MATerials MATemàtics
Volum 2009, treball no. 5, 33 pp. ISSN: 1887-1097
Publicació electrònica de divulgació del Departament de Matemàtiques
de la Universitat Autònoma de Barcelona
www.mat.uab.cat/matmat

Recordando a Erlang: Un breve paseo (sin esperas) por la Teoría de Colas

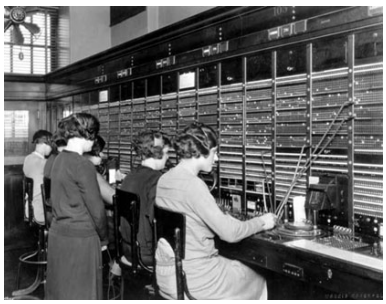
Rosario Delgado de la Torre

"Aunque algunos puntos dentro del campo de la Telefonía dan lugar a problemas cuya solución corresponde a la Teoría de la Probabilidad, ésta última no ha sido muy utilizada en este campo, hasta donde podemos ver. En este sentido, la Telephone Company of Copenhagen constituye una excepción puesto que su director gerente, el Sr. F. Johannsen, ha aplicado durante varios años los métodos de la Teoría de la Probabilidad a la solución de varios problemas de importancia práctica, además de incitar a otros a trabajar en investigaciones de características similares. Como creo que algún que otro punto de este trabajo puede resultar de interés, y no es en absoluto necesario un conocimiento especial de los problemas telefónicos para su entendimiento, daré cuenta de ello a continuación."

Esta es la introducción del artículo que el matemático danés Agner Krarup Erlang (Longborg, Dinamarca, 1878-1929) publicó en 1909[1] y que se

[1] "The Theory of Probabilities and Telephone Conversations", publicado originalmente en *Nyt Tidsskrift for Matematik* B, Vol. 20 (1909), p. 33. Su traducción al inglés se puede encontrar en [1], junto con otros trabajos y una biografía muy documentada de su autor.

considera el primer artículo de la Teoría de Colas. Así es, por tanto, que en 2009 celebramos el centenario de esta teoría, y la introducción del artículo de Erlang no puede ser más clara al relacionar de manera directa sus inicios con la resolución de problemas derivados de la telefonía.



El embrionario trabajo de Erlang fue continuado por diversos investigadores en la primera mitad del siglo XX, como Pollaczek, Kolmogorov y Khintchine, entre otros. A partir de la década de los años 50 hubo un considerable crecimiento de este área, y su interés ha ido yendo en aumento debido, en parte, al gran desarrollo del ámbito de las telecomunicaciones, uno de los campos donde la Teoría de Colas tiene una mayor implicación. Actualmente, la sofisticada teoría desarrollada durante estos años tiene un abanico muy amplio de aplicaciones, desde el campo de la telefonía y las telecomunicaciones donde se inició, hasta la ciencia de la computación, los procesos industriales manufactureros (cadenas de producción), el control de tráfico (por carretera, aéreo, de personas, de información a través de Internet,...) o la logística militar y civil, pasando por las empresas de servicios (hospitales, oficinas bancarias, restaurantes de comida rápida, supermercados, etc.).

Este artículo tiene un doble objetivo: por una parte, dar cuenta de la celebración del centenario de la Teoría de Colas; por otra, presentar una breve introducción a lo más clásico de dicha teoría, que pueda resultar de utilidad para aquéllos que teniendo unos conocimientos elementales de probabilidad desean introducirse en ella de manera rápida y sin demasiadas complicaciones.

1. Introducción

En los albores de este siglo XXI, como en el pasado, los ciudadanos del llamado “primer mundo” sufrimos como una lacra la lamentable pérdida de tiempo que nos supone el hecho de tener que esperar en una cola. Ya sea en los atascos diarios de entrada o salida de las grandes urbes, en los peajes de las autopistas, al ir a pagar nuestra compra en

el supermercado, al lavar el coche en un túnel de lavado, o en la peluquería, el tiempo perdido esperando a que nos sirvan siempre es enojoso y se nos antoja demasiado. Y esto parece ser una consecuencia directa e inevitable de la sociedad tecnológica en la que vivimos.

