UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

TEORÍA DE COLAS
Modelo M/M/s

Autor: Roger Pérez Parera

Director: Dr. José Manuel Corcuera Valverde
Realitzat a: Departament de Probabilitat i Estadística

Barcelona, 21 de junio de 2020

Abstract

In this Final Degree Project, the Queuing Theory is developed around its best known model: the $M/M/s$. To do this, two previous theory sections necessary for this model are presented.

The first of these sections is about the Poisson process, where definitions and results on the exponential distribution are included and then the Poisson process itself is presented. The Poisson process helps us determine the distribution of customer arrivals in $M/M/s$ queuing systems.

The second of these sections is about continuous time Markov chains. This section plays a central role in $M/M/s$ queuing systems, since these are a particular case of birth and death processes, which are nothing more than a specific type of continuous-time Markov chain.

With this, section number 4 focuses on the development of $M/M/s$ queuing systems with the objective of knowing measures of effectiveness in relation to these queues and knowing the most appropriate number of servers for a $M/M/s$ queue. To this end, birth and death processes, the concept of stationary distribution and Little's formulas are introduced earlier.

To give practical sense to the theoretical abstraction, in section number 6 a real case of application of the $M/M/s$ model is presented, before presenting in section number 5 the basic theory of queuing networks.

Resumen

En este Trabajo Final de Grado se encuentra desarrollada la Teoría de Colas alrededor de su más conocido modelo: el $M/M/s$. Para ello, se presentan dos secciones de teoría previas necesarias para este modelo.

La primera de estas secciones es acerca del proceso de Poisson, donde se incluyen definiciones y resultados sobre la distribución exponencial para luego presentar el proceso de Poisson propiamente. El proceso de Poisson nos sirve para determinar la distribución de la llegada de clientes en sistemas de colas $M/M/s$.

La segunda de estas secciones es acerca de las cadenas de Markov en tiempo continuo. Esta sección juega un papel central en los sistemas de colas $M/M/s$, puesto que éstos son un caso particular de los procesos de nacimiento y muerte, que no son más que un tipo específico de cadena de Markov a tiempo continuo.

Con ello la sección número 4 se centra en el desarrollo de los sistemas de colas $M/M/s$ con el objetivo de conocer medidas de efectividad con relación a estas colas y de conocer el número más adecuado de servidores para una cola $M/M/s$. Con este objetivo, se introducen antes los procesos de nacimiento y muerte, el concepto de distribución estacionaria y las fórmulas de Little.

Para dar sentido práctico a la abstracción teórica, en la sección número 6 se presenta un caso real de aplicación del modelo $M/M/s$, antes presentando en la sección número 5 la teoría básica sobre las redes de colas.

2010 Mathematics Subject Classification. 60K25, 60K30, 60J25, 60J27, 60G55

Agradecimientos

Al doctor José Manuel Corcuera Valverde por aceptar tutorizar este trabajo.

A mis seres queridos y amigos por acompañarme durante el largo viaje del Grado en Matemáticas.

Índice

1. Introducción

La **Teoría de Colas** afronta uno de los problemas más cotidianos de la vida moderna: las esperas, los retrasos y las demoras. Todos nosotros hemos esperado en la cola del banco, del supermercado o de la carnicería, a la vez que todos hemos tenido que esperar a ser atendidos en un hospital, que pedir hora con antelación en algún servicio público o que dedicado horas a hacer algún trámite burocrático. Las colas son tan importantes en nuestro día a día que son usadas habitualmente como testigo de la buena o mala gestión de los servicios públicos y, por ende, usadas para el debate político.

Las colas se forman porque hay más demanda del servicio que capacidad para atenderlo, es decir, los recursos son limitados. Por este motivo, la teoría de colas no solo tiene trascendencia matemática, sino que es de gran importancia para la gestión efectiva de los recursos y para la operativa y logística empresarial. La finitud de los recursos hace inevitable que tengamos que teorizar para optimizar las colas.

Figura 1: La cumbre del Everest en mayo de 2019

La imagen de una cola dio la vuelta al mundo el año 2019: La cola de la cumbre del Everest, a ~ 8848 metros, que causó varios fallecidos y reacciones políticas en Nepal y en el Tíbet. La Teoría de Colas permite conocer mejor los recursos disponibles, la capacidad de atención de la demanda y posibles soluciones a los problemas que aparecen en la gestión de servicios donde no todos los clientes pueden ser atendidos a la vez. El desarrollo económico y la competitividad nos obligan a administrar cada vez mejor y de manera más precisa estas colas. Por eso la cola en el Everest nos sirve de ejemplo: era impensable hace un siglo que un hecho así sucediera, pero ahora es una necesidad aportar soluciones. Y en este sentido hay que destacar la elevada utilidad práctica que tiene esta teoría. Su aplicación no solo puede mejorar la cuenta de resultados de una empresa, sino que también aumentar la satisfacción de los clientes, mejorar la administración pública, incentivar el consumo y estimular la economía, reducir los costes de oportunidad por las esperas y, por encima de todo, como en el caso del Everest, salvar vidas.

La Teoría de Colas fue desarrollada por el matemático danés A.K. Erlang en 1904 a través de su publicación *The theory of Probabilities and Telephone Conversations*, con el objetivo de determinar la capacidad necesaria para el sistema telefónico danés. Erlang se enfrentaba al siguiente problema: quería conocer lo grande que debía ser una central telefónica para mantener el número de llamadas en espera lo más bajo posible. Desde 1904 hasta la actualidad, la Teoría de Colas se ha ido desarrollando sobre todo a partir

de herramientas analíticas y teóricas y de simulaciones computacionales. Así esta teoría ha sido aplicada en bancos, aerolíneas, centros de logística, *call centers*, sistemas públicos de emergencia, gestión de centros de salud, etc. Su principal utilidad es identificar los niveles adecuados de personal y de infraestructura para lograr el rendimiento deseado de un servicio y, por consiguiente, poder tomar las decisiones. Las principales preguntas que pretende contestar son: *¿Cuánto tiempo deberá esperar un cliente que se una a la cola?* y *¿cuántas personas ocuparán esta cola?*. De esta manera, desde 1904, ha sido aplicada en múltiples ocasiones satisfactoriamente para mejorar la prestación de servicios en áreas como el tráfico de personas, de vehículos terrestres, de aeronaves y de comunicaciones; como la planificación de la atención a pacientes en hospitales, de la ejecución de tareas informáticas, del empleo de máquinas industriales; y como el diseño físico de bancos, oficinas de correos, parques de atracciones y restaurantes de comida rápida.

A partir de [Green, 2011], [Shortle u. a., 2018] y [Beasley, 2011], introducimos unas primeras nociones básicas sobre la teoría de colas: Una **cola** (o un retraso, demora, espera,...) es el resultado de la diferencia entre la demanda de un servicio y la capacidad disponible para atender esta demanda. Esta diferencia acostumbra a ser temporal y variable, dependiente ésta de los tiempos de la demanda y del tiempo necesario para proveer el servicio. A la vez, podemos considerar una cola como el conjunto ordenado formado por los clientes que en su llegada se han encontrado todos los servidores ocupados.
Entendemos por **sistema de colas** como el modelo de un servicio donde los **clientes** llegan a un **banco de servidores** y requieren de una prestación (o prestaciones) de uno (o varios) de los servidores. Un **cliente** es cualquier identidad requiriendo de un servicio. Un **servidor** es una identidad que provee el servicio. Notemos que tanto los clientes como los servidores pueden ser personas o *cosas*.
Para estructurar bien un sistema de colas necesitamos información acerca de:

- **El proceso de llegada**:
  - ¿Los clientes llegan individualmente o en grupos?
  - ¿Cómo está esta llegada distribuida en el tiempo? Es decir, cuál es la **distribución de probabilidad de la llegada de clientes**.
  - Si la muestra de potenciales clientes es finita o infinita.
  - Las posibles reacciones del cliente al llegar al sistema: ¿esperará en la cola sea cual sea el tiempo de espera? ¿renunciará a ser atendido si se cumplen algunos supuestos?
- **El servicio**:
  - Los recursos necesarios para dar el servicio.
  - ¿Cuál es el tiempo para realizar el servicio? Hablamos de la **distribución del tiempo de servicio**.
  - Número de servidores disponible y cuántos clientes puede atender simultáneamente un servidor.
  - Organización de los servidores: Hablamos de colas como **línea simple** si una sola cola es atendida por todos los servidores. Hablamos de colas en **paralelo** si cada servidor tiene su propia cola. Hablamos de colas en **red** si los clientes reciben el servicio de diferentes servidores de manera secuencial. Un cliente que requiere una intervención quirúrgica es un ejemplo de *cola en red*, dado que necesita de un gran número de servidores (especialistas médicos, quirófano libre,

cama, etc.). Observamos que todo sistema de colas puede ser descompuesto en subsistemas individuales de colas en línea simple [Beasley, 2011].

- Distinguimos si son con *derecho preferente* o no, es decir, si es posible que un cliente con una prioridad muy alta, al llegar, interrumpa el servicio dado a otro cliente.

- **Características de la cola**:

  - El conjunto de las reglas que determinan el orden en el que los clientes de la cola son atendidos, llamado **disciplina**. La disciplina más común es la regla *first-come, first served* (FCFS). Conocidas son también las reglas *last come, first served* (LCFS) o las *random selection for service* (RSS). También hay otras como las disciplinas basadas en la *prioridad*. Usada, por ejemplo, en el caso del sistema de triaje de los hospitales. La disciplina es un factor clave para la reducción de los tiempos de espera.
  - Hay que tener en cuenta la posibilidad que algunos clientes cambien de idea y finalmente no se unan a la cola, que algunos clientes abandonen la cola a media espera o que los clientes cambien de cola, en el caso de colas en paralelo.
  - Las *colas* pueden ser líneas físicas de personas u objetos, pero también líneas invisibles como la cola de espera de un centro de llamadas.
  - Si la capacidad de la cola es finita o infinita, es decir, si la sala de espera acepta ilimitados clientes o no.

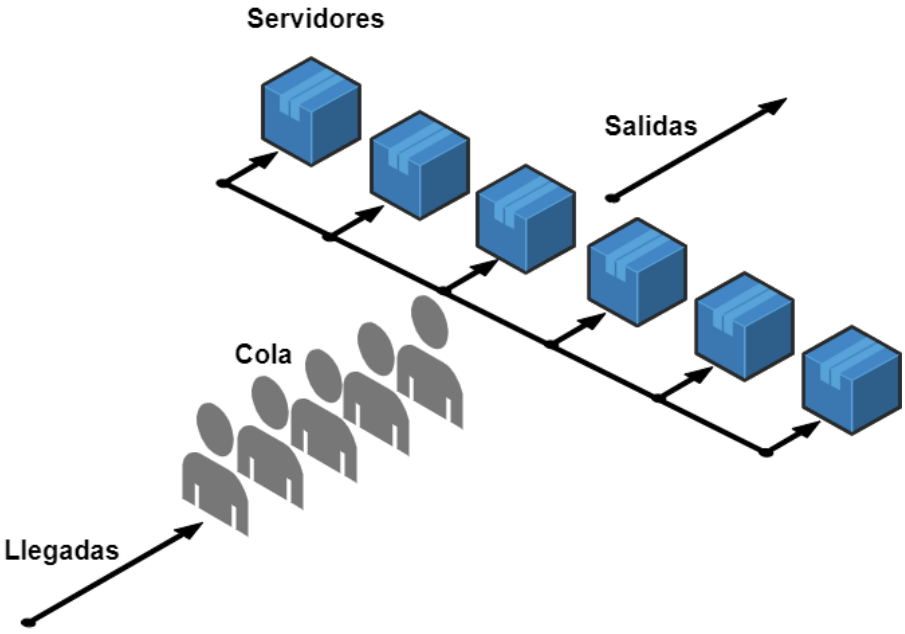

Figura 2: Diagrama conceptual de un sistema de colas.

Unos primeros sistemas de colas a tener en cuenta a modo de ejemplo serían:

- La cola del supermercado. Donde el *cliente* es el consumidor y el *servidor* es la caja. La prestación es el intercambio de unos productos por su valor monetario. La cola es una línea simple y física. La disciplina aplicada sería FCFS.

- Red de ordenadores. Donde el *cliente* es una tarea informática a ejecutar y el *servidor* es un ordenador conectado al sistema con capacidad para ejecutarla. La prestación es la ejecución de la tarea. La cola podría ser tanto simple, en paralelo o en red, y es invisible. La disciplina es arbitraria según se diseñe el sistema.

- Servicios de salud de urgencias. Donde el *cliente* es el paciente y el *servidor* es el conjunto de recursos físicos y humanos necesarios para atender al paciente. La prestación es el servicio médico de urgencia. La cola es en paralelo (dado que cada especialista tiene su propia cola) y en red (ya que el servicio involucra a más de un profesional y a más de una infraestructura), la cola es física. La disciplina es en prioridad con derecho preferente.

Continuando con el último ejemplo, los pacientes demandan el servicio sin cita previa, de manera impredecible y requiriendo servicios muy distintos (desde un tratamiento para la gripe como la identificación de síntomas de infarto). Es un sistema con alta variabilidad de servicios, donde destaca la impredecible llegada de la demanda, por tanto es complicado determinar los niveles de congestión provocados por las colas y cuál es la capacidad necesaria para satisfacer la demanda en determinados momentos. Hay que añadir a estas consideraciones la complejidad del propio servicio: cada consulta requerirá de diferentes camas, quirófanos, especialistas, enfermeras, equipos médicos, camillas, asistentes, ambulancias, etc. Además se contrapone la realidad presupuestaria de los servicios de salud y a su imperativo de mantener las cuentas, también, saneadas. Por eso, lograr el equilibrio entre utilización del servicio, capacidad media total, tiempos de espera y coste económico es uno de los retos de la Teoría de Colas.

A diferencia de metodologías que usan las simulaciones, los **modelos de colas** necesitan pocos datos y se presentan de una manera más simple y barata de utilizar. Hay muchos modelos de colas diferentes. El más conocido y aplicado, por su sencillez y utilidad, y en el que se basa este trabajo, es el modelo **M/M/s**, que definimos y desarrollamos en la sección 4. Las dos M en el nombre son por las dos suposiciones *Markovianas*: los clientes llegan de manera independiente entre ellos y siguiendo un proceso de Poisson con un parámetro constante; y las duraciones de los servicios son independientes entre sí y tienen una distribución exponencial. s denota el número de servidores, que en este modelo tienen que ser idénticos y capaces de realizar un solo servicio a la vez. Tiene una única cola y una sala de espera infinita. Otros conocidos modelos son:

- $M/M/s/K$ o cola con truncamiento. Es un modelo $M/M/s$ con un número máximo K de clientes en el sistema. El enfoque es el mismo que en el caso $M/M/s$, excepto por el parámetro de la distribución de la llegada de clientes: éste es 0 cuando el número de clientes en el sistema es K.

- $M/M/s/s$. Es un modelo de cola con truncamiento con $K = s$. En este caso, cuando hay s clientes en el sistema, en vez de formarse una cola, los nuevos clientes no acceden en el sistema.

- $M/G/1$. En este modelo las llegadas son Markovianas, es decir, son independientes entre sí y siguen un proceso de Poisson, pero los tiempos de servicio siguen una distribución arbitraria (o general) que cumple que son iid e independientes de las llegadas. Hay un único servidor.

- $G/M/1$. Como en el anterior, pero las llegadas siguen una distribución arbitraria (o

general) cumpliendo las mismas condiciones y los tiempos de servicio son Markovianos. Hay un único servidor.

- Hay muchas combinaciones distintas de los modelos anteriores: $M/M/\infty$, $M/G/s$, $M/G/\infty$, $G/M/s$, $G/G/1$. También hay otros modelos como $M/D/s$, donde hay s servidores, las llegadas son Markovianas y los tiempos de servicio son constantes.

2. Proceso de Poisson

Con el objetivo de modelizar un sistemas de colas con el modelo $M/M/s$, debemos introducir los conceptos necesarios para las suposiciones Markovianas. La suposición Markoviana respecto la llegada de los clientes tiene que ver con el **proceso de Poisson**, el cuál estudiaremos durante esta sección. Empezamos introduciendo unos primeros conceptos en el primer apartado, para luego definir el proceso de Poisson y justificar su uso en la Teoría de colas.

2.1. Distribución exponencial

Definición 2.1. *Una variable aleatoria T positiva tiene una **distribución exponencial** con parámetro λ si*

$$\mathbb{P}(T \leq t) = 1 - e^{-\lambda t}$$

para $t \geq 0$. Escribimos $T \sim exp(\lambda)$.

En esta última definición, $\mathbb{P}(T \leq t)$ es la **función de distribución** de T: $F(t) = \mathbb{P}(T \leq t)$. Equivalentemente podemos dar la definición a partir de la **función de densidad**:

$$f_T(t) = \lambda e^{-\lambda t} \mathbb{1}_{[0,\infty)}(t)$$

Destacamos una importante propiedad de la distribución exponencial: **no tiene memoria**. Es decir:

Proposición 2.2. *Si T es una variable aleatoria con una distribución exponencial, entonces*

$$P(T > t + s | T > t) = P(T > s)$$

Demostración. Recordamos que si $B \subset A$, entonces $\mathbb{P}(B|A) = \mathbb{P}(B)/\mathbb{P}(A)$. Así podemos demostrar:

$$\mathbb{P}(T > t + s | T > t) = \frac{\mathbb{P}(T > t + s)}{\mathbb{P}(T > t)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = \mathbb{P}(T > s)$$

□

Para definir el **proceso de Poisson**, necesitamos conocer la distribución de una suma de variables aleatorias con distribución exponencial. Así pues:

Definición 2.3. *Una variable aleatoria T tiene una **distribución gamma**, $gamma(n, \lambda)$, si T tiene función de densidad:*

$$f_T(t) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t} \mathbb{1}_{[0,\infty)}(t)$$

Teorema 2.4. *Sean $t_1, t_2, ..., t_n$ variables aleatorias independientes con distribución $exp(\lambda)$. La suma $t_1 + ... + t_n$ tiene distribución $gamma(n, \lambda)$.*

Para desarrollar la sección sobre *cadenas de Markov en tiempo continuo*, necesitamos el siguiente resultado:

Proposición 2.5. *Sean $t_1, t_2, ..., t_n$ n variables aleatorias con distribución $exp(\lambda_i)$ respectivamente. Entonces:*

$$\mathbb{P}(min(t_1, ..., t_n) > t) = \mathbb{P}(t_1 > t, ..., t_n > t) = \prod_{i=1}^{n} \mathbb{P}(t_i > t) = \prod_{i=1}^{n} e^{-\lambda_i t} = e^{-(\lambda_1 + ... + \lambda_n)t}$$

Por lo que $min(t_1, ..., t_n)$ tiene distribución exponencial con parámetro $\lambda_1 + ... + \lambda_n$

Finalmente, definimos la **distribución de Poisson**.

Definición 2.6. *Una variable aleatoria discreta T tiene una **distribución de Poisson** con parámetro λ, $Poisson(\lambda)$, si T tiene función de probabilidad:*

$$\mathbb{P}(T = k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ para } k \in \mathbb{Z}_+$$

2.2. Definiciones del proceso de Poisson

Este apartado lo desarrollamos siguiendo el criterio de [Durret, 1999]. Para empezar daremos dos definiciones distintas del **proceso de Poisson**.

Definición 2.7. *Sean $t_1, t_2, ...$ variables aleatorias independientes con distribución $exp(\lambda)$. Sean $T_n = t_1 + ... + t_n$ con $n \geq 1$, siendo $T_0 = 0$. Consideramos $N_s = max\{n : T_n \leq s\}$. N_s es llamado el **proceso de Poisson** de parámetro λ.*

Definición 2.8. *Sea $\{N_s, s \geq 0\}$ una sucesión de variables aleatorias. Entonces si se cumple:*
(i) $N_0 = 0$,
(ii) $N_{t+s} - N_s \sim Poisson(\lambda t)$, y
(iii) Los incrementos de N_t según $t \geq 0$ son independientes.
*Entonces decimos que N_s es un **proceso de Poisson** de parámetro λ.*

Observación 2.9. Podemos pensar los t_n como el tiempo entre llegadas de un cliente a un servidor. Por lo que, $T_n = t_1 + ... + t_n$ es el tiempo en el que ha llegado el cliente n y N_s es el número de llegadas hasta el tiempo s.

