo se comporta como un $M/M/s_i$ independiente de los demás, con tasa de llegadas Λ_i y tasa de servicio μ_i.

Observemos que podemos escribir de forma matricial la relación anterior para las tasas de llegada:

$$\Lambda = \lambda + P^t \Lambda$$

Esto quiere decir que podemos obtener las tasas totales de llegada como:

$$\Lambda = (I - P^t)^{-1} \lambda$$

Conociendo las tasas totales de llegada Λ_i podemos obtener los valores relevantes de cada subsistema (L_i, L_i^q, W_i, W_i^q) considerando un modelo $M/M/s_i$ con tasa de

llegada Λ_i y tasa de servicio μ_i, como hemos visto en la subsección 4.1.2. Además, podemos obtener los parámetros del sistema a partir de los parámetros de cada una de las colas:

$$L_T = \sum_{i=1}^{K} L_i \qquad L_T^q = \sum_{i=1}^{K} L_i^q \qquad W_T = \frac{L_T}{\lambda_T} \qquad W_T^q = \frac{L_T^q}{\lambda_T}$$

Observamos también la facilidad que proporciona este teorema para obtener las probabilidades asociadas a la red, siempre que las respectivas colas sean estacionarias, hipótesis que ya considerábamos en el estudio de colas simples. Las probabilidades de cada una de las colas son independientes entre ellas y son precisamente las estudiadas y obtenidas en el capítulo 3, simplificando el proceso de obtención de las probabilidades de la red. Observemos que las probabilidades para un instante dado no son independientes entre ellas: si una cola está sobrecargada, es fácil que la siguiente esté vacía. Sin embargo, estas dependencias desaparecen al considerar la estacionariedad, es decir, considerar las colas para largos intervalos de tiempo.

5.3 | Redes cerradas de Jackson

El análisis de las redes de Jackson cerradas es especialmente sencillo: únicamente tenemos que conocer el número de clientes dentro de la red, N, que es constante al no haber entradas ni salidas. Por tanto, $L_T = N$, y otros parámetros como W_T y W_T^q no tienen sentido.

Como los parámetros usuales de una red no tienen sentido en este caso, lo único que podemos hacer en este tipo de redes es obtener las distintas probabilidades del sistema. Para ello, consideramos cualquier solución Λ del sistema 5.2.1 con $\lambda = 0$, es decir, de:

$$\Lambda = P^t \Lambda$$

A partir de aquí, se pueden obtener las distintas probabilidades:

$$p(n_1, \ldots, n_K) = \frac{1}{G(N)} \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{a_i(n_i)}$$

En este caso hemos considerado:

$$a_i(n) = \begin{cases} n! & n \leq s_i \\ s_i! s_i^{n-s_i} & n \geq s_i \end{cases}$$

$$G(N) = \sum_{n_1 + \cdots + n_K = N} \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{a_i(n_i)}$$

5.4 | Colas en serie

El caso más sencillo de red de colas es el de colas en serie, donde los clientes entran a la red por la primera cola y van pasando progresivamente por todas las demás, en el mismo orden. Llamamos *red en serie* a este tipo de red. Podemos ver este tipo de red en la figura 5.2.

$$\longrightarrow 1 \longrightarrow \ldots \longrightarrow K \longrightarrow$$

Figura 5.2: Red de colas en serie.

En este caso, los parámetros de la red son:

$$\begin{cases} \lambda_1 = \lambda \\ \lambda_i = 0 \quad i = 2, \ldots, K \end{cases}$$

$$\begin{cases} p_{ij} = \delta_{i+1,j} \\ p_{0j} = \delta_{1,j} \\ p_{i0} = \delta_{i,K} \end{cases}$$

Bajo estas condiciones, podemos usar el teorema de Jackson, planteando el sistema de la ecuación 5.2.1, con lo que obtenemos:

$$\begin{cases} \Lambda_1 = \lambda \\ \Lambda_2 = \Lambda_1 \\ \vdots \\ \Lambda_K = \Lambda_{K-1} \end{cases}$$

La solución del sistema es trivialmente:

$$\Lambda_1 = \Lambda_2 = \cdots = \Lambda_K = \lambda$$

Por tanto, cada cola del sistema se comporta como una $M/M/s_i$ independiente con tasa de entrada λ y tasa de servicio μ_i.

Es relevante notar que como el recorrido de un cliente es secuencial (no hay bucles ni bifurcaciones), el tiempo medio de un cliente en el sistema es la suma de los tiempos medios en cada una de las colas:

$$W_T = \frac{L_T}{\lambda_T} = \frac{\sum_{i=1}^{K} L_i}{\lambda} = \sum_{i=1}^{K} \frac{L_i}{\lambda} = \sum_{i=1}^{K} \frac{L_i}{\Lambda_i} = \sum_{i=1}^{K} W_i$$

5.5 | Colas cíclicas

Ahora consideramos una red de Jackson cerrada con N clientes y con K nodos, conectada de manera que cada cliente pasa del nodo i al $i+1$, y del K al 1. Es decir, cada cliente va dando vueltas por la red en bucle. Llamamos *redes cíclicas* a este tipo de redes. Podemos ver un ejemplo de este tipo de red en la figura 5.3.