La razón de que se produzcan las colas de espera es obvia: éstas aparecen cuando la demanda de servicio por parte de los clientes supera la capacidad de servicio del sistema. También resulta obvio que la espera en cola no sólo disgusta a los clientes que la padecen sino también a los responsables del sistema que la provoca. Este problema se podría solventar añadiendo suficiente número de servidores como para eliminar casi completamente las colas, aunque por motivos económicos y de otra índole (espacio, etc.) hacer esto no suele ser posible. Sin embargo, sí pudiera ser conveniente realizar una mejora añadiendo algunos servidores. Es necesario, por tanto, estudiar estos sistemas con el fin de poder conocer su funcionamiento (por ejemplo, cuánto tiempo han de esperar los clientes o cuántos clientes hay en cola, en promedio) y así poder decidir si es factible o no, en función de la previsible consecuente mejora y de la inversión necesaria para ello, la ampliación del número de servidores del sistema.

La parte de la Teoría de la Probabilidad que estudia los modelos matemáticos que se utilizan para tratar estos sistemas en los que hay "colas" o "líneas de espera" de clientes se conoce como **Teoría de Colas** ("Queueing Theory" en inglés).

Los sistemas de colas que se consideran involucran "llamadas", "clientes" o "trabajos" que llegan para ser servidos. Puede ser que el servicio lo proporcionen uno o varios servidores. Cuando un cliente llega al sistema, si hay algún servidor libre pasa inmediatamente a ser servido, y cuando acaba abandona el sistema [2]. En caso contrario, pueden darse diversas situaciones:

(a) El cliente se queda esperando en el sistema. En los sistemas más simples, que son los únicos que consideraremos, los clientes en espera forman una única cola, independientemente del número de servidores. Cuando un servidor queda libre, el primer cliente de la cola (es decir, el primero en

[2] También se podría dar el caso de que cuando el cliente acaba su servicio con un servidor del sistema, necesite ser servido por otro servidor, o incluso por el mismo de nuevo. Este tipo de sistemas se conocen como "redes" y son más complicados de estudiar. En este artículo nos restringiremos al caso de que cada cliente sólo necesita ser servido una vez por algún servidor del sistema y todos esperan a ser servidos, si tienen que hacerlo, en una única cola.

llegar de todos los que están en ella) pasa a ser atendido por el servidor y el siguiente ocupa su lugar en la cola. Se dice entonces que se utiliza una disciplina de servicio **FIFO** (del inglés *First-In-First-Out*), es decir, el primero en llegar es el primero en ser servido. Notemos además que suponemos que ningún servidor se encuentra parado si hay clientes esperando para ser atendidos (se dice, en inglés, que la disciplina de servicio es *work-conserving* cuando se da esta circunstancia).

Ésta será una suposición básica a lo largo de todo el artículo: que la disciplina de servicio es FIFO y *work-conserving*.