Para justificar estas definiciones y entender la relación que tienen entre sí, destacamos los siguientes resultados:

Lema 2.10. *Sea N_s según establece la definición 2.7, entonces N_s tiene una distribución de Poisson de parámetro λs.*

Demostración. Observamos que $N(s) = n$ si, y solo si, $T_n \leq sT_{n-1}$, es decir, el cliente n llega antes del tiempo s, pero el cliente $n + 1$ llega después de s. Además, vemos que si

$T_n = t$, siendo $T_{n+1} > s$, entonces $t_{n+1} > s - t$. Con t_{n+1} independiente de T_n. Con todo esto tenemos que, usando el teorema 2.4:

$$\begin{aligned} P(N(s) = n) &= \int_0^s P(T_n = t)P(T_{n+1} > s | T_n = t)dt = \int_0^s P(T_n = t)P(t_{n+1} > s - t)dt = \\ &= \int_0^s \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}e^{-\lambda(s-t)}dt = \frac{\lambda^n}{(n-1)!}e^{-\lambda s}\int_0^s t^{n-1}dt = \\ &= e^{-\lambda s}\frac{(\lambda s)^n}{n!} \end{aligned}$$

Siendo ésta la función de probabilidad de la distribución de Poisson de parámetro λs. □

Lema 2.11. *Supongamos que N_t es un proceso de Poisson según la definición 2.7. Entonces:*

1. *Si consideremos $M_t = N_{t+s} - N_s$, con $t \geq 0$, entonces M es un proceso de Poisson con parámetro λ y es independiente de N_r, con $0 \leq r \leq s$.*
2. *Si $t_0 < t_1 < ... < t_n$, entonces $N_{t_k} - Nt_{k-1}$ son independientes para $k = 1, ..., n$ (N_s tiene incrementos independientes).*

Teorema 2.12. *Las definiciones de 2.7 y 2.8 son equivalentes.*

Demostración. Ya hemos visto, mediante los lemas 2.10 y 2.11, la primera implicación: 2.7 $\Rightarrow$ 2.8.
Ahora vemos que 2.7 $\Leftarrow$ 2.8. Sea T_n el tiempo en el que el último cliente ha llegado. Entonces vemos que:

1. Nos situamos en la primera llegada, es decir, cuando $n = 1$. Sabemos que ésta ocurre después del tiempo t si, y solo si, no hubo ninguna llegada en $[0, t]$. Usando la fórmula de la distribuión de Poisson tenemos:

   $$\mathbb{P}(T_1 > t) = \mathbb{P}(N(t) = 0) = e^{-\lambda t}$$

   Por lo que $T_1 = t_1$ es $exp(\lambda)$

2. Para $T_2 = t_1 + t_2 \Leftrightarrow t_2 = T_2 - T_1$,

   $$\begin{aligned} \mathbb{P}(t_2 > t | t_1 = s) &= \mathbb{P}(\text{no hubo llegadas en } (s, s+t] | t_1 = s) = \\ &= \mathbb{P}(N_{t+s} - N_s = 0 | N_r = 0 \text{ para } r < s, N_s = 1) = \\ &= \mathbb{P}(N_{t+s} - N_s = 0) = e^{-\lambda t} \end{aligned}$$

   Usando la propiedad (iii) de la segunda definición. Entonces, t_2 es $exp(\lambda)$ e independiente de t_1.

3. Si repetimos este argumento, vemos que $t_1, t_2, ...$ son $exp(\lambda)$ independientes.

Entonces, $N_s = max\{n : T_n \leq s\}$, con $T_n = t_1 + ... + t_n$, cumple la primera definición. □

En este punto, es automática la pregunta: *¿Por qué tiene importancia el Proceso de Poisson en la teoría de colas que nos ocupa?* Nuestra respuesta está basada en [Durret, 1999] y [Green, 2011]. Los sistemas de colas cotidianos acostumbran a ser muy complejos.

Para simplificar los modelos de estos sistemas, debemos hacer ciertas suposiciones sobre la distribución de probabilidad de la llegada de clientes y de la duración de los servicios. Las suposiciones más comunes y en ocasiones, como vemos más adelante, más ajustadas corresponden a las Markovianas. Esto se traduce en asumir que el número de llegadas hasta un momento determinado viene dado por una **distribución de Poisson** y que la duración del servicio viene dado por una **distribución exponencial**. Con ello se construye el modelo $M/M/s$ que veremos en detalle más adelante.
Observaremos que las suposiciones Markovianas son correspondidas por una gran parte de los sistemas de colas, y además así ha sido contrastado empíricamente múltiples veces según [Green, 2011]. Esto, unido a su sencillez, hace que el modelo $M/M/s$ sea el más usado para la modelización de sistemas de colas. Tomamos un ejemplo para ilustrar lo dicho en este último párrafo:

Ejemplo 2.13. Consideremos un restaurante, el servidor, y un conjunto de n potenciales clientes. Estos clientes de manera independiente deciden acudir al restaurante entre 12:00 y 13:00 con una probabilidad de $\frac{\lambda}{n}$ y que, si lo hacen, acuden a una hora escogida aleatoriamente entre el intervalo. La distribución binomial nos da la probabilidad de que exactamente k clientes acudan al servidor:

$$\frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$$

Que es igual a:

$$\frac{\lambda^k}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k} \tag{2.1}$$

Entonces vemos que:

1. El elemento $\frac{\lambda^k}{k!}$ no depende de n;
2. En el segundo término hay k elementos en el numerador y k elementos en el denominador. Por lo que lo podemos escribir como:

   $$\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n}$$

   Vemos fácilmente que si $n\to\infty$ entonces todo el segundo término converge a 1.
3. Es sabido que $\left(1-\frac{\lambda}{n}\right)^n$ converge a $e^{-\lambda}$ si $n\to\infty$
4. Por el mismo razonamiento que en 2., $\left(1-\frac{\lambda}{n}\right)^{-k}$ converge a 1 si $n\to\infty$

Por ende, (2.1) converge a

$$\frac{\lambda^k}{k!}\cdot 1\cdot e^{-\lambda}\cdot 1=\frac{e^{-\lambda}\lambda^k}{k!}$$

Tratándose así de una distribución de Poisson con parámetro λ.

3. Cadenas de Markov en tiempo continuo

3.1. Definiciones y ejemplos

Esta sección es desarrollada a partir de [Durret, 1999]. Antes de empezar con el cuerpo de esta sección, queremos recordar la definición de cadena de Markov a tiempo discreto:

Definición 3.1 (Propiedad de Markov)**.** *Sean $\{X_n\}_{n\geq 0}$ una sucesión de variables aleatorias con valores en $\{1, 2, ..., N\}$ y con $n \in \mathbb{Z}_+$. Decimos que $\{X_n\}_n$ tiene la* ***propiedad de Markov*** *si:*

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = p(i, j)$$

para cualquier valor de $i_0, i_1, ..., i_{n-1}, i, j$, y $n \in \mathbb{Z}_+$. Siendo $P(i, j) = (p(i, j), i, j)$ la ***matriz de probabilidades de transición*** *con:*

$$p(i, j) = \mathbb{P}(X_{n+1} = j | X_n = i)$$

Es decir, la matriz que da en la posición (i, j) la probabilidad de ir de i a j. De manera que, dado el estado actual X_n, los estados pasados son irrelevantes para predecir el estado siguiente X_{n+1}.

A partir de la definición en tiempo discreto, podemos conocer la definición en tiempo continuo:

Definición 3.2. *Sea $\{X_t\}_{t\geq 0}$ una cadena continua de variables aleatorias con $t \in \mathbb{R}_+$. Decimos que $\{X_t\}_{t\geq 0}$ es una* ***cadena de Markov en tiempo continuo*** *si para todos los $0 \leq s_0 < s_1 \cdots < s_n < s$ y todos los posibles estados $i_0, ..., i_n, i, j$ tenemos:*

$$\mathbb{P}(X_{t+s} = j | X_s = i, X_{s_n} = i_n, ..., X_{s_0} = i_0) = \mathbb{P}(X_t = j | X_0 = i)$$

De manera que, dado el estado actual, los estados pasados son irrelevantes para predecir el futuro.

A diferencia del caso discreto, en las cadenas de Markov en tiempo continuo no podemos definir una **probabilidad de transición** única.

Definición 3.3 (Probabilidad de transición)**.** *La* ***probabilidad de transición*** *de una cadena de Markov en tiempo continuo se define para cada $t > 0$ y se escribe:*

$$p_t(i, j) = \mathbb{P}(X_t = j | X_0 = i)$$

Tenemos el siguiente resultado:

Proposición 3.4 (Ecuación de Chapman-Kolmogorov)**.**

$$\sum_k p_s(i, k) p_t(k, j) = p_{s+t}(i, j)$$

Siendo $p_t(i, j)$ la probabilidad de transición y s y t tiempos > 0.

Demostración. De acuerdo con lo visto hasta ahora y usando la definición de probabilidad condicionada y de la propiedad de Markov:

$$P(X_{s+t} = j | X_0 = i) = \sum_k P(X_{s+t} = j, X_s = k | X_0 = i) =$$

$$= \sum_k P(X_{s+t} = j | X_s = k, X_0 = i) P(X_s = k | X_0 = i) = \sum_k p_t(k, j) p_s(i, k)$$

Habiendo usado que $\frac{\mathbb{P}(X_{s+t}=j, X_s=k, X_0=i)}{\mathbb{P}(X_0=i)} = \frac{\mathbb{P}(X_{s+t}=j | X_s=k, X_0=i)\mathbb{P}(X_s=k, X_0=i)}{\mathbb{P}(X_0=i)}$ en la segunda igualdad. □

La proposición 3.4 nos sugiere que si conocemos $p_t(i,j)$ para todo $t < t_0$, con un $t_0 > 0$ arbitrario, entonces podemos deducir $p_t(i,j)$ para todo t. En efecto, a continuación vemos como podemos reconstruir $p_t(i,j)$ para todo t a partir de su derivada en 0:

Definición 3.5 (Intensidad de transición)**.** *Si el límite* $\lim_{h\to 0} \frac{p_h(i,j)}{h}$ *existe, entonces definimos* ***intensidad de transición*** *de i a j como:*

$$q(i,j) = \lim_{h\to 0} \frac{p_h(i,j)}{h}$$

Definición 3.6. *Sea $q(i,j)$ la intensidad de transición de i a j de una cadena de Markov a tiempo continuo X_t. Definimos $\lambda_i = \sum_{j\neq i} q(i,j)$, el parámetro con el que X_t abandona el estado i. Entonces, suponiendo que $0 < \lambda_i < \infty$ (el caso $\lambda_i = \infty$ implica que se abandona el estado i inmediatamente y el caso $\lambda_i = 0$ que nunca abandonará), definimos la matriz con elementos*

$$r(i,j) = \frac{q(i,j)}{\lambda_i},$$

llamada la ***matriz de ruta*** *de X_t.*

Proposición 3.7. *Dadas las intensidades de transición $q(i,j)$ podemos construir una cadena de Markov a tiempo continuo que tiene estas intensidades de transición*

Demostración. Supongamos para simplificar que $\lambda_i > 0$ para todo i. Sea Y_n una cadena de Markov a tiempo discreto con probabilidad de transición $r(i,j)$, donde $r(i,j)$ es la matriz de ruta definida en la definición anterior. Esta cadena de Markov a tiempo discreto Y_n nos marcará el camino para determinar la cadena de Markov X_t en tiempo continuo que queremos. Para determinar cuánto tiempo el proceso permanece en cada estado, sean $\tau_0, \tau_1, \tau_2, ...$ variables aleatorias independientes con distribución $exp(1)$. Distinguimos:
En $t = 0$, el proceso se encuentra en el estado X_0. Permanece en este estado una cantidad de tiempo que es $exp(\lambda_{X_0})$. Así, definimos el tiempo que el proceso estará en X_0 como $t_1 = \frac{\tau_0}{\lambda_{X_0}}$
En un tiempo $t = t_1$, el proceso salta a X_1, estado en el que permanece durante un tiempo con parámetro λ_{X_1}. Así, definimos el tiempo que el proceso estará en X_1 como $t_2 = \frac{\tau_1}{\lambda_{X_1}}$.
De igual manera, en el tiempo $T_2 = t_1 + t_2$ el proceso salta al estado X_2, estado en el que permanece durante un tiempo con parámetro λ_{X_2}. Así, definimos el tiempo que el proceso estará en X_2 como $t_3 = \frac{\tau_2}{\lambda_{X_2}}$.
Así sucesivamente, definimos que la cantidad de tiempo que el proceso permanece en X_{n-1} como $t_n = \frac{\tau_{n-1}}{\lambda_{X_{n-1}}}$. Por lo que el proceso cambia al estado X_n en el tiempo

$$T_n = t_1 + \cdots + t_n$$

Así pues, si establecemos $T_0 = 0$, para $n \geq 0$ podemos definir la cadena de Markov X_t en tiempo continuo buscada como:

$$X_t = Y_n \text{ para } T_n \leq t \leq T_{n+1}$$

□

En [Takahara, 2017], [Durret, 1999] y [Mitrofanova, 2007] encontramos ejemplos que nos ilustran la relación existente entre las cadenas de Markov en tiempo continuo y el proceso de Poisson con los sistemas de colas. Los usamos para construir los siguientes ejemplos:

Ejemplo 3.8 (Modelo de colas como cadena de Markov)**.** Sea un sistema de colas con n servidores que atienden a clientes que esperan en una única cola, en el caso de encontrar todos los servidores ocupados. Asumimos que los clientes llegan conforme a un proceso de Poisson con parámetro λ y que los tiempos de servicio de los servidores corresponden a variables aleatorias independientes con distribución $exp(\mu)$.
Sea X_t una cadena de Markov que indica el número de clientes en el sistema (en la cola o siendo atendidos por los servidores) con intensidades de transición $q(i,j)$.
Buscamos $q(i,j)$ para las llegadas, es decir, buscamos $q(i,i+1)$ para todo estado i. Tenemos que la probabilidad de que se dé una llegada es igual a 1 menos la probabilidad de que haya 0 llegadas. Entonces, $p_h(i,i+1) = 1 - e^{-\lambda h}$. Así podemos escribir:

$$p_h(i,i+1) = 1 - e^{-\lambda h} = 1 - \sum_{n=0}^{\infty} \frac{(-\lambda h)^n}{n!} = -\sum_{n=1}^{\infty} \frac{(-\lambda h)^n}{n!} = \lambda h + o(h)$$

Donde $o(h)$ representa a los términos de orden mayor que h. Por tanto, siguiendo la definición 3.5,

$$q(i,i+1) = \lim_{h\to 0} \frac{p_h(i,i+1)}{h} = \lambda$$

Entonces, $\boxed{q(k,k+1) = \lambda}$, para todo $k \geq 0$.
Ahora buscamos $q(i,j)$ para las salidas, es decir, buscamos $q(i,i-1)$ para todo estado i. Suponemos que hay k servidores ocupados, con tiempos de servicio siguiendo exponenciales de parámetro μ. Considerando la propiedad 2.5 del mínimo de exponenciales, esta salida se producirá con ratio $k\mu$. Así pues:

$$\boxed{q(k,k-1) = \begin{cases} k\mu & \text{si } 0 \leq k \leq n; \\ n\mu & \text{si } k \geq n \end{cases}} \tag{3.1}$$

para todo $k \geq 1$, dado que como máximo hay n servidores atendiendo.
Dada 3.7, ya hemos definido la cadena de Markov relativa a este sistema.

Ejemplo 3.9. Como en el ejemplo anterior, supongamos que hay n servidores, que los clientes llegan conforme a un proceso de Poisson con parámetro λ y que los tiempos de servicio de los servidores corresponden a variables aleatorias independientes con distribución $exp(\mu)$. Suponemos, para simplificar el ejemplo, que estamos siempre en el caso $0 \leq k \leq n$ de la ecuación (3.1).
Queremos responder a la pregunta: *¿Cuántos servidores ocupados se encontrará el próximo cliente en llegar si hay k servidores ocupados?* Sea X_k la variable aleatoria que indica el número de servidores ocupados que encontrará el siguiente cliente si actualmente hay k servidores ocupados. Sea Y_k la **esperanza** de la variable aleatoria X_k. Distinguimos los casos:
$Y_0 = 0$. En efecto, dado que si no hay ningún servidor ocupado, en la siguiente llegada tampoco habrá ninguno.
Y_1. Si tenemos un servidor ocupado, el siguiente cliente encontrará 1 servidor ocupado si el tiempo que tarda en llegar es menor que el tiempo de servicio del servidor ocupado, en cambio, encontrará 0 servidores ocupados si el tiempo que tarda en llegar es mayor que el tiempo de servicio del servidor ocupado. Dada la propiedad 2.2 de falta de memoria, el tiempo de la siguiente llegada sigue una distribución $exp(\lambda)$ y el tiempo restante para completar el servicio sigue una $exp(\mu)$, por lo que, usando la proposición 2.5, $\lambda + \mu$ es el parámetro de la distribución exponencial del mínimo entre $exp(\lambda)$ y $exp(\mu)$. Finalmente:

$$Y_1 = (1)\frac{\lambda}{\lambda+\mu} + (0)\frac{\mu}{\lambda+\mu} = \frac{\lambda}{\lambda+\mu}$$

dado que la probabilidad de que el siguiente cliente encuentre 1 servidor ocupado es $\frac{\lambda}{\lambda+\mu}$ y la probabilidad de que el siguiente cliente encuentre 0 servidores ocupados es $\frac{\mu}{\lambda+\mu}$.
En general, para Y_k, si hay k servidores ocupados, tenemos $k+1$ exponenciales independientes: k $exp(\mu)$, para indicar el tiempo que queda para acabar los k servicios aun activos, y una $exp(\lambda)$ para el tiempo que tarda el siguiente cliente en llegar. Debemos dilucidar que sucede primero. Usamos la proposición 2.5 para ver que el mínimo de las k $exp(\mu)$ tienen distribución $exp(k\mu)$. Entonces, la probabilidad de que un servicio se complete antes de la llegada del siguiente cliente es la probabilidad de que una variable aleatoria $exp(k\mu)$ sea más pequeña que una variable aleatoria $exp(\lambda)$. Esta probabilidad es: $\frac{k\mu}{(k\mu+\lambda)}$. También, la probabilidad de que suceda la llegada antes que finalice un servicio es $\frac{\lambda}{(k\mu+\lambda)}$.
Además, si sucede antes la llegada del cliente: $Y_k = k$. En cambio, si sucede antes la finalización de un servicio: $Y_k = Y_{k-1}$. En efecto, dada la propiedad 2.2 de *memoryless* de la distribución exponencial: una vez suponemos que ha sucedido la finalización de un primer servicio antes que la llegada del cliente, para considerar la probabilidad de que se complete otro servicio antes de la llegada del cliente se *reinicia la cuenta* y se parte de Y_{k-1}, y así sucesivamente. Por tanto, tenemos

$$Y_k = Y_{k-1}\frac{k\mu}{k\mu+\lambda} + k\frac{\lambda}{k\mu+\lambda}$$

Para encontrar Y_n para cualquier $n \geq 2$ necesitamos resolver la recursión de la ecuación anterior. Así, primero vemos que

$$Y_2 = Y_1\frac{2\mu}{2\mu+\lambda} + \frac{2\lambda}{2\mu+\lambda} = \left(\frac{\lambda}{\mu+\lambda}\right)\left(\frac{2\mu}{2\mu+\lambda}\right) + \frac{2\lambda}{2\mu+\lambda}$$

y que

$$\begin{aligned} Y_3 &= Y_2\frac{3\mu}{3\mu+\lambda} + \frac{3\lambda}{3\mu+\lambda} = \\ &= \left(\frac{\lambda}{\mu+\lambda}\right)\left(\frac{2\mu}{2\mu+\lambda}\right)\left(\frac{3\mu}{3\mu+\lambda}\right) + \left(\frac{2\lambda}{2\mu+\lambda}\right)\left(\frac{3\mu}{3\mu+\lambda}\right) + \frac{3\lambda}{3\mu+\lambda} \end{aligned}$$

Así, a partir de los patrones que podemos observar, tenemos que

$$Y_n = \frac{n\lambda}{n\mu+\lambda} + \sum_{i=1}^{n-1}\frac{i\lambda}{i\mu+\lambda}\prod_{j=i+1}^{n}\frac{j\mu}{j\mu+\lambda}$$

Observamos los siguientes aspectos relevantes de la cadena $\{X_n\}_n$:

1. Cuando hay $i < n$ servidores ocupados, hay $i+1$ *relojes* que siguen una distribución exponencial independiente *corriendo*, siendo i de ellos de parámetro μ y 1 con parámetro λ. De esta manera, el tiempo por el que el proceso cambia al estado siguiente sigue una distribución exponencial de parámetro $i\mu+\lambda$. Si los n servidores están ocupados, entonces hay n distribuciones exponenciales corriendo y, así, el tiempo por el que sucede un cambio de estado sigue una distribución exponencial de parámetro $n\mu$.