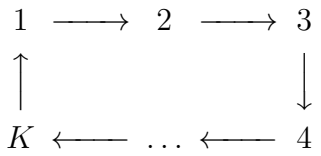

Figura 5.3: Red de colas cíclicas.

Las probabilidades asociadas a las transiciones en este tipo de red son:

$$p_{ij} = \begin{cases} 1 & j = i+1, 1 \leq i \leq K-1 \\ 1 & i = k, j = 1 \\ 0 & \text{en otro caso} \end{cases}$$

El sistema obtenido a partir del teorema de Jackson ahora es:

$$\begin{cases} \Lambda_1 = \Lambda_K \\ \Lambda_2 = \Lambda_1 \\ \vdots \\ \Lambda_K = \Lambda_{K-1} \end{cases}$$

La solución de este sistema es de nuevo trivial, con:

$$\Lambda_1 = \cdots = \Lambda_K = c$$

En este caso, c es una constante positiva arbitraria. Observamos que debido a la condición de que $\rho_i < 1$, necesariamente se tiene que $0 \leq c \leq \text{mín}\{\mu_i\}$. Si se cumple esta condición, podemos proceder como vimos en la sección 5.3.

Existe un caso particular de fácil análisis: si cada cola tiene un único servidor y las tasas de servicio de las colas coinciden, escogiendo $c = \mu$ tendríamos que todos los ρ_i son exactamente 1. Esto simplifica los cálculos porque $a_i(n) = 1$ para todos los valores posibles de i y n, con lo que tenemos:

$$p(n_1, \ldots, n_K) = \frac{1}{G(N)}$$

Es decir, todos los posibles estados del sistema son equiprobables. Además es fácil calcular las probabilidades ya que podemos obtener la constante $G(N)$:

$$G(N) = \sum_{n_1+\cdots+n_K=N} 1 = \binom{N+K-1}{N}$$

Por lo tanto, tenemos:

$$p(n_1, \ldots, n_K) = \frac{N!(K-1)!}{(N+K-1)!}$$

Además, en este caso es fácil obtener las probabilidades marginales. Por ejemplo, para la cola 1, podemos obtener:

$$p(n_1, \cdot, \ldots, \cdot) = \sum_{n_2+\cdots+n_K=N-n_1} p(n_1, \ldots, n_K) = \frac{1}{G(N)} \sum_{n_2+\cdots+n_K=N-n_1} 1 = \frac{\binom{N-n_1+K-2}{N-n_1}}{\binom{N+K-1}{N}}$$

5.6 | Red BCMP

En [BCMP75] apareció una extensión a las redes de Jackson: las redes BCMP (en honor a sus autores, Baskett, Chandy, Muntz y Palacios). Esta extensión permitía diferentes disciplinas en las colas, así como distribuciones diferentes a la exponencial. También incluían la posibilidad de redes abiertas, cerradas, e incluso mixtas.

Este tipo de red esta formado por centros de servicio de uno de los cuatro tipos posibles permitidos:

1. Cola FIFO en la que cada cliente tiene el mismo tiempo de servicio exponencial, que puede depender del número de clientes en el centro de servicio. También conocido como servidor de cola.
2. Cola de disciplina PS.
3. Cola de infinitos servidores. También llamada centro de retardo o servidor de retardo.
4. Cola LIFO de un servidor en la que el trabajo ya realizado con un cliente no se pierde.

Además, en todos los tipos menos en el primero se requiere que las distribuciones de tiempo tengan transformadas de Laplace racionales.

Este tipo de redes es relevante por sus usos y por su fácil tratamiento en base al siguiente teorema:

TEOREMA 5.2. *Sea una red BCMP de K colas. Entonces las probabilidades en el estado de equilibrio vienen dadas por:*

$$p(n_1, \ldots, p_K) = C \prod_{i=1}^{K} p_i(n_i)$$

En este caso, C es una constante elegida para hacer que las probabilidades sumen 1 y p_i representa la distribución en el equilibrio para la cola i-ésima.

6 | Modelos de colas en computación

A lo largo de los capítulos 3, 4 y 5 hemos visto la teoría básica de colas y redes de colas. En este capítulo usaremos lo visto para analizar algunos modelos referidos a la computación: analizaremos los recursos de una computadora, modelizándola como una red de colas. La motivación para usar redes de colas es simple: una computadora está compuesta por varios elementos que dependen entre ellos de manera que los inputs de uno de los elementos pueden ser los outputs de otro, de manera similar a las redes de colas que hemos visto.