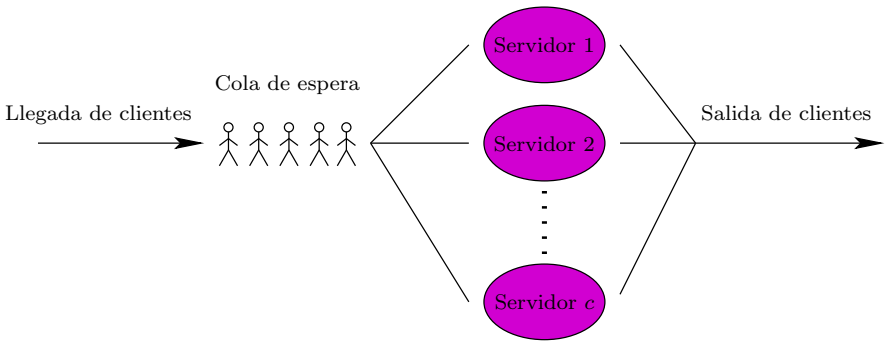


Figura 1: Esquema general de una cola con c servidores

(b) Como en el caso anterior, el cliente se queda esperando en el sistema pero al cabo de cierto tiempo sin haber sido servido, se cansa de esperar y se va (se habla de *clientes impacientes*). En este caso hay a su vez diversas posibilidades: se va y no vuelve, o se va pero se queda en el sistema formando lo que se conoce como "órbita" para, al cabo de cierto tiempo, volver a intentarlo de nuevo (son las llamadas *colas con reintentos*).

(c) El cliente que no puede ser atendido inmediatamente abandona el sistema (se produce lo que se llama una *pérdida*).

En este artículo nos centraremos en el caso (a), en la Sección 3 para un único servidor, y en la 4 para múltiples servidores, y también en el caso (c) en la Sección 5. Para este último caso, veremos la conocida *función de pérdida de Erlang* o *Erlang B* que apareció en el artículo de Erlang de 1917[3]. En

[3] "Solution of some problems in the Theroy of Probabilities of significance in automatic

la sección 4 también veremos otra famosa fórmula de Erlang introducida en el mismo trabajo, la llamada *Erlang C*. El caso (b) puede resultar mucho más complicado y no lo trateremos; en particular, un servidor se puede ver obligado a estar inactivo aunque haya algún cliente en el sistema, si éstos se encuentran en la "órbita", hasta que se produzca un reintento.

Asociado a un sistema de colas como los descritos, tenemos el correspondiente modelo matemático; como las llegadas de clientes al sistema y/o los tiempos de servicio de los clientes, son aleatorios, el modelo matemático adecuado es estocástico y ha de ser tratado con ayuda de la Teoría de la Probabilidad.

2. Los Procesos de Nacimiento y Muerte

Los modelos matemáticos más simples de la Teoría de Colas son los llamados Procesos de Nacimiento y Muerte. Estos procesos estocásticos son un caso particular de Cadena de Markov a tiempo continuo. Vamos a ver en primer lugar qué tipo de procesos son las Cadenas de Markov y luego daremos su definición.

2.1. Conceptos básicos

Definición: Una Cadena de Markov a tiempo continuo es un proceso estocástico $X = \{X_t,\, t \geq 0\}$ a valores en un conjunto de estados E finito o numerable[4], que cumple la llamada *"propiedad de Markov"* de independencia entre el futuro y el pasado conocido el presente; esta propiedad se expresa de la siguiente manera mediante la probabilidad condicionada[5]:

$$P\left(X_{t_{n+1}} = i_{n+1} \,/\, X_{t_n} = i_n,\, \ldots,\, X_{t_1} = i_1\right) = P\left(X_{t_{n+1}} = i_{n+1} \,/\, X_{t_n} = i_n\right)$$

telephone exchanges", publicado originalmente en *Elektroteknikeren* Vol. 13 (1917), p. 5. Su traducción al inglés se puede consultar en [1].

[4]Un proceso estocástico es una familia de variables aleatorias, todas en el mismo espacio de probabilidad y tomando valores en el mismo espacio de estados E; en este caso, la familia está indexada por $t \in [0, +\infty)$, que juega el papel del tiempo, esto es, para todo t, X_t es una variable aleatoria.

[5]Dados dos sucesos A y B, tales que $P(B) > 0$, se define la probabilidad condicionada de A a (o por) B así: $P(A/B) = \frac{P(A \cap B)}{P(B)}$ y representa la probabilidad de observar el suceso A sabiendo que se ha producido el suceso B. Se dice que los sucesos A y B són independientes si $P(A \,/\, B) = P(A) \cdot P(B))$, lo que equivale a decir que $P(A \,/\, B) = P(A)$ si $P(B) > 0$ y que $P(B \,/\, A) = P(B)$ si $P(A) > 0$.

si $P\left(X_{t_n} = i_n, \ldots, X_{t_1} = i_1\right) > 0$, $i_1, \ldots, i_n, i_{n+1} \in E$ y $0 \leq t_1 < t_2 < \cdots < t_n < t_{n+1}$.