2. Cuando el proceso se encuentra en el estado $i < n$, salta al estado $i+1$ con probabilidad $\lambda/(i\mu+\lambda)$ y al estado $i-1$ con probabilidad $i\mu/(i\mu+\lambda)$. Si los n servidores están ocupados, el proceso vuelve al estado $n-1$ con probabilidad $n\mu/n\mu = 1$.

3. Cuando el proceso se encuentra en i y cambia de estado, empiezan a correr de nuevo todos los *relojes* de las distribuciones que correspondan al nuevo estado.

A partir de esta descripción y usando las definiciones de este apartado, vemos que hemos hallado en $\{X_n\}_n$ una cadena de Markov a tiempo continuo.

3.2. Probabilidad de transición a partir de las intensidades de transición

El resultado más relevante del anterior apartado es la proposición 3.7 por la que hemos visto que, dadas las intensidades de transición $q(i,j)$, podemos construir una cadena de Markov en tiempo continuo que tiene estas intensidades de transición. A partir de [Durret, 1999], [Norris, 1997] y [Corcuera, 2019a], vemos el principal resultado de este apartado: las ecuaciones hacia delante y hacia atrás, tanto en el caso finito como en el infinito.

Empezamos este apartado presentando la relación entre las matrices de intensidades y las cadenas de Markov en tiempo continuo.

Definición 3.10 (Matriz de intensidades)**.** *Sea I un conjunto numerable. Una* ***matriz de intensidades*** *en I es una matriz $Q = (q(i,j)|i,j \in I)$ que satisface:*

1. $0 \leq -q(i,i) < \infty$ *para cualquier* $i \in I$*;*
2. $q(i,j) \geq 0$ *para cualquier* $i \neq j$ *de* I*;*
3. $\sum_{j \in I} q(i,j) = 0$ *para todo* $i \in I$*.*

Observación 3.11. La suma de todos los elementos, excepto el diagonal, de cada fila es finita:

$$q(i) = \sum_{j \neq i} q(i,j) < \infty$$

y el elemento de la diagonal $q(i,i)$ es $-q(i)$. Haciendo así que la suma de toda la fila sea cero.

Definición 3.12 (Matriz estocástica)**.** *Una matriz $P = (p(i,j)|i,j \in I)$ es estocástica si cumple:*

1. $0 \leq p(i,j) < \infty$ *para todo* $i,j \in I$*;*
2. $\sum_{j \in I} p(i,j) = 1$ *para todo* $i \in I$*.*

Observación 3.13. Si una matriz $P(t) = (p_t(i,j)|i,j \in I)$ tiene como elementos las probabilidades de transición de una cadena de Markov en tiempo continuo, entonces $P(t)$ es una matriz estocástica.
Si una matriz $Q = (q(i,j)|i,j \in I)$ tiene como elementos las intensidades de transición definidas en 3.5, entonces Q es una matriz de intensidades.

Observación 3.14. Sabemos que para una matriz $Q = (q(i,j)|i,j \in I)$, la serie:

$$\sum_{k=0}^{\infty} \frac{Q^k}{k!}$$

es convergente y su límite es e^Q. También que si dos matrices Q_1 y Q_2 conmutan, entonces

$$e^{Q_1+Q_2} = e^{Q_1} e^{Q_2}$$

Teorema 3.15. *Una matriz Q en un conjunto finito I es una matriz de intensidades si, y solo si, $P(t) = e^{tQ}$ es una matriz estocástica para todo $t \geq 0$.*

Demostración. Podemos encontrar la demostración de este teorema en la sección 2.1 de [Norris, 1997]. □

Teorema 3.16. *Sea Q una matriz en un conjunto finito I. Definimos $P(t) = e^{tQ}$. Entonces, el proceso $(P(t), t \geq 0)$ tiene las siguientes propiedades:*

1. *$P(s+t) = P(s)P(t)$ para todo s, t. Llamada propiedad del semigrupo.*
2. *$(P(t), t \geq 0)$ es la única solución de la ecuación diferencial de Kolmogorov hacia delante:*
$$\frac{d}{dt}P(t) = P(t)Q;\ P(0) = I \tag{3.2}$$
3. *$(P(t), t \geq 0)$ es la única solución de la ecuación diferencial de Kolmogorov hacia atrás:*
$$\frac{d}{dt}P(t) = QP(t);\ P(0) = I \tag{3.3}$$
4. *para $k = 0, 1, 2, ...$, tenemos*
$$\left(\frac{d}{dt}\right)^k_{|k=0} P(t) = Q^k \tag{3.4}$$

Demostración. Probamos el primer punto de la propiedad del semigrupo. Tenemos que para todo $s, t \in \mathbb{R}$, sQ y tQ conmutan. Por lo tanto
$$e^{sQ}e^{tQ} = e^{(s+t)Q}$$

Por la hipótesis tenemos que $P(t) = e^{tQ}$, es decir, podemos escribir
$$P(t) = \sum_{k=0}^{\infty} \frac{(tQ)^k}{k!}$$
Esta serie de potencias de matrices tiene un radio de convergencia infinito. Por lo que podemos derivar la serie componente a componente:
$$P'(t) = \sum_{k=1}^{\infty} \frac{t^{k-1}Q^k}{(k-1)!} = P(t)Q = QP(t)$$
Probando así las ecuaciones diferenciales de Kolmogorov hacia delante y hacia atrás. Si repitiéramos la derivación componente a componente, llegaríamos, de la misma manera, a demostrar el punto cuatro del teorema.
Falta por demostrar que $P(t)$ es la única solución de las ecuaciones hacia delante y hacia atrás. Supongamos que $M(t)$ es otra matriz que satisface la ecuación hacia delante, entonces
$$\begin{aligned}\frac{d}{dt}(M(t)e^{-tQ}) &= \left(\frac{d}{dt}M(t)\right)e^{-tQ} + M(t)\left(\frac{d}{dt}e^{-tQ}\right) = \\ &= M(t)Qe^{-tQ} + M(t)(-Q)e^{-tQ} = 0\end{aligned}$$
De esta manera, $M(t)e^{-tQ}$ es constante y, por tanto, $M(t) = P(t)$. De la misma manera podemos ver que $P(t)$ es la única solución de la ecuación hacia atrás. □

Teorema 3.17. *Sea I un conjunto finito. Dadas las probabilidades de transición $p_t(i,j)$ con $i,j \in I$, podemos definir la matriz de intensidades Q, entonces podemos construir una cadena de Markov en tiempo continuo con un conjunto finito I de posibles estados y matriz de probabilidades de transición $P(t) = (p_t(i,j)|i,j \in I)$. Entonces ésta cumple la ecuación diferencial de Kolmogorov hacia delante:*

$$P'(t) = QP(t) \tag{3.5}$$

Demostración. Empezamos tomando la ecuación de Chapman-Kolmogorov, vista en la proposición 3.4:

$$\sum_{k\in I} p_s(i,k)p_t(k,j) = p_{s+t}(i,j),$$

sacando el término $k = i$ fuera del sumatorio tenemos:

$$\begin{aligned} p_{t+h}(i,j) - p_t(i,j) &= \left(\sum_{k\in I} p_h(i,k)p_t(k,j)\right) - p_t(i,j) = \\ &= \left(\sum_{k\in I, k\neq i} p_h(i,k)p_t(k,j)\right) + [p_h(i,i) - 1]p_t(i,j) \end{aligned}$$

Dividimos ambos lados por h y hacemos $h \to 0$, obteniendo:
El término $p_{t+h}(i,j) - p_t(i,j)$:

$$\lim_{h\to 0} \frac{p_{t+h}(i,j) - p_t(i,j)}{h} = p_t^{'}(i,j), \tag{3.6}$$

por definición.
El término $\sum_{k\in I,k\neq i} p_h(i,k)p_t(k,j)$:

$$\lim_{h\to 0} \frac{\sum_{k\in I,k\neq i} p_h(i,k)p_t(k,j)}{h} = \sum_{k\in I,k\neq i} q(i,k)p_t(k,j), \tag{3.7}$$

usando la definición de intensidad de transición q(i,j).
El término $[p_h(i,i) - 1]p_t(i,j)$: Tenemos que

$$\lim_{h\to 0} \frac{p_h(i,i) - 1}{h} = -\lim_{h\to 0} \sum_{k\in I,k\neq i} \frac{p_h(i,k)}{h} = -\sum_{k\in I,k\neq i} q(i,k) = -\lambda,$$

usando que $1 - p_h(i,i) = \sum_{k\in I,k\neq i} p_h(i,k)$. Entonces:

$$\lim_{h\to 0} \frac{p_h(i,i) - 1}{h} p_t(i,j) = -\lambda_i p_t(i,j) \tag{3.8}$$

Finalmente usando las ecuaciones 3.6, 3.7 y 3.8, tenemos:

$$p_t^{'}(i,j) = \sum_{k\neq i} q(i,k)p_t(k,j) - \lambda_i p_t(i,j) = Qp_t(i,j)$$

□

A partir de este teorema, el teorema 3.16 para I finito nos sirve para ver que la ecuación hacia delante y la ecuación hacia atrás tienen la misma solución. Por tanto, la matriz de probabilidades de transición $P(t) = (p_t(i,j)|i,j \in I)$ también es solución de la ecuación hacia atrás:

$$P'(t) = P(t)Q$$

Este último resultado permite ver la ecuaciones diferenciales de Kolmogorov **hacia delante** y la de **hacia atrás**, pero en el caso de tener un conjunto de posibles estados I **finito**. Los teoremas que presentamos y demostramos a continuación nos sirven para demostrar el **caso infinito**.

Teorema 3.18. *Sea Q una matriz de intensidades. Entonces la ecuación hacia atrás*

$$P'(t) = QP(t);\ P(0) = I \tag{3.9}$$

tiene una solución no negativa $(P(t)|t \geq 0)$ y esta solución forma un semigrupo matricial, ya que cumple:

$$P(s)P(t) = P(s+t)$$

para cualquier $s, t \geq 0$.

Definición 3.19 (Semigrupo de una matriz de intensidades)**.** *Siguiendo el teorema anterior, llamamos a $(P(t)|t \geq 0)$ el **semigrupo** asociado a una matriz de intensidades Q.*

Teorema 3.20. *Sea $\{X_t\}_t$ una cadena de Markov a tiempo continuo con valores en I. Sea Q su matriz de intensidades y $P(t)$ su matriz de probabilidades de transición. Entonces, $P(t)$ satisface la ecuación (3.9) y es el semigrupo de Q.*

Demostramos a la vez los teoremas 3.18 y 3.20.

Demostración. Sea Q la matriz de intensidades de una cadena de Markov $\{X_t\}_t$ a tiempo continuo. Y sea $P(t)$ la matriz de transición, entonces los elementos de $P(t)$ son $p_t(i,j) = \mathbb{P}(X_t = j|X_0 = i)$. Vemos que $P(t)$ cumple las condiciones de los teoremas que queremos demostrar:

1. Vemos que $P(t)$ cumple la ecuación hacia delante. Supongamos que $X_0 = i$. Trabajamos a partir de un lema, cuya demostración encontramos en [Norris, 1997]:

   Lema 3.21. *Sea $\{X_t\}_t$ una cadena de Markov con probabilidades de transición $p_t(i,j) = \mathbb{P}(X_t = j|X_0 = i)$ y sea Q su matriz de intensidades. Usamos la notación:*

   $$q(i) = -q(i,i);\ m(i,k) = \frac{q(i,k)}{qi};\ \delta_{i,j} = \begin{cases} 1 \ si\ i = j \\ 0 \ si\ i \neq j \end{cases}$$

   Entonces

   $$p_t(i,j) = e^{-q(i)t}\delta_{ij} + \sum_{k \neq i} \int_0^t q(i) e^{-q(i)s} m(i,k) p_{t-s}(k,j) ds \tag{3.10}$$

   y

   $$e^{q(i)t} p_t(i,j) = \delta_{ij} + \int_0^t \sum_{k \neq i} q(i) e^{q(i)u} m(i,k) p_u(k,j) du \tag{3.11}$$

   *La ecuación (3.10) se llama **forma integral de la ecuación hacia atrás**.*

A partir de este lema y su ecuación (3.11), sabemos que $p_t(i,j)$ es continua en t por todo i,j, el integrando es una suma de funciones uniformemente convergente, por tanto, continua y $p_t(i,j)$ es diferenciable en t y cumple

$$e^{q(i)t}(q(i)p_t(i,j) + p_t^{'}(i,j)) = \sum_{k\neq i} q(i)e^{q(i)t}m(i,k)p_t(k,j)$$

Con la notación que hemos venido usando hasta ahora. De esta manera, podemos escribir

$$p_t^{'}(i,j) = \sum_{k\in I} q(i,k)p_t(k,j)$$

Por tanto, $P(t)$ cumple la ecuación hacia atrás.

2. $P(t)$ con $t \geq 0$ cumple la propiedad del semigrupo. En efecto, ya que, a partir de la propiedad de Markov, tenemos

$$p_{s+t}(i,j) = \mathbb{P}(X_{s+t} = j|X_0 = i) = \sum_{k\in I} \mathbb{P}(X_{s+t} = j|X_s = k, X_0 = i)\mathbb{P}(X_s = k|X_0 = i) =$$
$$= \sum_{k\in I} \mathbb{P}(X_s = k|X_0 = i)\mathbb{P}(X_t = j|X_0 = k) = \sum_{k\in I} p_s(i,k)p_t(k,j)$$

□

Acabamos de ver la ecuación hacia atrás en el caso infinito. Nos falta ver la ecuación hacia delante.

Teorema 3.22. *Una solución no negativa $(P(t)|t \geq 0)$ de la ecuación hacia atrás es también una solución no negativa de la ecuación hacia delante*

$$P'(t) = P(t)Q;\ P(0) = I$$

Demostración. Partimos de un lema demostrado en [Norris, 1997]:

Lema 3.23. *Sea $\{X_t\}_t$ una cadena de Markov con probabilidades de transición $p_t(i,j) = \mathbb{P}(X_t = j|X_0 = i)$ y sea Q su matriz de intensidades. Usamos la notación:*

$$q(i) = -q(i,i);\ m(i,k) = \frac{q(i,k)}{qi};\ \delta_{i,j} = \begin{cases} 1 \ si\ i = j \\ 0 \ si\ i \neq j \end{cases}$$

Entonces

$$p_t(i,j) = \delta_{ij}e^{-q(i)t} + \int_0^t \sum k \neq j p_{t-s}(i,k)q(k,j)e^{-q(j)s}ds \tag{3.12}$$

y

$$p_t(i,j)e^{q(j)t} = \lambda_{ij} + \int_0^t \sum_{k\neq j} p_u(i,k)q(k,j)e^{q(j)u}du \tag{3.13}$$

*La ecuación (3.12) se llama **forma integral de la ecuación hacia delante**.*

Del lema 3.21 de la demostración anterior, tenemos que $e^{q(i)t}p_t(i,k)$ es creciente para todo estado i,k. Por tanto, existen dos posibilidades: o bien, $\sum_{k\neq j} p_u(i,k)q(k,j)$ converge uniformemente para todo $u \in [0,t]$; o bien, $\sum_{k\neq j} p_u(i,k)q(k,j) = \infty$ para todo $u \geq t$. La última opción contradice a la ecuación (3.13) del lema anterior, dado que su componente

izquierda es finita para cualquier t.
De la demostración de 3.18 sabemos que $p_t(i,j)$ es continua para todo estado i,j. De esta manera, dada la convergencia uniforme, el integrando de la ecuación (3.13) es continuo y, por tanto, podemos diferenciarlo para obtener

$$p_t^{'}(i,j) + p_t(i,j)q(j) = \sum_{k \neq j} p_t(i,k)q(k,j)$$

Por tanto, vemos que $P(t)$ cumple la ecuación hacia delante. Tal y como queríamos demostrar. □

3.3. Comportamiento límite

En este apartado, siguiendo las ideas de [Norris, 1997] y [Durret, 1999], presentamos definiciones y resultados básicos que nos servirán para modelizar colas a partir de las cadenas de Markov en tiempo continuo.

Definición 3.24 (Recurrencia y transitoriedad)**.** *Sea $\{X_t\}_t$ una cadena de Markov en tiempo continuo con matriz de intensidades Q. Decimos que un estado i de la cadena es* ***recurrente*** *si*

$$\mathbb{P}(\textit{el conjunto } \{t \geq 0 | X_t = i\} \textit{ no está acotado}) = 1.$$

En cambio, decimos que un estado i es ***transitorio*** *si*

$$\mathbb{P}(\textit{el conjunto } \{t \geq 0 | X_t = i\} \textit{ no está acotado}) = 0.$$

Definimos el fenómeno por el que una cadena de Markov en tiempo continuo llega a recorrer un número infinito de estados durante un periodo finito de tiempo.

Definición 3.25 (Tiempo de explosión)**.** *Sean $J_0, J_1, J_2, ...$ los tiempos de transición de una cadena de Markov en tiempo continuo. Definimos tiempo de explosión ζ:*

$$\zeta = sup_n J_n$$

Definición 3.26. *Decimos que una cadena de Markov en tiempo continuo $\{X_t\}_t$ es explosiva si para algún estado $i \in I$ cumple:*

$$\mathbb{P}_i(\zeta < \infty) > 0$$

En este caso también decimos que su matriz de intensidades Q es explosiva.

Proposición 3.27. *Sea $\{X_t\}_t$ una cadena de Markov en tiempo continuo, en un conjunto I y con matriz de intensidades Q. Entonces $\{X_t\}_t$ no explota si cumple alguna de las siguientes condiciones:*

1. *I es finito.*
2. *$\sum_{i \in I} q(i) < \infty$, donde $q(i) = -q(i,i)$.*
3. *$X_0 = i$ y i es recurrente respecto la cadena.*

Demostración. La demostración de esta proposición la podemos encontrar en la sección 2.7 de [Norris, 1997]. □

Vemos que la recurrencia y la transitoriedad de la cadena de Markov en tiempo continuo $\{X_t\}_t$ están determinadas por un subconjunto en tiempo discreto de $\{X_t\}_t$.