Es realmente complicado obtener modelos relevantes para computación ya que las computadoras tienen un gran número de elementos interconectados de maneras muy complejas. Sin embargo, existen algunos modelos lo suficientemente simples pero con resultados relevantes, que estudiaremos a lo largo de este capítulo.

6.1 | Introducción

En el capítulo 5 hemos visto redes abiertas y cerradas, donde en las primeras había un número variable de clientes que recorrían la red y salían de ella después de recibir determinados servicios, mientras que en una red cerrada había un número constante de clientes que se repartían por las colas de la red. Sin embargo, existe un tercer tipo de red: la red mixta, donde algunos clientes entran desde fuera del sistema y desaparecen después, mientras otros clientes no escapan de la red y permanecen siempre en ella.

Debido a esto, es habitual denotar por C el número de clases de clientes, es decir, el número de tipos diferentes de clientes, cuyos comportamientos dentro de la red son esencialmente diferentes como en una red mixta. Por otra parte, K seguirá denotando el número de nodos en una red. Supondremos que el nodo k tiene un tiempo medio de servicio S_{ck} para clientes de la clase c, y que cada cliente

de la clase c realiza un total de V_{ck} visitas al nodo k, obteniendo así una demanda del servicio de $D_{ck} = V_{ck}S_{ck}$ para la clase c. Observemos que la demanda total en el nodo k es por tanto $D_k = \sum_c D_{ck}$. Por otra parte, si el sistema tiene un rendimiento λ, entonces podemos obtener el rendimiento de la clase c en el nodo k como $\lambda_{ck} = \lambda V_{ck}$.

Uno de los factores clave analizado en el rendimiento de una computadora es el *cuello de botella*: el nodo cuyo rendimiento es bajo y dificulta el correcto funcionamiento de los demás, resultando en un sistema saturado al estarlo uno o más de sus nodos. Este nodo se puede identificar fácilmente al ser el de más demanda. En el ámbito de la computación se suele resumir especificando si el cuello de botella es la CPU o el I/O (input/output, es decir, la lectura y/o escritura de datos).

Sin embargo, no es fácil determinar el cuello de botella en el rendimiento de una computadora. Imaginemos el caso de un servidor que todas las mañanas recibe datos, los procesa durante el día, y los envía de vuelta por la noche. Es de esperar que el cuello de botella en los envíos y recepciones de datos sea el I/O, mientras que en la etapa de procesamiento, el cuello de botella será la CPU. Es por esto que normalmente únicamente analizamos los cuellos de botella en las situaciones más exigentes para la computadora, ya que analizar todos los posibles cuellos de botella no sería factible.

6.2 | Modelo de población finita

En esta sección presentamos un modelo de población finita, en el que únicamente aparece una clase de trabajos. El lector interesado podrá encontrar más información de este modelo en [Mun75]. En este modelo de red, que podemos observar en la figura 6.1, presenta un procesador central al que los clientes acceden mediante N estaciones de trabajo. Cada uno de los clientes está en uno de los tres estados:

1. Introduciendo el trabajo en la estación de trabajo (este tiempo se denomina tiempo de trabajo t, con media $E[t] = T$).
2. Esperando a recibir servicio del procesador central.
3. Recibiendo servicio del procesador central.

Por otra parte, también suponemos que el cliente no puede empezar con un trabajo nuevo hasta que el anterior ha concluido.

Como se observa en la figura 6.1 y en los posibles estados establecidos para el cliente, el sistema es completamente cerrado. La red concreta que usemos dependerá

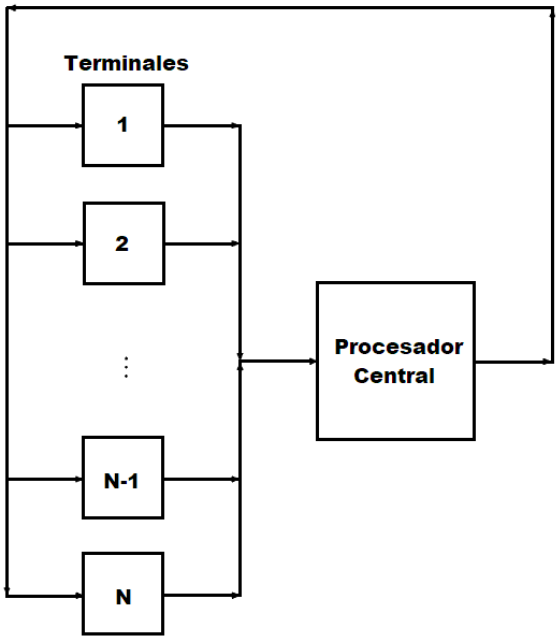

Figura 6.1: Modelo de población finita

del modelo de cola que tomemos para el procesador central. Uno de los modelos más simples y más usados es el que aparece al usar un modelo con tiempo de servicio exponencial.