Notemos que aquí el presente está representado por el instante t_n, el futuro por t_{n+1} y el pasado por los instantes $t_1, \ldots, t_{n-1}$, así que la propiedad de Markov nos dice que condicionar lo que pase en el instante t_{n+1} (el futuro) a lo que ha pasado anteriormente, es igual a condicionarlo sólo a lo que pasa en el presente (instante t_n).

Las *probabilidades de transición* de la cadena se definen así para $i, j \in E$ y $t \geq 0$:

$$P_{ij}(t) \stackrel{\text{def}}{=} P\left(X_{s+t} = j \,/\, X_s = i\right)$$

(la probabilidad de ir del estado i al estado j en un período de tiempo de longitud t) si la cadena es "homogénea", es decir, si las probabilidades anteriores no dependen del instante de inicio del período, $s \geq 0$.

Si para una cadena de Markov homogénea conocemos con qué probabilidad está la cadena en cada uno de los estados en un instante de tiempo fijado s (por ejemplo, en el origen $s = 0$), y también conocemos sus probabilidades de transición, entonces tenemos toda la información que necesitamos para poder saber, en cualquier otro instante de tiempo posterior, con qué probabilidad estará la cadena en cada uno de los estados posibles, esto es, para todo $t > s$ y para todo $j \in E$, podemos calcular usando la *Fórmula de Probabilidad Total*,

$$\begin{aligned} P(X_t = j) &= \sum_{i \in E} P(X_t = j \cap X_s = i) = \sum_{i \in E} P(X_t = j / X_s = i)\, P(X_s = i) \\ &= \sum_{i \in E} P_{ij}(t - s)\, P(X_s = i)\,. \end{aligned}$$

Por desgracia, en la mayoría de los casos no es posible conocer la expresión de las probabilidades de transición de la cadena en general, aunque a veces sí se pueden suponer conocidas (a partir del conocimiento de la evolución física del proceso que se pretende modelar) lo que se llama las *probabilidades de transición infinitesimales*, que son las $P_{ij}(h)$ para h pequeño. Esto es precisamente lo que sucede con el siguiente tipo particular de Cadena de Markov, los conocidos como Procesos de Nacimiento y Muerte:

Definición: Se dice que la Cadena de Markov homogénea a tiempo continuo $X = \{X_t,\, t \geq 0\}$, con espacio de estados E, es un Proceso de Nacimiento y Muerte si sus *probabilidades de transición infinitesimales* son de la siguiente forma para m y $n \in E$ y $h > 0$ pequeño, según estos dos casos:

(i) Si $E = \{0, 1, 2, \ldots\} = \mathbb{N} \cup \{0\}$,

$$P_{n\,m}(h) = \begin{cases} \lambda_n\, h + o(h) & \text{si } m = n+1,\ n \geq 0 \\ \mu_n\, h + o(h) & \text{si } m = n-1,\ n \geq 1 \\ 1 - (\lambda_n + \mu_n)\, h + o(h) & \text{si } m = n \geq 1 \\ 1 - \lambda_0\, h + o(h) & \text{si } m = n = 0 \\ o(h) & \text{en caso contrario} \end{cases}$$

donde $o(h)$ es la notación habitual para denotar una expresión función de h tal que $\lim\limits_{h\to 0} \dfrac{o(h)}{h} = 0$, para ciertas $\lambda_n > 0\,(n \geq 0)$ y $\mu_n \geq 0\,(n \geq 1)$.