Proposición 3.28. *Sea $\{X_t\}_t$ una cadena de Markov en tiempo continuo. Sea $h > 0$ y consideremos la cadena en tiempo discreto $Z_n = X_{nh}$. Entonces:*

1. *Si el estado i es recurrente para $\{X_t\}_t$ entonces i es recurrente para $\{Z_n\}_n$.*
2. *Si el estado i es transitorio para $\{X_t\}_t$ entonces i es transitorio para $\{Z_n\}_n$.*

Demostración. La demostración de esta proposición la podemos encontrar en la sección 3.4 de [Norris, 1997]. □

Definición 3.29 (Tiempo de parada)**.** *Sea $\{X_t\}_t$ una cadena de Markov en tiempo continuo. Sea T una variable aleatoria que toma valores en $[0, \infty) \cup \{\infty\}$. T es un tiempo de parada de $\{X_t\}_t$, si el evento $\{T \leq t\}$ depende solo de $(X_s | s \leq t)$, para todo $t \in [0, \infty)$.*

Teorema 3.30 (Propiedad fuerte de Markov)**.** *Sea $\{X_t\}_t$ una cadena de Markov a tiempo continuo con matriz de intensidades Q y sea T un tiempo de parada de esta cadena. Entonces, $\{X_{T+t}\}_t$, condicionada a $T < \infty$ y $X_T = i$, es una cadena de Markov a tiempo continuo con matriz de intensidades Q e independiente de $(X_s | s \leq T)$.*

Demostración. Podemos encontrar demostraciones de este teorema en las siguientes fuentes bibliográficas: [Neeman, 2019], [Khoshnevisan, 2011] y [Roch, 2012]. □

Definición 3.31 (Irreductibilidad)**.** *Sea $\{X_t\}_t$ una cadena de Markov en tiempo continuo. Decimos que X es **irreducible** si por cualquier estado x e y es posible ir de x a y en un número finito de transiciones. Es decir, si para cada pareja de estados x e y, existe una sucesión de estados $x_0 = x, x_1, ..., x_n = y$ tal que $q(x_{m-1}, x_m) > 0$ para $1 \leq m \leq n$*

Definición 3.32. *El vector de probabilidades π es la **distribución estacionaria** de una cadena de Markov a tiempo continuo X_t, si $\pi P(t) = \pi$ para todo $t > 0$.*

El siguiente resultado nos permite averiguar con más facilidad si un vector de probabilidades π es distribución estacionaria:

Teorema 3.33. *Sea Q la matriz de intensidades de una cadena de Markov en tiempo continuo $\{X_t\}_t$ recurrente e irreducible. π es una distribución estacionaria si, y solo si, $\pi Q = 0$.*

Demostración. Distinguimos el caso finito del infinito. En el caso finito, a partir de la ecuación diferencial de Kolmogorov hacia atrás:

$$\frac{d}{dt}\pi p_t = \pi p_t^{'} = \pi Q p_t$$

entonces si $\pi Q = 0 \Rightarrow$ la derivada de πp_t es cero y, por tanto, πp_t es constante. Además, este valor constante es π, dado que éste es su valor en $t = 0$. Por tanto, $\pi p_t = \pi$ para cualquier t.
En la otra dirección, a partir de la ecuación diferencial de Kolmogorov (3.5), multiplicando por $\pi(i)$ y sumando por cada i tenemos:

$$\sum_{i \in I} \pi(i) p_t^{'}(i, j) = \sum_{i,j \in I} \pi(i) p_t(i, k) Q(k, j)$$

con I, conjunto de estados posibles, finito. Suponemos que $\pi p_t = \pi$. Observamos que en la expresión de la izquierda de la última ecuación tenemos

$$\frac{d}{dt}\left(\sum_{t\in I}\pi(i)p_t(i,j)\right) = \frac{d}{dt}\pi(j) = 0.$$

En la expresión de la derecha tenemos que, usando $\sum_{i\in I}\pi(i)p_t(i,k) = \pi(k)$,

$$0 = \sum_{k\in I}\pi(k)Q(k,j) = (\pi Q)_j$$

Siendo $(\pi Q)_j$ el elemento j-ésimo del vector πQ. Entonces, así hemos visto que $\pi Q = 0$ en el caso finito.
Queda por ver el caso I infinito, dado que el intercambio de la derivada con el sumatorio en general no es posible en casos no finitos. Dado que Q es recurrente también es no explosiva, debida la proposición 3.27. A la vez, por la proposición 3.28, tenemos que $P(t)$ es recurrente. Por tanto, un π satisfaciendo $\pi Q = 0$ o $\pi P(t) = \pi$ es único, excepto por escalares múltiples.
Fijamos un estado i y definimos el vector $\mu = (\mu_j | j \in I)$ con

$$\mu_j = \mathbb{E}\int_0^{T_i} \mathbb{1}_{\{X_t=j\}}dt$$

siendo T_i un tiempo de parada. En este punto, presentamos un lema, demostración del cual podemos encontrar en la sección 3.5 de [Norris, 1997], para poder continuar con la demostración.

Lema 3.34. *Con un estado i fijado, tenemos:*

$$\mu_j = \mathbb{E}\int_0^{T_i} \mathbb{1}_{\{X_t=j\}}dt \Rightarrow \mu Q = 0$$

Por tanto, tenemos que $\mu Q = 0$. Así pues, si vemos que $\mu P(t) = \mu$, dada la unicidad salvo múltiples, habremos acabado la demostración. Por la propiedad fuerte de Markov (teorema 3.30) en T_i, podemos escribir

$$\mathbb{E}_i\int_0^{s} \mathbb{1}_{\{X_t=j\}}dt = \mathbb{E}_i\int_{T_i}^{T_i+s} \mathbb{1}_{\{X_t=j\}}dt$$

Finalmente por el teorema de Fubini, que nos permite calcular el valor de una integral múltiple, tenemos

$$\mu_j = \mathbb{E}_i\int_s^{s+T_i} \mathbb{1}_{\{X_t=j\}}dt = \int_0^{\infty}\mathbb{P}_i(X_{s+t}=j, t<T_i)dt =$$
$$= \int_0^{\infty}\sum_{k\in I}\mathbb{P}_i(X_t=k, t<T_i)p_s(k,j)dt = \sum_{k\in I}\left(\mathbb{E}_i\int_0^{T_i}\mathbb{1}_{\{X_t=k\}}dt\right)p_s(k,j) = \sum_{k\in I}\mu_k p_s(k,j)$$

Entonces, $\sum_{k\in I}\mu_k p_s(k,j) = \mu_j$. Dado que $\mu = (\mu_j | j \in I)$, tenemos $\mu P(t) = \mu$. Como queríamos demostrar. □

Teorema 3.35. *Si una cadena de Markov en tiempo continuo X_t es irreducible y tiene una distribución estacionaria π, entonces*

$$\lim_{t\to\infty} p_t(i,j) = \pi(j)$$

Demostración. Podemos encontrar una demostración de este teorema en la sección quinta de [Whitt, 2013]. □

Teorema 3.36. *Consideremos $q(i,j)$ las intensidades de transición de una cadena de Markov en tiempo continuo. Sea π un vector de probabilidades que cumple:*

$$\pi(k)q(k,j) = \pi(j)q(j,k) \text{ para todo } j,k \tag{3.14}$$

Entonces, π es una distribución estacionaria.

Demostración. Sumando la ecuación (3.14) para todo $k \neq j$ y usando la definición de λ_j, tenemos:

$$\sum_{k\neq j} \pi(k)q(k,j) = \pi(j)\sum_{k\neq j} q(j,k) = \pi(j)\lambda_j$$

Con lo que tenemos

$$(\pi Q)_j = \sum_{k\neq j} \pi(k)q(k,j) - \pi(j)\lambda_j = 0$$

□

Definición 3.37. *La ecuación (3.14) del teorema anterior se llama* ***condición de equilibrio****.*

4. Los sistemas de colas M/M/s

4.1. Procesos de nacimiento y muerte

Este apartado está fundamentado a partir de [Shortle u. a., 2018] y [Durret, 1999]. Un proceso de **nacimiento y muerte** es un tipo específico de cadena de Markov a tiempo continuo. Su estudio nos permite encontrar con facilidad la distribución estacionaria de modelos de colas pertenecientes a procesos de nacimiento y muerte. Algunos de estos modelos son los $M/M/s$, con los que nos centraremos en el siguiente apartado.
Un proceso de nacimiento y muerte consiste en un conjunto de estados $\{0,1,2,...\}$, normalmente llamados población. Cuando se produce un nacimiento en el estado n, el proceso pasa del estado n al estado $n+1$; y si se produce una muerte, al estado $n-1$. Siguiendo la línea de todo el trabajo, supondremos que los nacimientos se producen siguiendo una distribución exponencial de parámetro $\lambda_n > 0$ y las muertes una exponencial de parámetro $\mu_n > 0$, para cada n. En la teoría de colas, usamos estos procesos para hablar de llegadas de clientes en vez de nacimiento y de salidas de clientes en vez de muertes, considerando $\lambda_n = \lambda$ y $\mu = \mu_n$ para todo n.
Si escribimos la condición de equilibrio tenemos:

$$\pi(n) = \frac{\lambda_{n-1}}{\mu_n}\pi(n-1)$$
$$\pi(n-1) = \frac{\lambda_{n-2}}{\mu_{n-1}}\pi(n-2)$$

Por lo que

$$\pi(n) = \frac{\lambda_{n-1}}{\mu_n} \cdot \frac{\lambda_{n-2}}{\mu_{n-1}} \cdot \pi(n-2)$$

Así, repitiendo el proceso, tenemos que:

$$\boxed{\pi(n) = \frac{\lambda_{n-1} \cdot \lambda_{n-2} \cdots \lambda_0}{\mu_n \cdot \mu_{n-1} \cdots \mu_1} \pi(0) = \pi(0) \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}} \tag{4.1}$$

Dado que se trata de probabilidades, la suma total debe dar 1:

$$\sum_{n=0}^{\infty} \pi(n) = \pi(0) + \sum_{n=1}^{\infty} \left(\pi(0) \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right) =$$

$$= \pi(0) + \pi(0) \sum_{n=1}^{\infty} \left(\prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right) = \pi(0) \left(1 + \sum_{n=1}^{\infty} \left(\prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right) \right) = 1$$

Así pues, $\pi(0)$ debe ser:

$$\boxed{\pi(0) = \left(1 + \sum_{n=1}^{\infty} \left(\prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right) \right)^{-1}} \tag{4.2}$$

De esto se desprende que necesaria y suficientemente para la existencia de la distribución estacionaria necesitamos que la siguiente serie infinita converja:

$$\sum_{n=1}^{\infty} \left(\prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right) \tag{4.3}$$

Las ecuaciones (4.1) y (4.2) y la condición (4.3) son de gran utilidad para el estudio de sistemas de colas tal y como podemos ver en la siguiente sección.

4.2. Distribución estacionaria en los modelos de colas M/M/s

En este apartado, presentaremos ejemplos que podemos encontrar en [Durret, 1999] sobre la aplicación de la teoría de cadenas de Markov a tiempo continuo en los modelos de colas $M/M/s$, a la vez que usando razonamientos incluidos en [Shortle u. a., 2018].
Nuestro objetivo será encontrar la distribución estacionaria introducida en el apartado anterior. La distribución estacionaria es un vector que en su posición n nos dice cuál es la probabilidad de que haya n clientes utilizando el servicio o ocupando la cola (n es el número de llegadas). Además nos permite calcular otros valores relevantes.
Por [Omahen und Marathe, 1975], [Green, 2011] y otras fuentes mencionadas hasta ahora, un modelo $M/M/s$ consiste en:

1. Hay s servidores idénticos, cada uno capaz de realizar un solo servicio a la vez;
2. hay una única cola;
3. los clientes llegan de manera independiente entre ellos y siguiendo un proceso de Poisson con un parámetro constante;
4. y la duración de los servicios es independiente entre ellos y tienen una distribución exponencial.

Las dos últimas suposiciones son llamadas *Markovianas* y el número de servidores se suele denotar con s, por eso las dos M y la s en el nombre del modelo.
Suponemos durante toda la sección que el parámetro del proceso de Poisson de las llegadas es λ y que el parámetro de la distribución exponencial de los tiempos de servicio es μ.

Ejemplo 4.1 (Modelo M/M/1)**.** Siguiendo las suposiciones y el ejemplo 3.8, tenemos que las intensidades de transición son:

$$q(n, n+1) = \lambda \text{ si } n \geq 0$$
$$q(n, n-1) = \mu \text{ si } n \geq 1$$

Como hemos señalado en el apartado anterior, estamos en un caso particular de las cadenas de *nacimiento y muerte*, puesto que representa que la intensidad del nacimiento de individuos es $\lambda > 0$ y la intensidad de la muerte de individuos es $\mu > 0$. A partir de los pasos del apartado anterior de Procesos de nacimiento y muerte y dado que en el caso que estamos trabajando $\lambda_n = \lambda$ y $\mu_n = \mu$ para todo n:

$$\pi(n) = \left(\frac{\lambda}{\mu}\right)^n \pi(0)$$

Nos falta encontrar el valor de $\pi(0)$. Para ello, necesitamos suponer que $\lambda < \mu \Rightarrow \left(\frac{\lambda}{\mu}\right) < 1$. Así

$$\sum_{n=0}^{\infty} \pi(n) = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \pi(0) = \frac{\pi(0)}{1 - (\lambda/\mu)}$$

por lo que, para tener $\sum_{n=0}^{\infty} \pi(n) = 1$, tenemos que tomar $\pi(0) = 1 - \frac{\lambda}{\mu}$. Concluimos entonces que la distribución estacionaria es:

$$\boxed{\pi(n) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \text{ para } n \geq 0}$$

En el caso de $\lambda > \mu$, tenemos que no podemos encontrar ningún valor de $\pi(0)$ para hacer la suma 1. En este caso, no hay una distribución estacionaria para nuestro modelo y decimos que se trata de un modelo **transitorio**.

Ejemplo 4.2. Siguiendo con el escenario planteado por el ejemplo anterior, nos preguntamos: *¿cuál es la distribución de probabilidad del tiempo de espera W de un cliente?*
En el caso que nos ocupa, estamos ante una ley de probabilidad *mixta*, con masa de probabilidad en el caso $W = 0$ y con función de densidad en $W \in (0, +\infty)$. Tenemos que para que el tiempo de espera sea 0, necesitamos que en la cola haya $Q = 0$ personas esperando. Entonces escribimos:

$$\mathbb{P}(W = 0) = \mathbb{P}(Q = 0) = 1 - \frac{\lambda}{\mu}$$

En el caso de que ya haya alguien esperando en la cola, escribimos $f_W(x)$ para referirnos a la función de densidad del tiempo de espera W en $(0, \infty)$. Si hay n personas esperando en la cola, por lo visto en el teorema 2.4, el tiempo de espera tiene una densidad $gamma(n, \mu)$ y

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

Finalmente, haciendo un cambio de variables $m = n - 1$:

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

Así pues, podemos afirmar que la función de distribución de W, condicionada a $W > 0$, es exponencial con parámetro $\mu - \lambda$.

Ejemplo 4.3 (Modelo M/M/1 con una sala de espera finita)**.** En este caso, además, suponemos que la sala de espera tiene una capacidad finita de N clientes y un cliente encontrándose $\geq N$ clientes en la cola desiste y abandona. Así tenemos las siguientes intensidades de transición:

$$q(n, n+1) = \lambda \text{ si } 0 \leq n < N$$
$$q(n, n-1) = \mu \text{ si } 0 < n \leq N$$

Siendo los estados posibles del sistema $S = \{0, 1, ..., N\}$.
Tal y como sucede en el primer ejemplo, la distribución estacionaria para $1 \leq n \leq N$ es:

$$\pi(n) = \left(\frac{\lambda}{\mu}\right)^n \pi(0)$$

Pero, en cambio para encontrar $\pi(0)$, hacemos:

$$\sum_{n=0}^{N} \pi(n) = \sum_{n=0}^{N} \left(\frac{\lambda}{\mu}\right)^n \pi(0) \tag{4.4}$$

Dado que para $N < \infty$ y $\theta \neq 1$, $\sum_{n=0}^{N} \theta^n = \frac{1-\theta^{N+1}}{1-\theta}$, si suponemos que $\lambda \neq \mu$, para que la suma de 4.4 dé 1 necesitamos tomar $\pi(0) = \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{N+1}}$. Así pues, finalmente:

$$\boxed{\pi(n) = \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{N+1}} \left(\frac{\lambda}{\mu}\right)^n \text{ para } 0 \leq n \leq N}$$

Para el caso $\lambda = \mu$, como $\sum_{n=0}^{N} 1^n = N + 1$, tenemos que tomar $\pi(0) = \frac{1}{N+1}$, así pues:

$$\boxed{\pi(n) = \frac{1}{N+1} \text{ para } 0 \leq n \leq N}$$

Ejemplo 4.4 (Modelo M/M/s)**.** En este caso, a diferencia con el primer ejemplo, tenemos $s \geq 2$ servidores. No hay limitación de la capacidad de la cola. En este caso, claramente observamos que las intensidades de transición son:

$$q(n, n+1) = \lambda$$

$$q(n, n-1) = \begin{cases} \mu n \text{ si } n \leq s \\ \mu s \text{ si } n \geq s \end{cases}$$

La diferencia con los anteriores ejemplos es que hasta un máximo de s clientes pueden ser atendidos a la vez.
De la misma manera que en el caso $M/M/1$, a partir de los pasos del apartado anterior de Procesos de nacimiento y muerte y dado que en el caso actual

$$\lambda_n = \lambda \text{ y } \mu_n = q(n, n-1) = \begin{cases} \mu n \text{ si } n \leq s \\ \mu s \text{ si } n \geq s \end{cases}$$

para todo n:

$$\boxed{\pi(n) = \begin{cases} \frac{\lambda^n}{n!\mu^n}\pi(0) \text{ para } 0 \leq n < s \\ \frac{\lambda^n}{s^{n-s}s!\mu^n}\pi(0) \text{ para } n \geq s \end{cases}}$$

Para encontrar $\pi(0)$, usamos la condición de que la suma de probabilidades debe dar 1 y seguimos la fórmula 4.2 del Proceso de nacimiento y muerte:

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

Viendo que $\sum_{n=s}^{\infty} \frac{\lambda^n}{s^{n-s}s!\mu^n} = \frac{(\lambda/\mu)^s}{s!} \sum_{n=s}^{\infty} \left(\frac{\lambda/\mu}{s}\right)^{n-s} = \frac{(\lambda/\mu)^s}{s!} \sum_{m=0}^{\infty} \left(\frac{\lambda/\mu}{s}\right)^{m} = \frac{(\lambda/\mu)^s}{s!} \frac{1}{1-\frac{\lambda/\mu}{s}}$ podemos escribir:

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

De la última ecuación podemos llegar a la conclusión que busca este ejemplo. Si $\lambda < s\mu$ $\Rightarrow \frac{\lambda}{s\mu} < 1$, entonces es posible tomar un $\pi(0)$ tal que la suma fuera 1 (como lo que hemos hecho en los ejemplos anteriores). Podemos afirmar pues:

1. Si $\lambda < s\mu$, es decir, si el parámetro de la distribución del tiempo de servicio con todos los s servidores llenos es más grande que el parámetro de la distribución de la llegada, el modelo $M/M/s$ tiene una distribución estacionaria. En cambio, si $\lambda > s\mu$, el modelo $M/M/s$ es transitorio, es decir, no tiene distribución estacionaria.

2. El modelo $M/M/s$ con s servidores con parámetro de servicio μ aprovecha menos la capacidad disponible (es menos eficiente) que un modelo $M/M/1$ con 1 servidor con parámetro de servicio $s\mu$, ya que el parámetro de las salidas $q(n, n-1)$ es en algunos casos $n\mu$ en el primer caso, pero, en cambio, en el segundo caso, es siempre $s\mu$ siendo $n < s$.

Acabamos esta sección con un ejemplo que nos sirve para ilustrar otras propiedades del modelo $M/M/s$. Suponemos que $s = \infty$, modelo que sirve, por ejemplo, para estudiar el tráfico telefónico y determinar cuántas líneas necesitamos para tener la mayor parte del tiempo suficiente capacidad.