Si además suponemos que el tiempo de trabajo es exponencial, obtenemos un modelo de cola $M/M/1/K/K$, también conocido como el modelo de reparación de máquinas. En general, los resultados obtenidos para este modelo también son válidos en el caso de que el tiempo de trabajo sea general y mientras el procesador central mantenga su distribución exponencial, como se puede ver en [Lav83].

Pasamos ya al análisis del modelo: como el tiempo de servicio en las estaciones de trabajo es precisamente el tiempo de trabajo, con media $E[t] = T$, podemos usar la fórmula de Little, obteniendo:

$$W = \frac{N}{\lambda} - T$$

Además, tenemos que $\lambda W_s = 1 - p_0$, de manera que:

$$W = \frac{NW_s}{1 - p_0} - T$$

En este caso, se tiene:

$$p_0 = \left(N! \sum_{k=0}^{N} \frac{W_s^k}{(N-k)! T^k} \right)^{-1}$$

En este modelo también podemos encontrar un valor relevante para el análisis del modelo: N^* es el *número de saturación*, que refleja el punto a partir del cual el procesador central será el cuello de botella, y no las estaciones de trabajo. N^* se puede obtener con:

$$N = \frac{T + W_s}{W_s}$$

Este punto tiene una interpretación física entendible: representa el número de clientes a partir del cual los trabajos de los clientes interfieren entre sí, de manera que cada cliente retrasa a los demás progresivamente.

Si asumimos que el tiempo de trabajo es exponencial y que el procesador central tiene una disciplina PS (processor sharing), los resultados siguen siendo válidos bajo algunas condiciones en términos de las transformadas de Laplace-Stieltjes del tiempo de servicio, como podemos ver en [Kle76].

6.3 | Modelo de servidor central

Ahora discutimos un modelo de servidor central, usado para modelizar la memoria principal de una computadora a partir de una red de Jackson cerrada como la que se puede observar en la figura 6.2.

En este modelo, tenemos un servidor central representando la CPU, que trabaja con un número de programas (los clientes) constante, es decir, que tenemos una computadora multiprograma con N programas. Cada vez que un programa pasa de la CPU a la CPU de nuevo, entendemos que un programa ha terminado de ejecutarse y ha empezado la ejecución de otro programa diferente, por lo que suponemos que hay toda una cola de programas esperando para ejecutarse.

En este modelo consideramos $K-1$ servidores de I/O, cada uno con su propia cola, tasa de servicio μ_k y disciplina FIFO. Por otra parte, la CPU se toma con una distribución exponencial, aunque en [Lav83] se puede ver que también funciona una distribución general con una disciplina PS. Al completar un servicio en la CPU, un proceso completa su ejecución con probabilidad p_K, mientras que pasa al servidor I/O k con probabilidad p_k.

Podemos usar dos algoritmos para obtener las medidas de rendimiento del sistema.

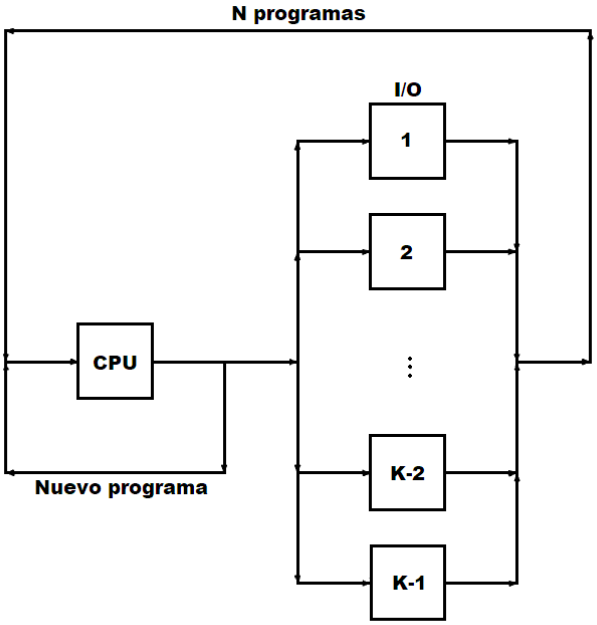

Figura 6.2: Modelo de servidor central

ALGORITMO 6.1. *Consideramos un modelo de servidor central. Supongamos que conocemos S_k y las probabilidades asociadas al modelo p_k. Así podemos obtener la demanda para cada uno de los k nodos.*

1. *Definimos el ratio de visitas a la CPU.*

$$V_K = \frac{1}{p_K}$$

2. *Obtenemos los ratios de visita a los servidores I/O:*

$$V_k = p_k V_K \qquad k = 1, \ldots, K-1$$

3. *Obtenemos las demandas de cada nodo:*

$$D_k = V_k S_k \qquad k = 1, \ldots, K$$

ALGORITMO 6.2. *Consideramos un modelo de servidor central. Supongamos que conocemos la demanda D_k para cada uno de los k nodos, y el nivel de multiprogramación N. A partir de aquí podemos obtener las medidas de rendimiento del sistema.*