(ii) Si $E = \{0, 1, 2, \ldots, M\}$, $M \geq 1$,

$$P_{n\,m}(h) = \begin{cases} \lambda_n\, h + o(h) & \text{si } m = n+1,\ 0 \leq n \leq M-1 \\ \mu_n\, h + o(h) & \text{si } m = n-1,\ 1 \leq n \leq M \\ 1 - (\lambda_n + \mu_n)\, h + o(h) & \text{si } 1 \leq m = n \leq M-1 \\ 1 - \lambda_0\, h + o(h) & \text{si } m = n = 0 \\ 1 - \mu_M\, h + o(h) & \text{si } m = n = M \\ o(h) & \text{en caso contrario} \end{cases}$$

para ciertas $\lambda_n > 0\,(0 \leq n \leq M-1)$ y $\mu_n \geq 0\,(1 \leq n \leq M)$.

Interpretación: En este tipo de procesos, X_t representa el número de individuos de cierto tipo en el instante t. En este caso, suponer la propiedad de Markov es algo bastante natural, así como la homogeneidad del proceso en el tiempo (la evolución de la población de individuos no depende de en qué instante se observa, y el futuro es independiente del pasado, conocida la población en el presente). El caso (i) corresponde a que no haya limitación en el número máximo de individuos permitidos en la población, mientras que en el caso (ii), sí.

En cuanto a las probabilidades de transición infinitesimales, se obtienen a partir de los siguientes principios básicos asumidos sobre la evolución real de la población:

(a) Cuando hay $n \geq 0$ individuos en la población y $n+1 \in E$, la probabilidad de que "nazca" o "llegue" un nuevo individuo en un intervalo de tiempo

de longitud $h > 0$ pequeña es aproximadamente proporcional a h, esto es, es de la forma: $\lambda_n\, h + o(h)$. λ_n se llama "tasa de nacimiento (o llegada) cuando hay n individuos".

(b) Cuando hay $n \geq 1$ individuos en la población, la probabilidad de que "muera" o "se vaya" uno de ellos en un intervalo de tiempo de longitud $h > 0$ pequeña es aproximadamente proporcional a h, de la forma: $\mu_n\, h + o(h)$. μ_n se llama "tasa de defunción (o salida, o servicio) cuando hay n individuos".

(c) La probabilidad de que pase más de un evento a la vez, en un intervalo de longitud $h > 0$ pequeña es $o(h)$, donde por "evento" entendemos un nacimiento (o llegada) o una muerte (o salida).

(d) Los individuos nacen (llegan), mueren (se van), o no hacen nada, independientemente los unos de los otros.

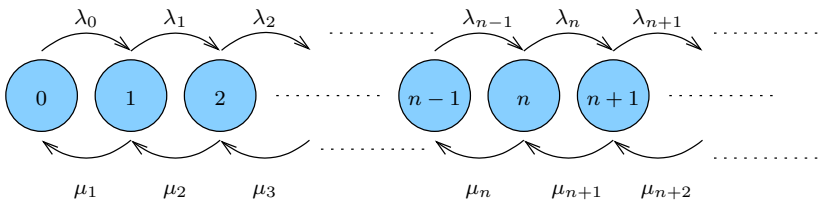


Figura 2: Tasas para un proceso de nacimiento y muerte, caso (i)

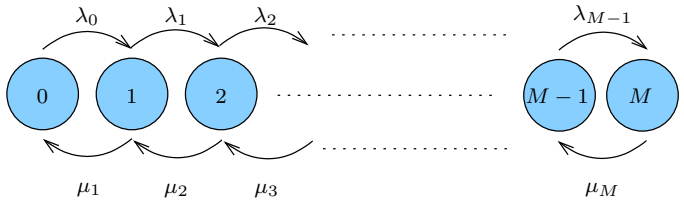


Figura 3: Tasas para un proceso de nacimiento y muerte, caso (ii)