Ejemplo 4.5 (Modelo M/M/∞)**.** En este modelo suponemos que hay infinitos servidores disponibles, por lo que nunca se forma cola y cada cliente es atendido inmediatamente. Claramente tenemos:

$$q(n, n+1) = \lambda \text{ si } n \geq 0$$
$$q(n, n-1) = \mu n \text{ si } n \geq 1$$

Así, a partir del primer ejemplo de esta sección y la fórmula 4.1 tenemos que:

$$\pi(n) = \frac{\lambda_{n-1} \cdots \lambda_0}{\mu_n \cdots \mu_1} \cdot \pi(0) = \frac{\lambda^n}{\mu^n \cdot n \cdot (n-1) \cdots 2 \cdot 1} \cdot \pi(0) = \frac{(\lambda/\mu)^n}{n!} \cdot \pi(0)$$

Para determinar el valor de $\pi(0)$ buscamos el que hace que la suma $\sum_{n=0}^{\infty} \pi(n) = \sum_{n=0}^{\infty} \left(\frac{(\lambda/\mu)^n}{n!}\right) \cdot \pi(0)$ sea 1. Dado que $\sum_{n=0}^{\infty} \frac{(\lambda/\mu)^n}{n!} = e^{\lambda/\mu}$, $\pi(0)$ debe ser $\pi(0) = e^{-\lambda/\mu}$. Por tanto, la distribución estacionaria queda:

$$\boxed{\pi(n) = e^{-\lambda/\mu} \frac{(\lambda/\mu)^n}{n!} \text{ para } n \geq 0}$$

4.3. Fórmulas de Little

Introducimos las siguientes variables aleatorias: T_q el tiempo que el cliente pasa en la cola; T el tiempo total que el cliente pasa en el sistema ($T = T_q + S$, donde S es el tiempo de servicio); y $W_q = \mathbb{E}[T_q]$ y $W = \mathbb{E}[T]$, el tiempo esperado (o media del tiempo) que un cliente pasará en la cola y en el sistema, respectivamente. Finalmente, denotamos con N y N_q las variables aleatorias sobre el número de clientes en el sistema y en la cola, respectivamente, estando éste en el estado estacionario y denotemos con L y L_q sus valores esperados. En este apartado demostraremos las **fórmulas de Little**: $L = \lambda W$ y $L_q = \lambda W_q$. Serán una herramienta muy útil para desarrollar el apartado siguiente de medidas de efectividad. Trabajamos a partir de [Sigman, 2009], [Stidham, 1972] y [Shortle u. a., 2018].
Primero enunciamos las fórmulas de Little, para luego introducir y demostrar resultados que nos servirán como prueba para las fórmulas. Esta sección se desarrolla a partir de un análisis *omega a omega*, por lo que trabajamos para cualquier $\omega \in \Omega$, siendo Ω el espacio muestral.

Teorema 4.6 (Fórmulas de Little)**.**

$$L = \lambda W$$

y

$$L_q = \lambda W_q$$

Definición 4.7. *Si indexamos a los clientes con $n \in \mathbb{N}$, su ausencia o presencia en el sistema para cualquier $\omega \in \Omega$ puede ser indicado por la siguiente variable aleatoria:*

$$\mathbb{I}_n(t) = \begin{cases} 1, & \text{si el cliente } n \text{ está en el sistema en el momento } t \\ 0, & \text{en caso contrario} \end{cases}$$

Definición 4.8. *Definimos los siguientes procesos estocásticos:*

1. $\{L(t), t \geq 0\}$, *donde $L(t)$ es el número de clientes presentes en el sistema en el momento t.*
2. $\{W_n, n \in \mathbb{N}\}$, *donde W_n es el tiempo que pasa en el sistema el cliente n.*
3. $\{W_n^\beta, n \in \mathbb{N}\}$, *para cada $\beta \geq 0$ donde W_n^β es el tiempo con un factor exponencial que definimos así:*
$$W_n^\beta = \int_0^\infty e^{-\beta t} \mathbb{I}_n(t) dt$$

Observación 4.9. Dadas las definiciones anteriores observamos que para cualquier $\omega \in \Omega$:

1. $L(t) = \sum_{n=1}^\infty \mathbb{I}_n(t)$ para $t \geq 0$.
2. $W_n = \int_0^\infty \mathbb{I}_n(t) dt$ para $n \in \mathbb{N}$.

Teorema 4.10. *Para cualquier $\omega \in \Omega$ y $\beta > 0$, la integral $\int_0^\infty e^{-\beta t} L(t)$ existe, aunque su valor puede ser infinito. Además*

$$\int_0^\infty e^{-\beta t} L(t) dt = \sum_{n=1}^\infty W_n^\beta \tag{4.5}$$

Demostración. A partir de las definiciones de $L(t)$ y W_n^β, del hecho que $\mathbb{I}_n(t)$ es siempre no negativa y que, fijado $\omega \in \Omega$, es $Borel - medible$ respecto a $t \geq 0$, esta demostración es inmediata por el teorema de Fubini. □

Definimos el siguiente conjunto: $A_n = \{t \geq 0 | \mathbb{I}_n(t) = 1\}$, dado un $\omega \in \Omega$.

Corolario 4.11. *Para todo $\omega \in \Omega$ y $\beta \geq 0$, si existe una sucesión $\{t_n, n \in \mathbb{N}\}$ no negativa y no decreciente, tal que $A_n = [t_n, t_n + W_n)$ para $n \in \mathbb{N}$, entonces:*

$$\int_0^\infty e^{-\beta t} L(t) dt = \sum_{n=1}^\infty e^{-\beta t_n}(1 - e^{-\beta W_n})/\beta$$

Demostración. De manera directa a partir del teorema anterior tenemos:

$$\int_0^\infty e^{-\beta t} L(t) dt = \sum_{n=1}^\infty \int_{t_n}^{t_n + W_n} e^{-\beta t} dt =$$
$$= \sum_{n=1}^\infty e^{-\beta t_n} \int_0^{W_n} e^{-\beta t} dt = \sum_{n=1}^\infty e^{-\beta t_n}(1 - e^{-\beta W_n})/\beta$$

□

Antes de llegar a la demostración de las fórmulas de Little, necesitamos enunciar y demostrar dos lemas previos.

Definición 4.12. *Supongamos que existe una sucesión no decreciente y no negativa $\{t_n, n \in \mathbb{N}\}$, donde t_n es el momento de la llegada del cliente n. Supongamos también que A_n es un intervalo: $A_n = [t_n, t_n + W_n)$ para $n \in \mathbb{N}$. Sea $t_0 = 0$ y definimos $N(t) = max\{n | t_n \leq t\}$, es decir, el número de llegadas en $[0, t]$, $t \geq 0$. Si existen, definimos los siguientes valores:*

1. $\hat{L} = \lim_{T\to\infty}(1/T)\int_0^T L(t)dt$.
2. $\hat{\lambda} = \lim_{T\to\infty} \frac{N(T)}{T}$.
3. $\hat{W} = \lim_{N\to\infty}(1/N)\sum_{n=1}^\infty W_n$.

Lema 4.13. *Para todo $\omega \in \Omega$ y para todo $\hat{\lambda}$, $0 \leq \hat{\lambda} \leq \infty$:*

$$\lim_{t\to\infty} \frac{N(t)}{t} = \hat{\lambda} \Leftrightarrow \lim_{n\to\infty} \frac{t_n}{n} = \frac{1}{\hat{\lambda}}$$

Demostración. Empezamos demostrando la implicación de la derecha. Supongamos que $\lim_{t\to\infty} \frac{N(t)}{t} = \hat{\lambda}$. Distinguimos dos casos: si existe un $T < \infty$ tal que $t_n \leq T$ para todo $n \in \mathbb{N}$, entonces $N(t) = \infty$ para todo $t \geq T$. Por consiguiente, $\hat{\lambda} = \infty$ y $\lim_{n\to\infty} \frac{t_n}{n} = 0 = \frac{1}{\hat{\lambda}}$. En el caso contrario, si $t_n \to \infty$ si $n \to \infty$, entonces $\lim_{n\to\infty} \frac{N(t_n)}{t_n} = \hat{\lambda}$. Por tanto, $\lim_{n\to\infty} \frac{t_n}{n} = \frac{1}{\hat{\lambda}}$, ya que $N(t_n) = n$. Demostrando así la primera implicación.
Demostramos la implicación de la izquierda. Supongamos que $t_n/n \to 1/\hat{\lambda}$ cuando $n \to \infty$. Distinguimos dos casos: si $N(T) = \infty$ para algún $T < \infty$, entonces $t_n \leq T$ para cualquier

$n \in \mathbb{N}$, $1/\hat{\lambda} = \lim_{n\to\infty} t_n/n = 0$ y $\lim_{t\to\infty} N(t)/t = \infty = \hat{\lambda}$. En cambio, si $N(t) < \infty$ para todo $t \geq 0$, entonces tenemos la siguiente desigualdad $t_{N(t)} \leq t \leq t_{N(t)+1}$. Por tanto,

$$\frac{t_{N(t)}}{N(t)} \leq \frac{t}{N(t)} \leq \left(\frac{t_{N(t)+1}}{N(t)+1}\right)\left(\frac{N(t)+1}{N(t)}\right) \tag{4.6}$$

Como t_n está bien definida y es finita para todo $n \in \mathbb{N}$, entonces $N(t) \to \infty$ si $t \to \infty$. Así, finalmente, dado que $\lim_{t\to\infty} \frac{N(t)+1}{N(t)} = 1$ y que $\lim_{t\to\infty} \frac{t_{N(t)}}{N(t)} = \lim_{t\to\infty} \frac{t_{N(t)+1}}{N(t)+1} = \lim_{n\to\infty} \frac{t_n}{n} = \frac{1}{\hat{\lambda}}$ y dadas las desigualdades de la (4.6), tenemos $\lim_{t\to\infty} \frac{N(t)}{t} = \hat{\lambda}$. □

Lema 4.14. *Para todo $\omega \in \Omega$, si $\hat{\lambda} < \infty$ y $\hat{W} < \infty$, entonces*

$$\lim_{N\to\infty} \frac{W_N}{t_N} = 0$$

Demostración. Para cualquier $N \in \mathbb{N}$, tenemos $\frac{W_N}{t_N} = \frac{W_N}{N}\frac{N}{t_N}$. Entonces,

$$\begin{aligned} \frac{W_N}{N} &= \frac{\sum_{n=1}^{N} W_n - \sum_{n=1}^{N-1} W_n}{N} \leq \\ &\leq \left(\frac{\sum_{n=1}^{N} W_n}{N}\right) - \left(\frac{\sum_{n=1}^{N-1} W_n}{N}\right) \end{aligned} \tag{4.7}$$

Dado que $\lim_{N\to\infty}(N-1)/N = 1$ y que

$$\lim_{N\to\infty} \frac{\sum_{n=1}^{N} W_n}{N} = \lim_{N\to\infty} \frac{\sum_{n=1}^{N-1} W_n}{N} = \lim_{N\to\infty} \frac{\sum_{n=1}^{N-1} W_n}{N-1} = \hat{W} < \infty$$

Pasando a límite (4.7) tenemos

$$\lim_{N\to\infty} \frac{W_N}{N} = 0$$

A partir del lema 4.13, $\lim_{N\to\infty} \frac{N}{t_N} = \hat{\lambda} < \infty$, entonces

$$\lim_{N\to\infty} \frac{W_N}{t_N} = 0$$

□

Teorema 4.15. *Si $\hat{\lambda} < \infty$ y $\hat{W} < \infty$, entonces $\hat{L} < \infty$ y*

$$\hat{L} = \hat{\lambda} \cdot \hat{W}$$

Demostración. Definimos $U(t) = \sum_{n=1}^{N(t)} W_n$ para todo $t \geq 0$. Primero, observamos que como t_n es finito y bien definido para cualquier $n \in \mathbb{N}$, entonces $N(T) \to \infty$ cuando $T \to \infty$ y, por tanto, $\hat{W} = \lim_{T\to\infty} \sum_{n=1}^{N(T)} \frac{W_n}{N(T)}$. Esto implica que, cuando $\hat{\lambda} < \infty$ y cuando $\hat{W} < \infty$,

$$\begin{aligned} \hat{\lambda} \cdot \hat{W} &= \lim_{T\to\infty} \frac{N(T)}{T} \frac{1}{N(T)} \sum_{n=1}^{N(T)} W_n = \\ &= \lim_{T\to\infty} \frac{1}{T} \sum_{n=1}^{N(T)} W_n = \lim_{T\to\infty} \frac{U(T)}{T} \end{aligned}$$

También tenemos que $\lim_{T\to\infty} U(T)/T < \infty$, ya que $\hat{\lambda}$ y $\hat{W}$ son finitos. Entonces

$$\int_0^\infty e^{-\beta t} dU(t) < \infty$$

para cualquier $\beta > 0$. Por tanto, ya que $\int_0^\infty e^{-\beta t} dU(t) = \sum_{n=1}^\infty e^{-\beta t_n} W_n$ y a partir de un teorema abeliano que encontramos en la sección XIII.5 de [Feller, 1966], tenemos

$$\hat{\lambda} \cdot \hat{W} = \lim_{T\to\infty} \frac{U(T)}{T} = \lim_{\beta\to 0_+} \beta \sum_{n=1}^\infty e^{-\beta t_n} W_n$$

Además, por el corolario 4.11, tenemos que $\int_0^\infty e^{-\beta t} L(t) dt = \sum_{n=1}^\infty e^{-\beta t_n} \int_0^{W_n} e^{-\beta t} dt$ para todo $\beta > 0$, entonces como $\int_0^{W_n} e^{-\beta t} dt = \frac{1-e^{-\beta W_n}}{\beta} < W_n$, tenemos

$$\int_0^\infty e^{-\beta t} L(t) dt = \sum_{n=1}^\infty e^{-\beta t_n} \int_0^{W_n} e^{-\beta t} dt \leq \sum_{n=1}^\infty e^{-\beta t_n} W_n < \infty$$

Así pues, en caso de existir el límite $\lim_{\beta\to 0_+} \beta \int_0^\infty e^{-\beta t} L(t) dt$ y, también, a partir de un teorema tauberiano que encontramos en la sección XIII.5 de [Feller, 1966], tenemos

$$\begin{aligned} \hat{L} &= \lim_{T\to\infty} \frac{1}{T} \int_0^T L(t) dt = \lim_{\beta\to 0_+} \beta \int_0^\infty e^{-\beta t} L(t) dt \leq \\ &\leq \lim_{\beta\to 0_+} \beta \sum_{n=1}^\infty e^{-\beta t_n} W_n = \hat{\lambda} \cdot \hat{W} < \infty \end{aligned}$$

También a partir del corolario 4.11, para todo $\beta > 0$

$$\int_0^\infty e^{-\beta t} L(t) dt \geq \sum_{n=1}^\infty e^{-\beta (t_n + W_n)} W_n$$

En este punto, hemos probado que para demostrar el resultado del teorema es suficiente probar que

$$\lim_{\beta\to 0_+} \beta \sum_{n=1}^\infty e^{-\beta (t_n + W_n)} W_n = \hat{\lambda} \cdot \hat{W}$$

Fijamos un $\epsilon > 0$ arbitrario. Dado el lema 4.13, existe un $N \in \mathbb{N}$ tal que $W_n/t_n < \epsilon$ para todo $n > N$. Entonces, podemos escribir para cualquier $\beta > 0$

$$\begin{aligned} &\beta \sum_{n=1}^\infty e^{-\beta t_n} W_n \geq \beta \sum_{n=1}^\infty e^{-\beta (t_n + W_n)} W_n > \\ &> \beta \sum_{n=1}^N e^{-\beta (t_n + W_n)} W_n + \beta \sum_{n>N} e^{-\beta t_n (1+\epsilon)} W_n = \\ &= \beta \sum_{n=1}^N e^{-\beta (t_n + W_n)} W_n - \frac{1}{1+\epsilon} \hat{\beta} \sum_{n=1}^N e^{-\hat{\beta} t_n} W_n + \frac{1}{1+\epsilon} \hat{\beta} \sum_{n=1}^\infty e^{-\hat{\beta} t_n} W_n \end{aligned}$$

siendo $\hat{\beta} = (1+\epsilon)\beta$. Aquí observamos:

1. Dado que ϵ y N no dependen de β, el primer y el segundo término de la parte derecha de la desigualdad tienden a 0 cuando $\beta \to 0_+$.

2. El tercer término de la última parte de la desigualdad tiende a $\hat{\lambda} \cdot \hat{W}/(1+\epsilon)$.

3. La parte izquierda de la desigualdad tiende a $\hat{\lambda} \cdot \hat{W}$

De esta manera, tenemos

$$\hat{\lambda} \cdot \hat{W} \geq \lim_{\beta \to 0_+} \beta \sum_{n=1}^{\infty} e^{-\beta(t_n + W_n)} W_n > \frac{1}{1+\epsilon} \hat{\lambda} \cdot \hat{W}$$

Finalmente, para acabar la demostración, dado que ϵ es arbitrario, la anterior desigualdad implica

$$\lim_{\beta \to 0_+} \beta \sum_{n=1}^{\infty} e^{-\beta(t_n + W_n)} W_n = \hat{\lambda} \cdot \hat{W}$$

□

En este punto, hemos demostrado que los límites de la media de los procesos estocásticos vinculados a L, λ y W cumplen que para todo $\omega \in \Omega$, $\hat{L} = \hat{\lambda}\hat{W}$. De esta manera, tenemos que $L = \lambda W$ casi seguramente. Y así podemos dar por probada la primera fórmula de Little. Podemos encontrar más detalles acerca de las justificaciones probabilísticas para las hipótesis anteriores de los lemas en [Stidham, 1972].
De la misma manera que los resultados anteriores se han desarrollado a partir de $\hat{L}$, $\hat{\lambda}$ y $\hat{W}$, respecto a los clientes que se encuentran en el sistema (en la cola o siendo atendidos), se puede desarrollar respecto a los clientes que se encuentran en la cola para obtener el resultado $\hat{L}_q = \hat{\lambda}_q \hat{W}_q$ y probar la segunda fórmula de Little $L_q = \lambda W_q$.

4.4. Medidas de efectividad

Este apartado está fundamentado a partir de [Shortle u. a., 2018]. A partir de las distribuciones estacionarias encontradas en el apartado anterior, podemos encontrar medidas que nos permiten evaluar la efectividad de un sistema de colas. Nos preocupamos de encontrar el *esperado número de clientes dentro del sistema* y *el número esperado de clientes en la cola*, en ambos casos encontrándose el sistema en el estado estacionario. Siguiendo con las suposiciones realizadas hasta ahora: asumimos que estamos en modelos de colas donde las llegadas siguen un proceso de Poisson con parámetro λ y que los tiempos de servicio siguen una distribución exponencial con parámetro μ. Introducimos el símbolo ρ en la notación para designar la *intensidad de uso*, es decir, $\rho = \lambda/s\mu$ en el caso de sistemas con $s \geq 1$ servidores (posible ya que $\lambda, \mu > 0$). También continuamos con las notaciones de T_q, T, W_q y W del apartado anterior. Empezamos viendo dos resultados útiles para este apartado:

Proposición 4.16.

$$L - L_q = \rho$$

Demostración. Dado que $T = T_q + S$, tenemos que $\mathbb{E}[T] = \mathbb{E}[T_q] + \mathbb{E}[S]$ o, equivalentemente, $W = W_q + 1/\mu$ (dado que S es una variable aleatoria con distribución exponencial). De esta manera, es directa la demostración a partir de las fórmulas de Little:

$$L - L_q = \lambda(W - W_q) = \lambda(1/\mu) = \lambda/\mu = \rho$$

□

Vemos primero las medidas de efectividad en el modelo $M/M/1$.