1. *Inicializamos:*

$$L_k[0] = 0 \qquad k = 1, \ldots, K$$

2. *Para $n = 1, \ldots, N$ iteramos:*

$$\begin{cases} W_k[n] = D_k(1 + L_k[n-1]) & k = 1, \ldots, K \\ W[n] = \sum_{k=1}^{K} W_k[n] & \\ \lambda[n] = \frac{n}{W[n]} & \\ L_k[n] = \lambda[n] W_k[n] & k = 1, \ldots, K \end{cases}$$

3. *Obtenemos las medidas de rendimiento:*

$$\begin{cases} W = W[N] \\ \lambda = \lambda[N] \\ \rho_k = \lambda D_k \qquad k = 1, \ldots, K \end{cases}$$

Aunque estos algoritmos nos permiten obtener el rendimiento del modelo, este presenta algunas flaquezas, como por ejemplo el hecho de que cada I/O puede atender únicamente un programa diferente, que no haya prioridades en los servidores I/O, o que cada programa vaya a un servidor I/O aleatoria. También tiene algunas ventajas: permite estudiar fácilmente el incremento de rendimiento al disponer de más servidores o de un mejor nivel de multiprogramación, o al incrementar los niveles de servicio en cualquiera de los servidores.

Podemos extender este modelo a un *modelo abierto de servidor central*, en el que en vez de un bucle que representa el fin de la ejecución de un programa y el inicio de uno nuevo, incluimos después de la CPU una salida del sistema con la misma probabilidad p_K, y entradas a la CPU con tasa de llegada λ.

En este caso, podemos obtener las demandas como anteriormente, y a partir de aquí medir el rendimiento del sistema a partir del siguiente algoritmo:

Algoritmo 6.3. *Consideramos un modelo abierto de servidor central. Podemos obtener las medidas de rendimiento del sistema con:*

1. *Para $k = 1, \ldots, K$, calculamos:*

$$\begin{cases} \rho_k = \lambda D_k \\ L_k = \frac{\rho_k}{1-\rho_k} \\ W_k = \frac{D_k}{1-\rho_k} \end{cases}$$

2. *A partir de aquí obtenemos:*

$$W = \sum_{k=1}^{K} W_k$$

$$L = \lambda W$$

3. *Por otra parte, podemos obtener el cuello de botella:*

$$\lambda_{max} = \frac{1}{D_{max}}$$

6.4 | Modelos BCMP

6.4.1 | Modelo de una clase cerrado

El modelo cerrado BCMP para una clase es el mostrado en la figura 6.3. Podemos observar que este modelo se obtiene al sustituir el procesador central de la figura 6.1 por la variante abierta de la figura 6.2. Los N servidores son estaciones de retardo (tipo 3), la CPU se puede considerar exponencial (tipo 1) o PS (tipo 2), y los servidores I/O son exponenciales (tipo 1). La caja discontinua de la figura 6.3 representa el sistema central de la computadora. En este modelo no contemplamos un límite para el número de procesos que puede haber simultáneamente en la computadora, cosa que sí haremos en modelos posteriores.

Como en los casos anteriores, podemos obtener las medidas de rendimiento de este modelo con el siguiente algoritmo:

Algoritmo 6.4. *Sea un modelo BCMP de una clase cerrado como el de la figura 6.3. Suponemos que el tiempo medio de trabajo es $E[t] = T$ para cada uno de los N terminales, que son de tipo 3. La CPU, representada por el nodo $N+1$, es tipo 1 o tipo 2 con demanda $D = D_{N+1}$ dada. También conocemos las demandas D_k para los servidores I/O. Podemos calcular las medidas de rendimiento del sistema con:*

1. *Para $k = N+1, \ldots, K$, inicializamos:*

$$L_k[0] = 0$$

2. *Para $n = 1, \ldots, N$, calculamos:*

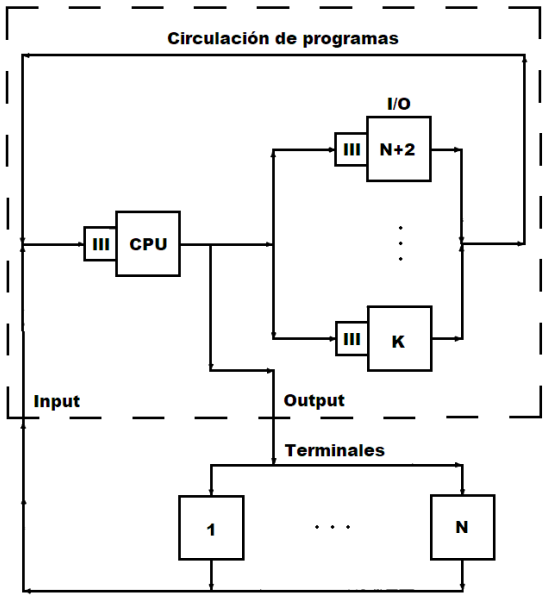

Figura 6.3: Modelo BCMP cerrado

$$\begin{cases} W_k[n] = D_k(1 + L_k[n-1]) & k = N+1, \ldots, K \\ W[n] = \sum_{k=N+1}^{K} W_k[n] & \\ \lambda[n] = \frac{n}{W[n]+T} & \\ L_k[n] = \lambda[n] W_k[n] & k = N+1, \ldots, K \end{cases}$$