Ejemplo 4.17 (Medidas de efectividad en el modelo M/M/1)**.** Recordemos que por el apartado anterior:

$$\pi(n) = (1-\rho)\rho^n$$

En este caso, $\rho = \lambda/\mu$ y trabajamos en la hipótesis que $\rho < 1$, pues como hemos visto en el ejemplo 4.1 es condición necesaria para que exista distribución estacionaria. Entonces,

$$L = \mathbb{E}[N] = \sum_{n=0}^{\infty} n\pi(n) = (1-\rho)\sum_{n=0}^{\infty} n\rho^n$$

Observando que $\sum_{n=0}^{\infty} n\rho^n = \rho\sum_{n=1}^{\infty} n\rho^{n-1}$ y que $\sum_{n=1}^{\infty} n\rho^{n-1}$ es la derivada de $\sum_{n=0}^{\infty}\rho^n$. Finalmente, dado que $\rho < 1$, $\sum_{n=0}^{\infty}\rho^n = \frac{1}{1-\rho}$. Entonces:

$$\sum_{n=1}^{\infty} n\rho^{n-1} = \frac{1}{(1-\rho)^2} \Rightarrow \sum_{n=0}^{\infty} n\rho^n = \frac{\rho}{(1-\rho)^2}$$

Así pues

$$\boxed{L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}}$$

es el **valor esperado de clientes en el sistema en el estado estacionario**.
Busquemos ahora el valor esperado de clientes en la cola en el estado estacionario (notar que anteriormente se trataba de los clientes en *el sistema*, es decir siendo atendidos o esperando en la cola). Por la proposición 4.16, $L_q = L - \rho$, entonces:

$$\boxed{L_q = L - \rho = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}}$$

Para acabar, buscamos el valor de W y W_q, el tiempo esperado (o media del tiempo) que un cliente pasará en el sistema y en la cola, respectivamente. A partir de la fórmula de Little en el teorema 4.6, directamente obtenemos que:

$$\boxed{W = \frac{L}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu-\lambda}}$$

y

$$\boxed{W_q = \frac{L_q}{\lambda} = \frac{\rho}{\mu(1-\rho)} = \frac{\rho}{\mu-\lambda}}$$

Observación 4.18. Añadimos conclusiones importantes para el ejemplo anterior. Primero: remarcamos la importancia de la condición $\rho < 1$. En caso de tener $\rho \geq 1$, las fórmulas anteriores no tendrían sentido. Segundo: observando las fórmulas se desprende que el valor de ρ, es decir, la relación entre λ y μ determina L, L_q, W y W_q. Así:

1. Para mantener el número de clientes dentro del sistema (y en la cola) bajo nos interesa tener un ρ lejano a 1, pues si $\rho \to 1 \Rightarrow L \to \infty$, que significa que las esperas son muy largas.

2. A la vez, al tratarse $\rho = \lambda/\mu$ de la intensidad de uso del sistema, sabemos que cuan más cercano a 1 más ocupados estarán los servidores y, por tanto, más alta será la productividad.

3. Se trata de encontrar un equilibrio entre ocupación del sistema y productividad.

En el siguiente ejemplo generalizamos el caso de $M/M/1$ en el caso de tener s servidores.

Ejemplo 4.19 (Medidas de efectividad en el modelo M/M/s)**.** Continuamos con las suposiciones anteriores: las llegadas de clientes siguen un proceso de Poisson con parámetro λ y los tiempos de servicio una distribución exponencial de parámetro μ. Suponemos que tenemos $s > 1$ servidores. En este caso, $\rho = \lambda/s\mu < 1$, ya que estamos en el estado estacionario. Recordemos que por el apartado anterior:

$$\pi(n) = \begin{cases} \frac{\lambda^n}{n!\mu^n}\pi(0) \text{ para } 0 \leq n < s \\ \frac{\lambda^n}{s^{n-s}s!\mu^n}\pi(0) \text{ para } n \geq s \end{cases}$$

con

$$\pi(0) = \left(\frac{(\lambda/\mu)^s}{s!} \cdot \frac{1}{1-\rho} + \sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} \right)^{-1} \tag{4.8}$$

Primero nos ocupamos de encontrar L_q. Dado que estamos en el caso de que hay cola en el sistema, suponemos que $n \geq s+1$. Así pues

$$L_q = \mathbb{E}[N_q] = \sum_{n=s+1}^{\infty} (n-s)\pi(n) = \sum_{n=s+1}^{\infty} (n-s)\frac{\lambda^n}{s^{n-s}s!\mu^n}\pi(0) =$$

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

Y usando que $\sum_{m=1}^{\infty} m\rho^{m-1} = \frac{d}{d\rho}\left(\sum_{m=1}^{\infty} \rho^m\right) = \frac{d}{d\rho}\left(\frac{1}{1-\rho} - 1\right) = \frac{1}{(1-\rho)^2}$ tenemos:

$$\boxed{L_q = \left(\frac{(\lambda/\mu)^s\rho}{s!(1-\rho)^2} \right)\pi(0)}$$

Para encontrar W_q usamos la fórmula de Little:

$$\boxed{W_q = \frac{L_q}{\lambda} = \left(\frac{(\lambda/\mu)^s}{s!(s\mu)(1-\rho)^2} \right)\pi(0)}$$

Para encontrar W, usamos que $W = W_q + 1/\mu$:

$$\boxed{W = \frac{1}{\mu} + \left(\frac{(\lambda/\mu)^s}{s!(s\mu)(1-\rho)^2} \right)\pi(0)}$$

Finalmente, para encontrar L usamos de nuevo la fórmula de Little $L = \lambda W$:

$$\boxed{L = \frac{\lambda}{\mu} + \left(\frac{(\lambda/\mu)^s\rho}{s!(1-\rho)^2} \right)\pi(0)}$$

4.5. El número de servidores

Desarrollamos este apartado a partir de [Shortle u. a., 2018]. En la gestión de colas es muy importante determinar el número c más apropiado de servidores para nuestro sistema. Tenemos que encontrar el equilibrio entre: un c muy alto que mejora la calidad del servicio hacia los clientes pero supone un alto coste para el propietario del sistema y un c muy bajo que dificulta la prestación del servicio pero que supone un ahorro.
Estudiamos el número de servidores más adecuado para el modelo $M/M/s$, así pues con las suposición que las llegadas de clientes siguen un proceso de Poisson de parámetro λ y que los tiempos de servicio siguen una distribución exponencial de parámetro μ. Recordamos la notación $\rho = \lambda/s\mu$. Para este apartado introducimos una nueva notación: usamos r para $r = \lambda/\mu$ (en el caso particular de $M/M/1$, $r = \rho$). r es llamada *carga ofrecida*, dado que de media cada cliente requiere $1/\mu$ unidades de tiempo y de media en cada unidad de tiempo llegan λ clientes, pues el producto $\lambda(1/\mu)$ es la cantidad de trabajo necesario para satisfacer las llegadas por unidad de tiempo.
Suponiendo que estamos en el estado estacionario, tenemos que para que la cola sea estable

$$c = r + \triangle$$

siendo $\triangle > 0$ el número de servidores adicionales añadidos a r para llegar a c. Así para hallar c, debemos hallar $\triangle$.
Debemos empezar este apartado introduciendo una fórmula que nos será útil. A partir de las notaciones del apartado anterior, consideremos $W_q(0)$ como la probabilidad de que un cliente tenga 0 espera antes de recibir el servicio. Por ende, $1 - W_q(0)$ es la probabilidad que un cliente tenga que esperar en la cola antes de ser atendido.

Proposición 4.20 (Fórmula C-Erlang)**.** *La probabilidad que un cliente tenga que esperar un tiempo mayor que cero en la cola es determinado por la fórmula:*

$$C(c,r) = 1 - W_q(0) = \frac{\frac{r^c}{c!(1-\rho)}}{\left(\frac{r^c}{c!(1-\rho)} + \sum_{n=0}^{c-1} \frac{r^n}{n!}\right)}$$

Demostración. Rescatando las notaciones del apartado anterior, definimos T_q como la variable aleatoria del tiempo que el cliente pasa en la cola y W_q como $W_q = \mathbb{E}[T_q]$, es decir, el tiempo esperado que un cliente pasará en la cola. Primero, buscamos $W_q(0)$:

$$W_q(0) = \mathbb{P}\{T_q = 0\} = \mathbb{P}\{\text{número de clientes en el sistema} \leq c-1\} =$$
$$= \sum_{n=0}^{c-1} \pi(n) = \pi(0) \sum_{n=0}^{c-1} \frac{r^n}{n!}$$

En la fórmula 4.8 aparece también el término $\sum_{n=0}^{c-1} \frac{r^n}{n!}$, de esta manera:

$$\sum_{n=0}^{c-1} \frac{r^n}{n!} = \frac{1}{\pi(0)} - \frac{r^c}{c!(1-\rho)}$$

Por lo tanto, podríamos escribir:

$$W_q(0) = \pi(0)\left(\frac{1}{\pi(0)} - \frac{r^c}{c!(1-\rho)}\right) = 1 - \frac{r^c \pi(0)}{c!(1-\rho)}$$

Y así, tomando 4.8, tenemos:

$$1 - W_q(0) = \frac{\frac{r^c}{c!(1-\rho)}}{\left(\frac{r^c}{c!(1-\rho)} + \sum_{n=0}^{c-1} \frac{r^n}{n!}\right)}$$

□

Volviendo al problema que quiere tratar este apartado, distinguimos tres enfoques para encontrar el número de servidores *adecuado*:

1. **Enfoque de calidad**. El objetivo es proveer un servicio de alta calidad aunque suponga un elevado coste. Para esto, debemos fijar el nivel de congestión ρ deseado y mantenerlo constante respecto las variaciones de r. Puesto que $\rho = r/s$, mantener ρ constante es que el número de servidores y la carga ofrecida r sean proporcionales. Entonces aunque la carga ofrecida aumente la probabilidad de retraso disminuye (si $r \to \infty$ entonces $1 - W_q(0) \to 0$).

   Proposición 4.21. *En el enfoque de calidad, la fórmula para c es:*

   $$\boxed{c = \frac{r}{\rho}}$$

   Demostración. En efecto, si $c = r/\rho$, $\rho = r/c = \lambda/c\mu$ es constante respecto a r. □

2. **Enfoque de eficiencia**. El objetivo es reducir los costes aunque suponga una reducción de la calidad del servicio. Para esto, debemos fijar un $\triangle$ deseado y mantenerlo constante respecto las variaciones de r. De esta manera, si $r \to \infty$, entonces $\rho \to 1$ y $1 - W_q(0) \to 1$.

   Proposición 4.22. *En el enfoque de eficiencia, fijando $\triangle$, la fórmula para c es:*

   $$\boxed{c = r + \triangle}$$

   Demostración. Puesto que fijamos $\triangle$, ante las variaciones de r, solo nos queda aplicar la fórmula (4.5) para conocer c. □

3. **Enfoque de calidad y eficiencia**, con las siglas *QED* en inglés. Busca el equilibrio entre los enfoques de calidad y de eficiencia. El objetivo es mantener constante la calidad del sistema, es decir, mantener $\alpha := 1 - W_q(0)$ constante respecto las variaciones de r.

   Teorema 4.23. *Consideremos una sucesión de sistemas de colas $M/M/c$ indexado a partir del parámetro $n = 1, 2, 3, ...$ Suponemos que la cola n tiene $c_n = n$ servidores y una carga ofrecida r_n. Entonces*

   $$lim_{n\to\infty} C(c_n = n, r_n) = \alpha \tag{4.9}$$

   para $0 < \alpha < 1$, si, y solo si,

   $$\lim_{n\to\infty} \frac{n - r_n}{\sqrt{n}} = \beta \tag{4.10}$$

para $\beta > 0$. Donde $C(c, r)$ es la C-Fórmula de Erlang introducida en la proposición 4.20 y donde α y β son constantes relacionadas entre sí:

$$\alpha = \frac{\phi(\beta)}{\phi(\beta) + \beta\Phi(\beta)} \tag{4.11}$$

Donde ϕ y Φ son la función de densidad y la función de distribución de la distribución normal estándar.

Demostración. La demostración de este teorema se puede encontrar en la sección segunda de [Shlomo und Whitt, 1981]. □

Corolario 4.24. *De (4.9) se desprende que $C(c, r) = 1 - W_q(0)$ es aproximadamente constante durante la sucesión de sistemas de colas.*

Corolario 4.25 (Regla de la raíz cuadrada)**.** *En el enfoque de calidad y eficiencia, fijando un nivel de calidad $\alpha = 1 - W_q(0)$ y encontrando β a partir del teorema 4.23, la fórmula para c es:*

$$\boxed{c \approx r + \beta\sqrt{r} \text{ o } \triangle \approx \beta\sqrt{r}} \tag{4.12}$$

Demostración. Dada la fórmula (4.10), tenemos que $n - r_n \approx \beta\sqrt{n}$ o $n \approx r_n + \beta\sqrt{n}$. Siguiendo las suposiciones del teorema, si reemplazamos $\sqrt{n}$ por $\sqrt{r_n}$, entonces obtenemos la regla de la raíz cuadrada. □

Observación 4.26. Los parámetros α y β puede ser interpretados de la siguiente manera: $\alpha = 1 - W_q(0)$ es la probabilidad de que haya una espera superior a cero minutos en la cola; y β es la constante relacionada con α de la manera descrita en el teorema.

Observación 4.27. La Regla de la raíz cuadrada puede ser usada sin necesidad de especificar el valor de las constantes α y β. Se puede usar para interpretar la eficiencia y calidad del sistema en caso de variar la carga ofrecida r. Por ejemplo, si un número de servidores c satisface la actual carga ofrecida, si esta última se dobla entonces el número de servidores debe ser aumentado con un factor de $\sqrt{2}$.
A la vez, las constantes puedes ser utilizadas para aproximar el número de servidores más adecuado para una calidad de servicio α: primero hay que encontrar la β que satisface la ecuación (4.11) del teorema y, entonces, usar la fórmula (4.12) para hallar c.

Proposición 4.28. *Fijado un nivel de calidad de servicio α, una buena aproximación a la β definida en el teorema anterior es que β sea el $(1 - \alpha)$ cuantil de la distribución normal estándar.*

Demostración. En la referencia [Kolesar und Green, 1998] encontramos una extensa demostración de esta proposición. □

5. Redes de colas

A partir de [Durret, 1999], desarrollamos esta sección sobre sistemas con más de una cola. Continuamos con la notación usada en la sección anterior y en el supuesto que las llegadas de clientes siguen un proceso de Poisson de parámetro λ y que los tiempos de servicio siguen una distribución exponencial de parámetro μ.

5.1. Reversibilidad

Antes de profundizar en las redes de colas, necesitamos presentar algunos resultados previos. El principal es el primero, el cual demostraremos al final de este apartado después de demostrar un lema vinculado a él.

Teorema 5.1. *Si $\lambda < \mu s$, la salida de clientes en un sistema de colas $M/M/s$ en estado estacionario sigue un proceso de Poisson con parámetro λ. Este resultado también se cumple en el caso de tener μ dependiente del número n de clientes presentes en el sistema, $\mu = \mu(n)$.*

Lema 5.2. *Sea $\{X_t\}_t$ una cadena de Markov a tiempo continuo con intensidades de transición $q(i,j)$ y probabilidades de transición $p_t(i,j)$ con la distribución estacionaria como distribución inicial, es decir, $\mathbb{P}(X_0 = i) = \pi(i)$. Si fijamos un tiempo t y definimos $Y_s = X_{t-s}$ para $0 \leq s \leq t$. Entonces, recorrer Y_s es recorrer X_t pero en dirección contraria y tiene probabilidad de transición:*

$$\hat{p}_t(i,j) = \frac{\pi(j)p_t(j,i)}{\pi(i)}$$

Además, si la distribución estacionaria π satisface la condición de equilibrio $\pi(i)q(i,j) = \pi(j)q(j,i)$ para todo i,j (ecuación (3.14)), entonces Y_s tiene probabilidad de transición $p_t(i,j)$.

Demostración. Demostraremos este lema en su versión en tiempo discreto. La demostración en tiempo continuo está basada en las mismas ideas.
Probamos el siguiente resultado: Sea $\{X_n\}_n$ una cadena de Markov a tiempo discreto. Sea $\pi(i)$ su distribución estacionaria y sean $p(i,j)$ sus probabilidades de transición con la distribución estacionaria como distribución inicial, es decir, $\mathbb{P}(X_0 = i) = \pi(i)$. Entonces, si fijamos un estado n y definimos $Y_m = X_{n-m}$ para $0 \leq m \leq n$, Y_m es una cadena de Markov en tiempo discreto con probabilidad de transición

$$\hat{p}(i,j) = \mathbb{P}(Y_{m+1} = j | Y_m = i) = \frac{\pi(j)p(j,i)}{\pi(i)}$$

Empezamos calculando la probabilidad condicionada siguiente:

$$\begin{aligned}
&\mathbb{P}(Y_{m+1} = i_{m+1} | Y_m = i_m, Y_{m-1} = i_{m-1}, ..., Y_0 = i_0) = \\
&= \frac{\mathbb{P}(X_{n-(m+1)} = i_{m+1}, X_{n-m} = i_m, X_{n-m+1} = i_{m-1}, ..., X_n = i_0)}{\mathbb{P}(X_{n-m} = i_m, X_{n-m+1} = i_{m-1}, ..., X_n = i_0)} = \\
&= \frac{\pi(i_{m+1})p(i_{m+1}, i_m)\mathbb{P}(X_{n-m+1} = i_{m-1}, ..., X_n = i_0 | X_{n-m} = i_m)}{\pi(i_m)\mathbb{P}(X_{n-m+1} = i_{m-1}, ..., X_n = i_0 | X_{n-m} = i_m)} = \\
&= \frac{\pi(i_{m+1})p(i_{m+1}, i_m)}{\pi(i_m)}
\end{aligned}$$

Así hemos visto que Y_m es una cadena de Markov en tiempo discreto con la probabilidad de transición indicada.
Vemos que la fórmula $\hat{p}(i,j) = \frac{\pi(j)p(j,i)}{\pi(i)}$ tiene sentido: puesto que $\pi p = \pi$, tenemos

$$\sum_j \hat{p}(i,j) = \sum_j \frac{\pi(j)p(j,i)}{\pi(i)} = \frac{\pi(i)}{\pi(i)} = 1$$

Por último, comprobamos que si π satisface la condición de equilibrio, entonces Y_s tiene probabilidad de transición $p_t(i,j)$. Cuando π satisface la condición de equilibrio, cumple

$$\pi(i)p(i,j) = \pi(j)p(j,i)$$

para todo i, j. De esta manera,

$$\hat{p}(i,j) = \frac{\pi(j)p(j,i)}{\pi(i)} = p(i,j)$$

□

Tenemos que si nos desplazamos en dirección contraria por el recorrido de la cadena de Markov X_t de una cola $M/M/s$ y tomamos como distribución inicial la distribución estacionaria, entonces tenemos un proceso aleatorio con la misma distribución que X_t.

Demostración. Demostramos el teorema 5.1. Tal y como hemos visto en la sección anterior, en un sistema de cola $M/M/s$ si $\lambda < \mu s$, entonces la cola es un proceso de nacimiento y muerte con una distribución estacionaria π que satisface la condición de equilibrio (3.14). Si tomamos la cadena de Markov vinculada a esta cola $M/M/s$ con distribución inicial la estacionaria, a partir del lema anterior, si damos la vuelta a los tiempos (si vamos hacia atrás), las llegadas se convierten en salidas. Por tanto, las salidas siguen un proceso de Poisson de parámetro λ. □

Teorema 5.3. *Consideremos $N(t)$ como el número de salidas en una cola $M/M/1$ y consideremos $X(t)$ como la longitud de esta cola, desde $t = 0$ hasta $t = n$. Teniendo la distribución estacionaria como distribución inicial. Entonces, $\{N(s)|0 \leq s \leq t\}$ y $X(t)$ son independientes.*

Demostración. Tomando la cadena de Markov en tiempo continuo $\{X_t\}$, si damos la vuelta al tiempo, las salidas antes del momento t se convierten en llegadas después del momento t. Entonces, tanto las llegadas como las salidas son independientes de la longitud de la cola en el momento t, $X(t)$. □

5.2. Redes de colas

En este apartado, estudiamos dos ejemplos de redes de colas. Una red de colas es un sistema de colas que contiene más de una cola, por lo que en el sistema hay diferentes tiempos de servicio μ y el tiempo de llegada de los clientes a cada cola depende del orden de éstas. Un ejemplo canónico de red de colas sería el del servicio de urgencias de un hospital, donde una vez superada la cola del triaje, el cliente tiene que esperar en una segunda cola para ser atendido por el especialista pertinente. Hay muchos ejemplos de redes de colas muy complejas, pero en este apartado nos centramos en las redes de colas en tándem.

Ejemplo 5.4 (Red de dos servidores en tándem)**.** Como hemos supuesto en el inicio de la sección, los clientes llegan al servidor 1 siguiendo un proceso de Poisson de parámetro λ, donde son atendidos con un tiempo de servicio que sigue una distribución exponencial de parámetro μ_1. Una vez el servicio en este servidor 1 es completado, se unen a la cola para acceder al servidor 2 y son atendidos en éste con un tiempo de servicio que sigue

una distribución exponencial de parámetro μ_2.
Como hemos visto en la sección 4.2. de este texto, la primera cola, al no verse afectada por la segunda, para tener distribución estacionaria debe cumplir que $\lambda < \mu_1$. Entonces, la distribución del número de clientes N_t^1 en la primera cola, viene dado por:

$$\mathbb{P}(N_t^1 = m) = \left(\frac{\lambda}{\mu_1}\right)^m \left(1 - \frac{\lambda}{\mu_1}\right)$$

Ahora, a partir del teorema 5.1, sabemos que en estado estacionario la salida de los clientes del servidor 1 sigue un proceso de Poisson de parámetro λ. Entonces, el número de clientes N_t^2 de la segunda cola es una cola $M/M/1$ con parámetro de llegada λ y parámetro de servicio μ_2. De nuevo, a partir de la sección 4.2.:

$$\mathbb{P}(N_t^2 = n) = \left(\frac{\lambda}{\mu_2}\right)^n \left(1 - \frac{\lambda}{\mu_2}\right)$$

Para acabar de especificar la distribución estacionaria de esta red de colas, necesitamos dar la distribución de N_t^1 y N_t^2 conjuntamente. Dado el teorema 5.3, el número de salidas hasta el momento t es independiente de N_t^1. Entonces, dado que N_t^2 está determinado por el proceso de salida de los clientes de la cola 1, éste es independiente de N_t^1, resultando que N_t^2 y N_t^1 lo son para $t \geq 0$. Así pues

$$\mathbb{P}(N_t^1 = m, N_t^2 = n) = \left(\frac{\lambda}{\mu_1}\right)^m \left(1 - \frac{\lambda}{\mu_1}\right) \left(\frac{\lambda}{\mu_2}\right)^n \left(1 - \frac{\lambda}{\mu_2}\right)$$

Por lo que podemos escribir

$$\pi(m, n) = c\frac{\lambda^{m+n}}{(\mu_1^m \mu_2^n)}$$

con $c = (1 - \lambda/\mu_1)(1 - \lambda/\mu_2)$.

Ejemplo 5.5 (Red de s servidores en tándem)**.** Siguiendo el ejemplo anterior, en caso de tener s servidores los cuales el cliente ha de recorrer en un cierto orden, con λ como parámetro del proceso de Poisson de la llegada de clientes al servidor 1 y $\mu_1, \mu_2, ..., \mu_s$ como los parámetros de las distribuciones exponenciales que siguen los tiempos de servicio, tenemos

$$\pi(n_1, n_2, ..., n_s) = c\frac{\lambda^{n_1+n_2+\cdots+n_s}}{\mu_1^{n_1}\mu_2^{n_2}\cdots\mu_s^{n_s}}$$

con $c = \prod_{k=1}^{s}(1 - \lambda/\mu_k)$.

6. Ejemplo con aplicación práctica

6.1. Descripción del caso

Desde julio del 2019 trabajo en una empresa llamada Giesecke+Devrient Mobile Security Iberica, que es parte de la multinacional alemana Giesecke+Devrient, con más de 11.300 empleados en 32 países[2]. Formo parte del departamento de R&D desempeñando el rol de Test Mánager, por el que soy uno de los encargados de la calidad del producto que desarrollamos: sistemas operativos para tarjetas inteligentes, o *smartcards*, destinadas a teléfonos inteligentes, *smartwatches*, tarjetas bancarias, tarjetas sanitarias, tarjetas de

[2]Más información en: https://www.gi-de.com/es/es/mobile-security/?informaci %C3 %B3n %C2 %BB=

transporte, coches con conectividad, etc. La seguridad de estos productos es esencial y por ello el proceso de certificación de calidad es muy importante.
Esta certificación de calidad se realiza de dos maneras distintas:

1. *Integración continua*. Cada vez que en el departamento se realiza un cambio en el producto, un sistema automático realiza ejecuciones de test para comprobar que ese cambio no ha implicado ninguna regresión. Este sistema requiere de un complejo soporte de *software* y *hardware* para funcionar: aproximadamente 60 servidores están dedicados exclusivamente a la integración continua y se usa la plataforma *Jenkins* para la gestión de las ejecuciones de test y de sus ordenadores.

2. *Test manuales*. Son test ejecutados de manera programada y discrecional.

En el caso de la **integración continua**, dado el alto volumen de ejecuciones de test que se acumulan y dado el número limitado de ordenadores dedicados, se forma una **cola**. Esta cola será objeto de estudio en esta sección. Siguiendo la nomenclatura usada durante todo el trabajo, llamaremos *clientes* a las tareas de test o ejecuciones que entran en el sistema y llamaremos *servidores* a los ordenadores que procesan estas ejecuciones. Para preservar la política de confidencialidad de Giesecke+Devrient, en ningún momento se mencionarán nombres propios de ningún tipo, usando en todo momento nombres genéricos y etiquetas numéricas.

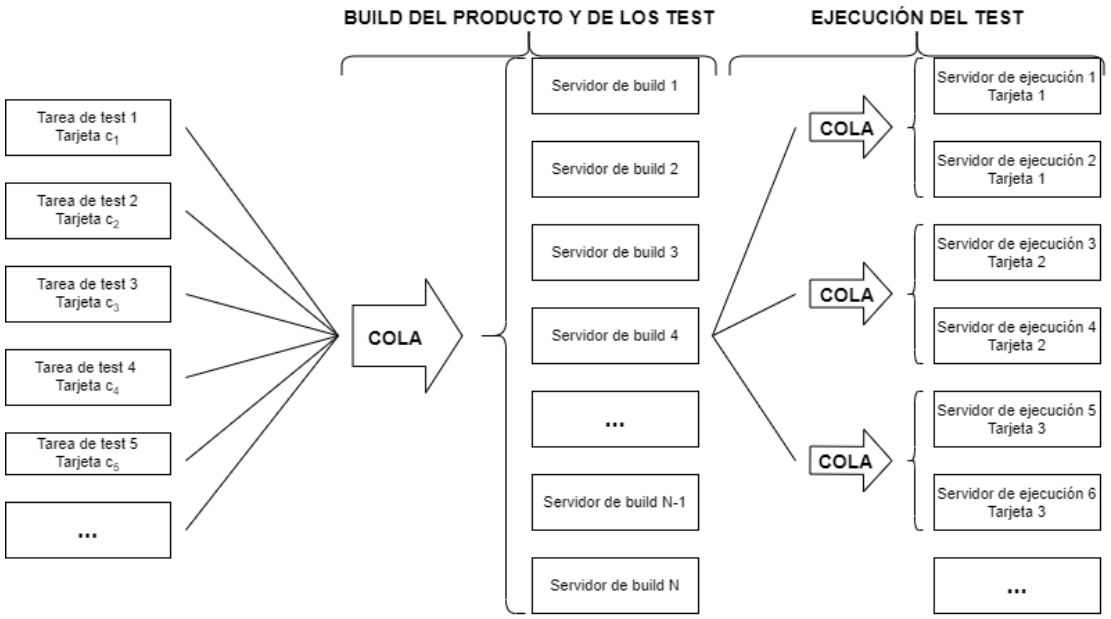

Figura 3: Diagrama del sistema de colas de integración continua

Cada tarea de test tiene asignada una etiqueta que indica en qué tarjeta se ejecutará. Hay 3 tarjetas diferentes en el sistema. Estas tareas de test, antes de su ejecución deben pasar por una fase previa donde se lleva a cabo el **build** o compilación del sistema operativo y de los test que se ejecutarán. El build se lleva a cabo en ordenadores del sistema, los *servidores de build*, y tiene una cola común para todas las ejecuciones. En la siguiente fase, donde hay la **ejecución** del test, cada servidor de ejecución tiene una tarjeta concreta, siendo posible que más de un servidor tenga el mismo tipo de tarjeta. Cada ejecución se incorpora a la cola perteneciente a los servidores que coinciden con su

tarjeta. Al haber 3 tarjetas, hay un total de 3 colas de ejecución de test. De esta manera, el build actúa como cuello de botella al impedir que ejecuciones de test con tarjetas con baja ocupación tengan que completar la cola de los build, antes de acceder a la cola de ejecución. Múltiples tareas de test suelen ser lanzadas en el mismo tiempo en forma de paquetes, por lo que varias decenas de tareas se unen de golpe a la cola.
El objetivo de esta sección es, a partir de la descripción proporcionada, hacer un planteamiento teórico que sirva para construir un modelo suficientemente representativo de la cola y, con ello, calcular las medidas de efectividad y el número de servidores más adecuado, mediante la recopilación de datos que nos proporciona el propio sistema de integración continua.

6.2. Planteamiento teórico

A partir de la teoría desarrollada en este trabajo, queremos modelizar el sistema de colas descrito. Para ello queremos modelizarlo como una red de colas $M/M/s$. Esta red sería de la siguiente manera:

1. Una primera cola $M/M/s$ perteneciente a la cola que se forma para el *build del producto y de los test*. Para ello necesitamos encontrar el parámetro λ_0 del proceso de Poisson que sigue la llegada de clientes (o de tareas de test) y encontrar el parámetro μ_0 de la distribución exponencial que sigue el tiempo de los servicios. En este punto, debemos ver que $\lambda_0 < \mu_0 s$ para probar que esta cola es estable y tiene distribución estacionaria, siendo s el número de servidores de build.

2. Siguiendo el teorema 5.1 y el apartado 5.2, tenemos que la salida de clientes de esta primera cola se produce siguiendo un proceso de Poisson de parámetro λ_0, dado que trabajamos en el caso $\lambda_0 < \mu_0 s$. Como cada servidor de ejecución tiene un solo tipo de tarjeta y hay 3 tarjetas distintas en el sistema, se forman 3 colas distintas en la fase de ejecución de test y cada una de estas colas tiene s_i servidores de ejecución con $i = 1, 2, 3$. De esta manera, se forman 3 sistemas de colas $M/M/s_i$ con λ_i (dependiente de λ) como parámetro del proceso de Poisson que sigue la llegada de clientes a cada cola y con μ_1, como parámetro de la distribución exponencial que siguen los tiempos de servicio (el tiempo de servicio es independiente de la tarjeta donde se ejecute el test). Sabiendo con qué probabilidad p_i una tarea de test tiene asignada la tarjeta i (siendo $p_1 + p_2 + p_3 = 1$), entonces $\lambda_i = \lambda p_i$. De esta manera, faltaría por conocer μ_1.

Por tanto, si conocemos los parámetros λ_0, μ_0, λ_i y μ_1 mencionados habremos modelizado la red de colas y con ellos podremos calcular las medidas de efectividad y el número de servidores más adecuado.
Usamos el método del estimador máximo verosímil para el cálculo de estos parámetros [Corcuera, 2019b].

Proposición 6.1. *Sea un modelo de n observaciones x iid exponenciales con parámetro $\mu > 0$. El estimador máximo verosímil $\bar{\mu}$ de μ es*

$$\bar{\mu}(x) = \frac{n}{\sum_{i=1}^{n} x_i}$$

Sea un modelo de n observaciones x iid Poisson con parámetro $\lambda > 0$. El estimado máximo verosímil $\bar{\lambda}$ de λ es

$$\bar{\lambda}(x) = \frac{\sum_{i=1}^{n} x_i}{n}$$

En caso de ser estimadores sesgados, tomaremos el estimador máximo verosímil corregido.

Proposición 6.2. *El estimador máximo verosímil $\bar{\mu}$ es sesgado. El estimador corregido es*

$$\bar{\mu}^*(x) = \frac{n-1}{\sum_{i=1}^{n} x_i}$$

El estimador $\bar{\lambda}$ ya es insesgado.

Para el caso del parámetro μ_0 y μ_1, el estimador máximo verosímil corregido lo podemos hallar sumando los tiempos de build o de ejecución de una muestra de n tiempos suficientemente grande y dividiéndola entre $n-1$. En ambos casos, los estimadores buscados son el inverso de este cálculo.
En cambio, para el cálculo del parámetro λ_0 del proceso de Poisson, sabemos por el lema 2.10 que sigue una distribución de Poisson de parámetro $\lambda_0 t$. Entonces el EMV será el número de llegadas totales durante un periodo suficientemente largo dividido entre el número de minutos transcurridos.

6.3. Cálculos

Antes de todo, constatamos que hay 13 servidores de build y 48 servidores de ejecución en el sistema, de los cuales 17 son dedicados a la tarjeta 1, 18 a la tarjeta 2 y 13 a la tarjeta 3. En esta sección, las medidas de tiempo estarán en minutos y los parámetros seguirán las ratios *llegadas/minuto* y *servicios/minuto*.

1. **Cálculo de λ_0.** A partir de los datos de 3612 tareas de test ejecutadas, tenemos que la media de ejecuciones es

   179,7977 al día
   0,1249 al minuto

   Por tanto

   $$\bar{\lambda_0} = 0{,}1249$$

2. **Cálculo de μ_0.** A partir de los datos de 2268 builds realizados, tenemos que la suma de los tiempos que necesita un servidor de ejecución para ejecutar una tarea dividido entre 2267 es

   93,6206 minutos

   Por tanto

   $$\bar{\mu_0}^* = \frac{1}{93{,}6206} = 0{,}0107$$

3. **Cálculo de λ_i.** A partir del análisis de la cantidad de tareas que se lanzan de cada tarjeta tenemos la siguiente información:

   a) El 41 % de las tareas son dirigidas a la tarjeta 1
   b) El 26 % de las tareas son dirigidas a la tarjeta 2
   c) El 33 % de las tareas son dirigidas a la tarjeta 3

De esta manera, dado que $\bar{\lambda}_0 = 0{,}1249$ es el promedio total de salidas al minuto y por lo visto en el apartado anterior, tenemos

$$\bar{\lambda_1} = \bar{\lambda_0} \cdot 0{,}41 = 0{,}0512$$
$$\bar{\lambda_2} = \bar{\lambda_0} \cdot 0{,}26 = 0{,}0325$$
$$\bar{\lambda_3} = \bar{\lambda_0} \cdot 0{,}33 = 0{,}0412$$

4. **Cálculo de** μ_1. A partir de los datos de 2268 ejecuciones realizadas, tenemos que la suma de los tiempos que necesita un servidor de ejecución para ejecutar una tarea dividido entre 2267 es

   $$303{,}1497 \text{ minutos}$$

   Por tanto

   $$\bar{\mu_1}^* = \frac{1}{303{,}1497} = 0{,}00329$$

5. **Medidas de efectividad** Aplicamos las fórmulas contenidas en el ejemplo 4.19.

   a) Respecto la **cola de los build**. En esta fase tenemos $s = 13$ servidores. Empezamos calculando ρ_0:

   $$\rho_0 = \frac{\lambda_0}{s\mu_0} = \frac{0{,}1249}{13 \cdot 0{,}0107} = \frac{0{,}1249}{0{,}1391} = 0{,}8979$$

   Dado que $\rho_0 < 1$, la cola es estable y tiene distribución estacionaria. Por lo que podemos calcular las medidas de efectividad.
   Seguimos con $\pi(0)$:

   $$\pi(0) = \left(\frac{(\lambda_0/\mu_0)^s}{s!} \cdot \frac{1}{1-\rho_0} + \sum_{n=0}^{s-1} \frac{\lambda_0^n}{n!\mu_0^n} \right)^{-1} =$$
   $$= \left(\frac{(0{,}1249/0{,}0107)^{13}}{13!} \cdot \frac{1}{1-0{,}8979} + \sum_{n=0}^{12} \frac{0{,}1249^n}{n!0{,}0107^n} \right)^{-1} =$$
   $$= (11996{,}1606 \cdot 9{,}7943 + 71971{,}6200)^{-1} = \frac{1}{189465{,}6158} = 0{,}000005278$$

   De aquí, calculamos L_q:

   $$L_q = \left(\frac{(\lambda_0/\mu_0)^s \rho_0}{s!(1-\rho_0)^2} \right) \pi(0) = \left(\frac{(0{,}1249/0{,}0107)^{13} \cdot 0{,}8979}{13!(1-0{,}8979)^2} \right) 0{,}000005278 = 5{,}4536$$

   Así W_q, por la fórmula de Litlle:

   $$W_q = \frac{L_q}{\lambda_0} = \frac{5{,}4536}{0{,}1249} = 43{,}6642$$

   Como $W = W_q + 1/\mu_0$,

   $$W = 43{,}6642 + \frac{1}{0{,}0107} = 137{,}1221$$

   Finalmente, por la fórmula de Little,

   $$L = \lambda_0 \cdot W = 0{,}1249 \cdot 137{,}1221 = 17{,}1266$$

   Con todo tenemos

   $$\boxed{L_q = 5{,}4536; L = 17{,}1266; W_q = 43{,}6642; W = 137{,}1221}$$

b) Respecto la **cola de la tarjeta 1**. En la fase de ejecución con esta tarjeta tenemos un total de $s = 17$ servidores. Empezamos calculando ρ_1:

$$\rho_1 = \frac{\lambda_1}{s\mu_1} = 0{,}9154$$

Dado que $\rho_1 < 1$, la cola es estable y tiene distribución estacionaria. Por lo que podemos calcular las medidas de efectividad.
Seguimos con $\pi(0)$:

$$\pi(0) = \left(\frac{(\lambda_1/\mu_1)^s}{s!} \cdot \frac{1}{1-\rho_1} + \sum_{n=0}^{s-1} \frac{\lambda_1^n}{n!\mu_1^n} \right)^{-1} =$$

$$= \left(\frac{(0{,}0512/0{,}00329)^{17}}{17!} \cdot \frac{1}{1-0{,}9154} + \sum_{n=0}^{16} \frac{0{,}0512^n}{n!0{,}00329^n} \right)^{-1} =$$

$$= (517833{,}9932 \cdot 11{,}8203 + 3495082{,}501)^{-1} = \frac{1}{9616035{,}651} = 0{,}000000103$$

De aquí, calculamos L_q:

$$L_q = \left(\frac{(\lambda_1/\mu_1)^s \rho_1}{s!(1-\rho_1)^2} \right) \pi(0) = \left(\frac{(0{,}0512/0{,}00329)^{17} 0{,}9154}{17!(1-0{,}9154)^2} \right) \cdot 0{,}000000103 =$$

$$= 66230912{,}45 \cdot 0{,}000000103 = 6{,}8875$$

Así W_q, por la fórmula de Litlle:

$$W_q = \frac{L_q}{\lambda_1} = \frac{6{,}8875}{0{,}0512} = 134{,}5215$$

Como $W = W_q + 1/\mu_1$,

$$W = 134{,}5215 + \frac{1}{0{,}00329} = 438{,}4728$$

Finalmente, por la fórmula de Little,

$$L = \lambda_1 \cdot W = 0{,}0512 \cdot 438{,}4728 = 22{,}4498$$

Con todo tenemos

$$\boxed{L_q = 6{,}8875; L = 22{,}4498; W_q = 134{,}5215; W = 438{,}4728}$$

c) Respecto la **cola de la tarjeta 2**. En la fase de ejecución con esta tarjeta tenemos un total de $s = 18$ servidores. Empezamos calculando ρ_2:

$$\rho_2 = \frac{\lambda_2}{s\mu_1} = 0{,}5488$$

Dado que $\rho_2 < 1$, la cola es estable y tiene distribución estacionaria. Por lo que podemos calcular las medidas de efectividad.
Seguimos con $\pi(0)$:

$$\pi(0) = \left(\frac{(\lambda_2/\mu_1)^s}{s!} \cdot \frac{1}{1-\rho_2} + \sum_{n=0}^{s-1} \frac{\lambda_2^n}{n!\mu_1^n} \right)^{-1} =$$

$$= \left(\frac{(0{,}0325/0{,}00329)^{18}}{18!} \cdot \frac{1}{1-0{,}5488} + \sum_{n=0}^{17} \frac{0{,}0325^n}{n!0{,}00329^n} \right)^{-1} =$$

$$= (125{,}3237 \cdot 2{,}2163 + 19255{,}3914)^{-1} = \frac{1}{19533{,}1463} = 0{,}0000512$$

De aquí, calculamos L_q:

$$L_q = \left(\frac{(\lambda_2/\mu_1)^s \rho_2}{s!(1-\rho_2)^2}\right)\pi(0) = \left(\frac{(0{,}0325/0{,}00329)^{18} 0{,}5488}{18!(1-0{,}5488)^2}\right) \cdot 0{,}0000512 =$$
$$= 337{,}8386 \cdot 0{,}0000512 = 0{,}0173$$