3. *Obtenemos las medidas de rendimiento:*

$$\begin{cases} \lambda = \lambda[N] & \\ W = W[N] & \\ \rho_k = \lambda D_k & k = N+1, \ldots, K \\ \lambda_k = \lambda V_k & k = N+1, \ldots, K \\ L_k = L_k[N] & k = N+1, \ldots, K \\ W_k = W_k[N] & k = N+1, \ldots, K \end{cases}$$

La mayor debilidad de este modelo es que no existen problemas de memoria: es decir, siempre hay lugar para cada trabajo que llega. Este supuesto es razonable para modelos con poca carga de trabajo, pero en la práctica las computadoras deben limitar el nivel de multiprogramación y este efecto sí es relevante cuando la carga de trabajo es alto.

En el caso de que la carga de trabajo sea baja, podemos usar el algoritmo anterior. En otro caso, podemos fijar un nivel de multiprogramación (por ejemplo, el más alto que creamos posible) y usar los algoritmos del modelo de servidor central cerrado para calcular λ, y a partir de aquí obtener $W = N/\lambda - T$ para calcular el tiempo medio de respuesta.

Por otra parte, existe otro problema, en el que el nivel de multiprogramación es alto pero tiene una media relativamente baja. Durante la computadora está en activo supondremos que este nivel fluctuará entre cero y m, dependiendo de si la computadora está parada o si el máximo nivel ha sido alcanzado y hay una cola de procesos esperando para entrar en la memoria principal.

Es habitual usar una técnica para manejar el control de memoria en dos pasos: en el primero se sustituye la computadora central mostrada en el cuadrado discontinuo de la figura 6.3 y la cola (la que aparece cuando el nivel m ya se ha alcanzado) por un *FESC* (flow equivalent service center), una caja negra que, dada la carga de trabajo del sistema, devuelve los mismos parámetros que el sistema real. El segundo paso consiste en modelizar el sistema de la figura 6.3 usando el FESC en vez de el sistema central.

Todo esto se puede resumir en el siguiente algoritmo:

ALGORITMO 6.5. *Consideramos el sistema BCMP cerrado de la figura 6.3. Podemos obtener las medidas de rendimiento del sistema con:*

1. *Usamos el algoritmo 6.2 para calcular* $\mu[j] = \lambda[j]$ *para* $j = 1, \ldots, m$.
2. *Inicializamos* $q[0] = 1$ *y* $\mu[j] = \mu[m]$ *para* $j = m+1, \ldots, N$.
3. *Para* $n = 1, \ldots, N$, *repetimos:*

   a) *Inicializamos* $W = 0$ *y para* $j = 1, \ldots, n$ *hacemos:*

   $$W = W + \frac{j}{\mu[j]} q[j-1]$$

   b) *Obtenemos:*

   $$\lambda[n] = \frac{n}{W+T}$$

   c) *Inicializamos* $s = 0$, *y para* $j = n, n-1, \ldots, 1$, *obtenemos:*

   $$q[j] = \frac{\lambda[n]}{\mu[j]} q[j-1]$$

   $$s = s + q[j]$$

d) *Recalculamos:*

$$q[0] = 1 - s$$

4. *Para finalizar:*

$$\begin{cases} \lambda = \lambda[N] & \\ \rho_k = \lambda D_k & k = N+1, \ldots, K \\ \lambda_k = \lambda V_k & k = N+1, \ldots, K \\ p_j = q[j] & j = 0, \ldots, N \end{cases}$$

En la vida real, pocas computadoras están enfocadas únicamente a un proceso, por lo que es conveniente usar los llamados modelos multiclase, que reflejan más acertadamente las propiedades de las computadoras. El problema de estos métodos es determinar las proporciones entre las clases, lo que puede llevar a algunos errores de cálculo del rendimiento.

6.4.2 | Modelo multiclase abierto

Como su propio nombre indica, un modelo multiclase abierto únicamente tiene clases abiertas. Estos modelos son populares porque su resolución es sencilla, aunque evidentemente no podrán representar un modelo real con tanta precisión como un modelo multiclase mixto.