Así W_q, por la fórmula de Litlle:

$$W_q = \frac{L_q}{\lambda_2} = \frac{0{,}0173}{0{,}0325} = 0{,}5322$$

Como $W = W_q + 1/\mu_1$,

$$W = 0{,}5322 + \frac{1}{0{,}00329} = 304{,}4836$$

Finalmente, por la fórmula de Little,

$$L = \lambda_2 \cdot W = 0{,}0325 \cdot 304{,}4836 = 9{,}8957$$

Con todo tenemos

$$\boxed{L_q = 0{,}0173; L = 9{,}8957; W_q = 0{,}5322; W = 304{,}4836}$$

d) Respecto la **cola de la tarjeta 3**. En la fase de ejecución con esta tarjeta tenemos un total de $s = 13$ servidores. Empezamos calculando ρ_3:

$$\rho_3 = \frac{\lambda_3}{s\mu_1} = 0{,}9633$$

Dado que $\rho_3 < 1$, la cola es estable y tiene distribución estacionaria. Por lo que podemos calcular las medidas de efectividad.
Seguimos con $\pi(0)$:

$$\pi(0) = \left(\frac{(\lambda_3/\mu_1)^s}{s!} \cdot \frac{1}{1-\rho_3} + \sum_{n=0}^{s-1} \frac{\lambda_3^n}{n!\mu_1^n}\right)^{-1} =$$
$$= \left(\frac{(0{,}0412/0{,}00329)^{13}}{13!} \cdot \frac{1}{1-0{,}9633} + \sum_{n=0}^{12} \frac{0{,}0412^n}{n!0{,}00329^n}\right)^{-1} =$$
$$= (29911{,}4054 \cdot 27{,}2476 + 141763{,}3056)^{-1} = \frac{1}{956777{,}3154} = 0{,}000001045$$

De aquí, calculamos L_q:

$$L_q = \left(\frac{(\lambda_3/\mu_1)^s \rho_3}{s!(1-\rho_3)^2}\right)\pi(0) = \left(\frac{(0{,}0412/0{,}00329)^{13} 0{,}9633}{13!(1-0{,}9633)^2}\right) \cdot 0{,}000001045 =$$
$$= 21392732 \cdot 0{,}000001045 = 22{,}3554$$

Así W_q, por la fórmula de Litlle:

$$W_q = \frac{L_q}{\lambda_3} = \frac{22{,}3554}{0{,}0412} = 542{,}6069$$

Como $W = W_q + 1/\mu_1$,

$$W = 542{,}6069 + \frac{1}{0{,}00329} = 846{,}5583$$

Finalmente, por la fórmula de Little,

$$L = \lambda_3 \cdot W = 0{,}0412 \cdot 846{,}5583 = 34{,}8782$$

Con todo tenemos

$$\boxed{L_q = 22{,}3554; L = 34{,}8782; W_q = 542{,}6069; W = 846{,}5583}$$

6. **Número de servidores** Calculamos el número más adecuado de servidores para las 4 colas estudiadas, siguiendo el enfoque de calidad y eficiencia presentado en la sección 4.5. Tanto el enfoque de calidad como el enfoque de eficiencia, nos dan resultados interesantes en caso de haber una variación de la intensidad del tráfico $r = \lambda/\mu$ del sistema, por lo que nos centraremos solo en el enfoque de **calidad y eficiencia**. Fijamos para ello un nivel de calidad deseado $\alpha = 0{,}2$, es decir, que solo el 20 % de los clientes que entran en el sistema tengan que esperar un tiempo $t > 0$ en la cola. Si $\alpha = 0{,}2$, siguiendo la ecuación (4.11),

$$\beta = 1{,}0615$$

ya que

$$\alpha = \frac{\phi(\beta)}{\phi(\beta) + \beta\Phi(\beta)} = \frac{0{,}2271}{0{,}2271 + 1{,}0615 \cdot 0{,}8558} = 0{,}2$$

Recordemos que con r notamos $r = \lambda/\mu$.

a) **Cola del build**.
Primero, calculamos el nivel de calidad actual con la fórmula C-Erlang

$$C(c,r) = 1 - W_q(0) = \frac{\frac{r^c}{c!(1-\rho)}}{\left(\frac{r^c}{c!(1-\rho)} + \sum_{n=0}^{c-1} \frac{r^n}{n!}\right)} =$$

$$= \frac{\frac{11{,}6729^{13}}{13!(1-0{,}8979)}}{\left(\frac{11{,}6729^{13}}{13!(1-0{,}8979)} + \sum_{n=0}^{12} \frac{11{,}6729^n}{n!}\right)} = \frac{117115{,}6869}{117115{,}6869 + 71972} = 0{,}6193$$

Para conseguir que el nivel de calidad α sea 0,2, el número de servidores debería ser

$$s = r + \beta\sqrt{r} = 11{,}6729 + 1{,}0615 \cdot \sqrt{11{,}6729} = 11{,}6729 + 3{,}6266812 = 15{,}2996$$

Actualmente esta cola tiene 13 servidores.

b) **Cola de la tarjeta** 1.
Primero, calculamos el nivel de calidad actual con la fórmula C-Erlang

$$C(c,r) = 1 - W_q(0) = \frac{\frac{r^c}{c!(1-\rho)}}{\left(\frac{r^c}{c!(1-\rho)} + \sum_{n=0}^{c-1} \frac{r^n}{n!}\right)} =$$

$$= \frac{\frac{15{,}5623^{17}}{17!(1-0{,}9154)}}{\left(\frac{15{,}5623^{17}}{17!(1-0{,}9154)} + \sum_{n=0}^{16} \frac{15{,}5623^n}{n!}\right)} = \frac{6120902{,}119}{6120902{,}119 + 3495053{,}118} = 0{,}6365$$

Para conseguir que el nivel de calidad α sea 0,2, el número de servidores debería ser

$$s = r + \beta\sqrt{r} = 15{,}5623 + 1{,}0615 \cdot \sqrt{15{,}5623} = 15{,}5623 + 4{,}1875 = 19{,}7498$$

Actualmente esta cola tiene 17 servidores.

c) **Cola de la tarjeta** 2.
Primero, calculamos el nivel de calidad actual con la fórmula C-Erlang

$$C(c,r) = 1 - W_q(0) = \frac{\frac{r^c}{c!(1-\rho)}}{\left(\frac{r^c}{c!(1-\rho)} + \sum_{n=0}^{c-1} \frac{r^n}{n!}\right)} =$$

$$= \frac{\frac{9{,}8784^{18}}{18!(1-0{,}5488)}}{\left(\frac{9{,}8784^{18}}{18!(1-0{,}5488)} + \sum_{n=0}^{17} \frac{9{,}8784^n}{n!}\right)} = \frac{277{,}7466}{277{,}7466 + 19260} = 0{,}0142$$

Para conseguir que el nivel de calidad α sea 0,2, el número de servidores debería ser

$$s = r + \beta\sqrt{r} = 9{,}8784 + 1{,}0615 \cdot \sqrt{9{,}8784} = 9{,}8784 + 3{,}3363 = 13{,}2147$$

Actualmente esta cola tiene 18 servidores.

d) **Cola de la tarjeta** 3.
Primero, calculamos el nivel de calidad actual con la fórmula C-Erlang

$$C(c,r) = 1 - W_q(0) = \frac{\frac{r^c}{c!(1-\rho)}}{\left(\frac{r^c}{c!(1-\rho)} + \sum_{n=0}^{c-1} \frac{r^n}{n!}\right)} =$$

$$= \frac{\frac{12{,}5228^{13}}{13!(1-0{,}9633)}}{\left(\frac{12{,}5228^{13}}{13!(1-0{,}9633)} + \sum_{n=0}^{12} \frac{12{,}5228^n}{n!}\right)} = \frac{815027{,}7557}{815027{,}7557 + 141760} = 0{,}8518$$

Para conseguir que el nivel de calidad α sea 0,2, el número de servidores debería ser

$$s = r + \beta\sqrt{r} = 12{,}5228 + 1{,}0615 \cdot \sqrt{12{,}5228} = 12{,}5228 + 3{,}7564 = 16{,}2792$$

Actualmente esta cola tiene 13 servidores.

6.4. Conclusiones prácticas

A partir de los cálculos realizados, se desprenden las siguientes ideas:

1. Hemos visto que las 4 colas estudiadas son estables y tienen distribución estacionaria, por lo que el sistema es funcional y cumple su objetivo.

2. Tanto la cola de los build, como las colas de la tarjeta 1 y 3 tienen niveles de ocupación y tiempos de espera superiores a los deseados. En cambio, la cola de la tarjeta 2 tiene mucha menos intensidad de uso, con un tiempo medio de espera en la cola de 0,5 minutos.

3. En el caso de la cola de los build, a partir de las medidas de efectividad calculadas, constatamos el hecho de que esta cola supone un cuello de botella para colas menos ocupadas, como es la de la tarjeta 2.

4. Una mejor distribución de los ordenadores dedicados a cada cola mejoraría el rendimiento y calidad de éstas: la cola de la tarjeta 2 podría tener hasta 5 ordenadores menos; en cambio, la cola de los build y las de las tarjetas 1 y 3 necesitarían cada una entre 2 y 3 ordenadores dedicados más

Tabla resumen de los cálculos realizados:

	ρ	L_q	L	W_q	W	s actuales	s según QED
Cola de build	0,8979	5,45	17,13	43,66	137,12	13	15,30
Cola tarjeta 1	0,9154	6,89	22,45	134,52	438,47	17	19,75
Cola tarjeta 2	0,5488	0,02	9,89	0,53	304,48	18	13,21
Cola tarjeta 3	0,9633	22,35	34,87	542,61	846,56	13	16,30

7. Conclusiones

Durante este trabajo hemos tratado las principales ideas en relación con el proceso de Poisson y con las cadenas de Markov en tiempo continuo para desarrollar el objetivo principal: los sistemas de colas $M/M/s$. A partir de más de 19 fuentes bibliográficas hemos logrado introducir al lector los conceptos más básicos de la Teoría de Colas y presentar una realidad empresarial, a modo de ejemplo práctico, donde esta teoría es de gran utilidad.

Hemos empezado este texto con la metáfora de la cola en el monte Everest el año 2019. La rescatamos para ratificar la idea, constatada en este trabajo, que la Teoría de Colas tiene una enorme aplicación práctica y ayuda a resolver problemas actuales y compartidos por todos. Uno puede acudir a las numerosas conferencias y artículos publicados, como por ejemplo los que podemos encontrar en la Real Sociedad Matemática Española[3] o la *Royal Society*[4], para percatarse del extraordinario papel que cada vez más están realizando las Matemáticas para mejorar la vida de la gente. Es mi parecer pensar que hemos de potenciar el poder que muchas áreas de las Matemáticas tienen para incrementar el bienestar de la sociedad. Sin duda, la Teoría de Colas están entre estas áreas, ya que bienestar también es que una empresa funcione más eficientemente gracias a la modelización de sus colas, es que un hospital atienda mejor y más rápido a sus pacientes o que las infraestructuras y servicios públicos sean fluidos y absorban a toda la demanda.

A fecha de redacción de este trabajo, está sucediendo un extraordinario y devastador evento global: la pandemia del Covid-19. Ya hemos mencionado el papel importante que juega la Teoría de Colas en los sistemas de salud. Por ejemplo, en un artículo del doctor Adolfo Crespo de la Universidad de Sevilla [5] se hace hincapié en la relevancia que tienen modelos analíticos de sistemas de colas para afrontar epidemias, permitiendo tomar decisiones óptimas a la hora de destinar los recursos sanitarios, calcular la capacidad que un sistema sanitario tiene para afrontar ciertos niveles de infección y analizar con detalle

[3]Real Sociedad Matemática Española: rsme.es

[4]Royal Society: royalsociety.org

[5]Título del artículo: A COVID-19 Recovery Strategy Based on the Health System Capacity Modeling. Implications on Citizen Self-Management. Enlace: https://idus.us.es/handle/11441/95407

las fortalezas y debilidades de cada organización sanitaria regional. Sin olvidar también la gestión de las colas que se forman para realizar pruebas PCR o para procesar éstas en los laboratorios. Hay muchos otros estudios sobre la aplicación de la Teoría de Colas en las pandemias víricas. Queremos destacar uno de la Universidad Vrije de Amsterdam[6] sobre la relación entre la transmisión de enfermedades infecciosas y los modelos $M/G/1$ de colas y otro de la Universidad Tecnológica de Malasia[7] sobre la modelización y el análisis del virus del Ébola a partir de la Teoría de Colas con la que, afirma, reduce el coste computacional y mejora la predicción frente a otro tipo de métodos como los que usan ecuaciones diferenciales ordinarias. Es decir, la Teoría de Colas no solo es útil para la gestión de los recursos necesarios, sino que también para estudiar propiamente la pandemia y ser capaces de explicar la transmisión y las posibilidades que hay de controlarla. Como hemos vivido todos nosotros, una mejor gestión de una pandemia no solo supone proteger la salud de los ciudadanos, sino que también asegurar que las libertades constitucionales no han de ser restringidas para evitar una saturación del sistema. Creo pues que ésta es una prueba definitiva del valor útil de la Teoría de Colas. Esta conclusión hace que me sienta muy satisfecho de haber escogido este tema para el trabajo.

Durante la realización del trabajo he tenido que afrontar retos que hasta ahora no me había encontrado (o no tanto) en el Grado en Matemáticas. La redacción de resultados matemáticos y sus demostraciones de manera rigurosa, la recopilación de fuentes bibliográficas en la biblioteca y por la red, la síntesis de la diferente información, el uso de nuevas herramientas como *Tex*, el trabajo coordinado y sostenido con el tutor y el contraste entre fuentes bibliográficas de distinto valor, serían un ejemplo de ellos. A lo largo del trabajo, he ido cometiendo incorrecciones por la falta de experiencia y de conocimiento, de los cuales me ha ido advirtiendo el tutor o me he dado cuenta yo mismo al ir avanzando. De esta manera, el proceso me ha permitido aprender a través de los errores y asimilar no solo los conocimientos teóricos propios del tema del trabajo, sino que otras muchas enseñanzas como: capacidad de análisis y de resolución de problemas en el ámbito académico, organización del calendario para cumplir con la entrega, capacidad de interpretación y síntesis de fuentes bibliográficas, herramientas para el lenguaje matemático para enunciar resultados y demostrarlos, asimilación del rigor matemático y, sobre todo, experiencia para poder llevar a cabo otros futuros trabajos académicos. De esta manera, hago un balance muy positivo del Trabajo Final de Grado: todos estos aprendizajes junto con la implicación personal y la curiosidad sobre la materia, además del vínculo profesional con el caso práctico, han hecho que me haya sentido muy realizado y acabe satisfecho con los aprendizajes y conocimientos obtenidos.

Pese a no ser fácil y suponer todo un reto, creo que he cumplido los objetivos de este trabajo. Entre ellos, especialmente el relacionado con mi actividad profesional en Giesecke+Devrient Mobile Security Iberica, por el que he podido conectar la parte más abstracta con una realidad que observo día tras día. Trasladaré las conclusiones prácticas de este trabajo a mis compañeros para así aplicar *in situ* los resultados.

Con todo, este Trabajo Final de Grado cierra una importante etapa en el Grado en Matemáticas, pero será también la cabecera de las nuevas etapas que están por venir.

[6] A useful relationship between epidemiology and queueing theory: The distribution of the number of infectives at the moment of the first detection. Enlace: http://www.few.vu.nl/∼rplanque/resources/PapersForProject/trapman.pdf

[7] Queueing theory based model and network analysis for predicting the transmission and control of Ebola virus desease. Enlace: http://eprints.utm.my/id/eprint/79267/1/ChinyereOgochukwuDikePFS2018.pdf

Referencias

[Beasley 2011] Beasley, JE: *Operations Research notes. Queuing theory.* 2011. – URL `http://people.brunel.ac.uk/ mastjjb/jeb/or/contents.html`. – Department of Mathematical Sciencies, Brunel University. West London, Reino Unido.

[Corcuera 2019a] Corcuera, José M.: *A Course in Stochastic Processes.* 2019. – Facultat de Matemàtiques i Enginyeria informàtica. Universitat de Barcelona, España.

[Corcuera 2019b] Corcuera, José M.: *Statistics.* 2019. – Facultat de Matemàtiques i Enginyeria informàtica. Universitat de Barcelona, España.

[Durret 1999] Durret, Rick: *Essentials of Stochastic Processes.* Department of Mathematics, Cornell University, New York : Springer, 1999

[Feller 1966] Feller, William: *An Introduction to Probability Theory and Its Applications. Volume II. Second edition.* 1966. – Department of Mathematics. Princenton University. Nueva Jersei, Estados Unidos.

[Green 2011] Green, Linda: *Queuing theory and modeling.* Graduate School of Business, Columbia University, New York, New York. : Yuehwern Yih, 2011

[Khoshnevisan 2011] Khoshnevisan, Davar: *The Strong Markov Property.* 2011. – URL `https://www.math.utah.edu/ davar/math7880/S11/Chapters/Ch9.pdf`. – Probability and Statistics School. University of Utah. Utah, Estados Unidos.

[Kolesar und Green 1998] Kolesar, Peter ; Green, Linda: *Insights on service system design from a normal approximation to Erlang's delay formula.* 1998. – Graduate School of Business, Columbia University. New York, New York, Estados Unidos.

[Mitrofanova 2007] Mitrofanova, Antonina: *Lecture 3: Continuous times Markov chains. Poisson Process. Birth and Death process.* 2007. – URL `https://cs.nyu.edu/mishra/COURSES/09.HPGP/scribe3`. – Department of Computer Science. New York University, New york, Estados Unidos

[Neeman 2019] Neeman, Joe: *The Markov property.* 2019. – URL `https://wt.iam.uni-bonn.de/fileadmin/WT/Inhalt/people/Joe_Neeman/introduction.pdf`. – Institute for Applied Mathematics. University of Bonn. Bonn, Alemania.

[Norris 1997] Norris, J.R.: *Markov Chains.* 1997. – Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, Reino Unido.

[Omahen und Marathe 1975] Omahen, K. ; Marathe, V.: *Analysis and Applications of the Delay Cycle for the M/M/c Queueing System.* 1975. – Department of Computer Science. Purdue University, Indiana, Estados Unidos.

[Roch 2012] Roch, Sebastien: *Lecture 22: Strong Markov Property.* 2012. – URL `https://www.math.wisc.edu/ roch/teaching_files/275b.1.12w/lect22-web.pdf`. – Department of Mathematics. University of Wisconsin-Madison. Wisconsin, Estados Unidos.

[Shlomo und Whitt 1981] Shlomo, Halfin ; Whitt, Ward: *Heavy-Traffic Limits for Queues with Many Exponential Servers. Operations Research.* 1981. – Bell Laboratories. Holmdel, New Jersey, Estados Unidos

[Shortle u. a. 2018] Shortle, J. F. ; Thompson, J. M. ; Gross, D. ; Harris, C. M.: *Fundamentals of queueing theory (Vol. 399)*. 2018. – John Wiley and Sons. Hoboken, New Jersey, Estados Unidos.

[Sigman 2009] Sigman, Karl: *Notes on Little's Law*. 2009. – URL `http://www.columbia.edu/ ks20/stochastic-I/stochastic-I-LL.pdf`. – Department of Industrial Engineering and Operations Research. Columbia University, Estados Unidos.

[Stidham 1972] Stidham, Jr. S.: *$L = \lambda W$: A discounted analogue and a New Proof*. 1972. – Institute for Operations Research and the Management Sciences (INFORMS). Maryland, Estados Unidos.

[Takahara 2017] Takahara, Glen: *Stochastic Processes. Continuous Time Markov chains, set 5*. 2017. – URL `https://mast.queensu.ca/ stat455/lecturenotes/set5.pdf`. – Department of Mathematics and Statistics. Queen's University, Ontario, Canada.

[Whitt 2013] Whitt, Ward: *Continuous-time Markov chains*. 2013. – Department of Industrial Engineering and Operations Research. Columbia University, New york, Estados Unidos.