Supongamos que hay C clases, cada una con tasa de llegadas λ_c. Además, cada clase tiene cierta demanda D_{ck} para cada centro de servicio k. Supondremos además que el sistema es estacionario, es decir, que $\rho_k < 1$ para cada k. Recordemos que el uso de cada centro de servicio k viene dado por:

$$\rho_k = \sum_{c=1}^{C} \lambda_c D_{ck} \qquad k = 1, \ldots, K$$

En este caso, tenemos:

$$W_{ck} = \begin{cases} D_{ck} & \text{para servidores de retardo} \\ \dfrac{D_{ck}}{1 - \sum_{j=1}^{C} \lambda_j D_{jk}} & \text{para servidores de cola} \end{cases}$$

Por la ley de Little, tenemos:

$$L_{ck} = \lambda_c W_{ck}$$

El tiempo medio de respuesta para la clase c es:

$$W_c = \sum_{k=1}^{K} W_{ck}$$

Por otro lado, usando de nuevo la ley de Little, tenemos:

$$L_c = \lambda_c W_c$$

6.4.3 | Modelo multiclase cerrado

Este tipo de modelos es mucho más difícil de resolver que la variante abierta, tanto en términos de complejidad como de coste.

Para el siguiente algoritmo denotaremos por N_c el número de clientes del tipo c en el sistema (que es fijo), y denotaremos $N = (N_1, \ldots, N_C)$. Denotaremos por $N - 1_c$ la población anterior pero sin un cliente del tipo c.

Algoritmo 6.6. *Sea un sistema como el de la figura 6.3. Supongamos que hay C clases de clientes, con N_c clientes en cada clase, y que cada clase tiene una demanda D_{ck}. Podemos obtener las medidas de rendimiento del sistema repitiendo los siguientes pasos con todas las poblaciones n en las que $n_c \leq N_c$.*

1. *Para $k = 1, \ldots, K$ y $c = 1, \ldots, C$ calculamos:*

$$W_{ck}[n] = \begin{cases} D_{ck} & \text{para servidores de retardo} \\ D_{ck}[1 + L_k[n - 1_c]] & \text{para servidores de cola} \end{cases}$$

2. *Para $c = 1, \ldots, C$, calculamos:*

$$W_c[n] = \sum_{k=1}^{K} W_{ck}[n]$$

$$\lambda_c[n] = \frac{N_c}{T_c + W_c[n]}$$

3. *Para $k = 1, \ldots, K$, calculamos:*

$$L_k[n] = \sum_{c=1}^{C} \lambda_c[n] W_{ck}[n]$$

Este algoritmo es exacto, y es razonablemente eficiente, aunque para sistemas grandes, está lejos de ser rápidos. Por eso presentamos un algoritmo aproximado que proporciona valores mucho más rápidamente.

ALGORITMO 6.7. *Sea un sistema como el de la figura 6.3. Supongamos que hay C clases de clientes, con N_c clientes en cada clase, y que cada clase tiene una demanda D_{ck}. Sea ε el criterio de error deseado en el cálculo. Entonces podemos aproximar las medidas de rendimiento del sistema.*

1. *Para $c = 1, \ldots, C$ y $k = 1, \ldots, K$ calculamos:*

$$L_{ck}^{(1)}[N] = \frac{N_c}{K}$$

2. *Para $c = 1, \ldots, C$ y $k = 1, \ldots, K$ calculamos:*

$$A_{ck}[N] = \frac{N_c - 1}{N_c} L_{ck}^{(1)}[N] + \sum_{j \neq c}^{C} L_{jk}^{(1)}[N]$$

3. *Para $c = 1, \ldots, C$ y $k = 1, \ldots, K$ calculamos:*

$$W_{ck}[N] = \begin{cases} D_{ck} & \text{para servidores de retardo} \\ D_{ck}[1 + A_{ck}[N]] & \text{para servidores de cola} \end{cases}$$

4. *Para $c = 1, \ldots, C$ calculamos:*

$$W_c[N] = \sum_{k=1}^{K} W_{ck}[N]$$

$$\lambda_c[N] = \frac{N_c}{T_c + W_c[N]}$$

5. *Para $c = 1, \ldots, C$ y $k = 1, \ldots, K$ calculamos:*

$$L_{ck}^{(2)}[N] = \lambda_c[N] W_{ck}[N]$$

6. *Si $|L_{ck}^{(1)}[N] - L_{ck}^{(2)}[N]| \geq \varepsilon$ para algunos $c = 1, \ldots, C$ y $k = 1, \ldots, K$, volvemos a empezar desde el paso dos, reemplazando $L_{ck}^{(1)}[N]$ por $L_{ck}^{(2)}[N]$. En caso contrario, los valores que habíamos obtenido en el último paso de $W_{ck}[N], W_c[N]$ y $\lambda_c[N]$ son los valores deseados, y podemos obtener:*

$$L_k[N] = \sum_{c=1}^{C} \lambda_c[N] W_{ck}[N]$$

6.4.4 | Modelo multiclase mixto

Los métodos multiclase mixtos presentan tanto clases abiertas como clases cerradas. Aunque, como hemos visto, según [BCMP75] este tipo de modelos tiene solución, estas tienden a ser soluciones en forma de aproximaciones. La idea es reducir la solución a la de un modelo multiclase cerrado para resolverlo con el algoritmo exacto o con el aproximado que hemos visto anteriormente. La influencia de las clases abiertas se introduce inflando la demanda de servicio de las clases cerradas. Denotaremos por O las clases abiertas y por C las clases cerradas.

Algoritmo 6.8. *Consideramos un modelo multiclase BCMP. Suponemos que las clases cerradas tienen demandas D_{ck}. Podemos obtener una solución aproximada a partir de los siguientes pasos:*

1. *Para cada k, calculamos:*

$$\rho_{ck} = \lambda_c D_{ck} \qquad \forall c \in O$$

También obtenemos el uso total por parte de las clases abiertas:

$$\rho_{Ok} = \sum_{c \in O} \rho_{ck}$$

2. *Cambiamos las demandas de las clases cerradas por:*

$$D^*_{ck} = \frac{D_{ck}}{1 - \rho_{Ok}} \qquad \forall c \in C, k = 1, \ldots, K$$

3. *Usamos el algoritmo exacto para modelos multiclase cerrados.*

4. *Los parámetros de las clases cerradas son los obtenidos al usar el algoritmo exacto. Para los abiertos, podemos calcular:*

$$W_{ck} = \frac{D_{ck}[1 + L_{Ck}]}{1 - \rho_{Ok}} \qquad \forall c \in O$$

$$L_{ck} = \lambda_c W_{ck} \qquad \forall c \in O$$

Es fácil implementar este algoritmo a partir del algoritmo exacto usado en modelos multiclase cerrados, aunque este algoritmo tampoco es perfecto, ya que prioriza las clases abiertas sobre las cerradas, independientemente de las prioridades del sistema original, debido a la manera en la que el algoritmo infla las demandas de las clases cerradas.

Índice analítico

proceso estocástico, 5
  de conteo, 7
    de Poisson, 8
  de Markov, 11
    cadena de Markov, 11
    matriz de transición, 13
    tasas de transición, 14
  de nacimiento y muerte, 18
    tasa de muerte, 18
    tasa de nacimiento, 18
  espacio de estados, 5
  espacio de tiempos, 5
  estacionario
    débilmente, 7
    fuertemente, 7

red de colas, 47
  abierta, 47
  BCMP, 53
    servidor de cola, 53
    servidor de retardo, 53
  cíclica, 52
  cerrada, 47
  de Jackson, 49
  en serie, 51

sistema de cola, 23
  capacidad, 22
  cliente, 22
  cola, 23
  disciplina, 22
    FIFO, 22
    GD, 22
    LIFO, 22
    PR, 22
    PS, 22
    SIRO, 22
  llegadas, 21
  población, 21
  servicio, 22
    servidores, 22

Bibliografía

[Aba02] R.C. Abad. *Introducción a la Simulación y a la Teoría de Colas*. Catálogo General. Netbiblo, 2002.

[All90] A.O. Allen. *Probability, Statistics, and Queueing Theory: With Computer Science Applications*. Computer science and scientific computing. Academic Press, 1990.

[BCMP75] Forest Baskett, K. Mani Chandy, Richard R. Muntz, and Fernando G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *J. ACM*, 22(2):248–260, April 1975.

[Bha15] U.N. Bhat. *An Introduction to Queueing Theory: Modeling and Analysis in Applications*. Statistics for Industry and Technology. Birkhäuser Boston, 2015.

[CT83] M.L. Chaudhry and J.G.C. Templeton. *A first course in bulk queues*. Wiley, 1983.

[Erl09] Agner K. Erlang. The Theory of Probabilities and Telephone Conversations. *Nyt Tidsskrift for Matematik*, 20(B):33–39, 1909.

[GH98] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory*. Wiley Series in Probability and Statistics. Wiley, 1998.

[Jac04] James R. Jackson. Jobshop-like queueing systems. *Management Science*, 50(12_supplement):1796–1802, 2004.

[Ken53] David G. Kendall. Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. *Ann. Math. Statist.*, 24(3):338–354, 09 1953.

[Kle76] Leonard Kleinrock. *Queueing Systems, Volume II: Computer Applications*. John Wiley, New York, 1976.

[Lav83] Stephen S. Lavenberg. *Computer Performance Modeling Handbook*. Academic Press, Inc., Orlando, FL, USA, 1983.

[Lit61] John D. C. Little. A Proof for the Queuing Formula: $L = \lambda W$. *Operations Research*, 9(3):383–387, 1961.

[MOVG10] Fermín Mallor, Edward A. M. Omey, and Stefan Van Gulck. *Introduction to the Modelling and Analysis of Queuing Systems. A text book for students of engineering, economics and management sciences*. Colección Estadística. Universidad Pública de Navarra, 2010.

[Mun75] R. R. Muntz. Analytic modeling of interactive systems. *Proceedings of the IEEE*, 63(6):946–953, 1975.