2.2. Casos particulares interesantes

(i) Los **procesos de nacimiento y muerte lineales**, en los que $E = \mathbb{N} \cup \{0\}$, $\lambda_n = \lambda\, n$ y $\mu_n = \mu\, n$, con $\lambda > 0$, $\mu > 0$. La interpretación es sencilla: si cada individuo tiene probabilidad $\lambda\, h + o(h)$ de dar lugar a un nuevo individuo en un intervalo de tiempo de longitud $h > 0$ pequeña, y hay n individuos en el sistema, entonces la probabilidad de que se produzca un nacimiento en ese intervalo de tiempo será $\lambda\, n\, h + o(h)$. Lo mismo con las defunciones: si cada individuo tiene probabilidad $\mu\, h + o(h)$ de fallecer en un intervalo de tiempo de longitud $h > 0$ pequeña, y hay n individuos en el sistema, entonces la probabilidad de que se produzca un fallecimiento en ese intervalo de tiempo será $\mu\, n\, h + o(h)$.

(ii) Los **procesos de nacimiento puro lineales**. Es el caso particular del apartado anterior en el que $\mu = 0$, es decir, sólo se producen nacimientos, no defunciones. En esta situación, a partir de las probabilidades de transición infinitesimales se pueden determinar el resto, es decir, las probabilidades $P_{n\,m}(t)$ para todo tiempo $t > 0$, y todo n, $m \in E = \mathbb{N} \cup \{0\}$, aunque en general esto no es posible, ni siquiera para procesos de nacimiento puro no lineales cualesquiera. Por ello el estudio de los procesos de nacimiento y muerte en general se limita al de su *comportamiento asintótico* (cuando t crece). Podemos considerar que el proceso de nacimientos, considerado de manera aislada del resto, es un proceso de este tipo.

(iii) El **proceso de Poisson** es un proceso de nacimiento puro no lineal, en el que $E = \mathbb{N} \cup \{0\}$ y las tasas de llegadas son todas constantes e iguales a cierta tasa λ llamada "intensidad" del proceso, es decir,

$$\lambda_n = \lambda > 0 \quad \forall n \geq 0\,.$$

Este proceso modela bien las llegadas "totalmente al azar"[6] al sistema de clientes o llamadas. Es precisamente el artículo de Erlang de 1909 el primero en el que se justifica este hecho. Para este caso particular, a

[6] Que las llegadas se produzcan "totalmente al azar" es una manera de expresar el hecho de que se supone que el número de llegadas en intervalos disjuntos de tiempo son variables aleatorias independientes. Dadas dos variables aleatorias Y, Z en el mismo espacio de probabilidad y con espacio de estados S finito o numerable se dice que son *independientes* si para todo $y, z \in S$, los sucesos $A = \{Y = y\}$ y $B = \{Z = z\}$ son independientes, esto es, $P(A \cap B) = P(A) \cdot P(B)$.

partir de las probabilidades de transición infinitesimales se obtiene que el número de llegadas $\{X_t,\, t \geq 0\}$ cumple:

- $X_0 = 0$
- los incrementos son independientes (és decir, si $t_1 < t_2 < t_3 < t_4$, las variables aleatorias $X_{t_2} - X_{t_1}$ y $X_{t_4} - X_{t_3}$ son independientes).
- $X_t - X_s$ es una variable de Poisson de paràmetro $\lambda\,(t-s)$, $\forall 0 \leq s < t$, donde esto último quiere decir que

$$P(X_t - X_s = k) = e^{-\lambda\,(t-s)}\,\frac{(\lambda\,(t-s))^k}{k!} \quad \forall k \in \{0,\, 1,\, 2,\, \ldots\}\,.$$

Además, un hecho notable es que lo anterior es equivalente decirque si consideramos los instantes entre llegadas consecutivas de clientes o llamadas al sistema, esto es, los tiempos de espera entre llegadas, digamos τ_1, τ_2, ..., entonces estas variables aleatorias son **independientes e idénticamente distribuidas**, con distribución Exponencial de parámetro λ (la intesidad del proceso de Poisson!), esto es,

$$P(\tau_n \leq t) = 1 - e^{-\lambda\, t} \quad \forall t > 0 \text{ y } \forall n \geq 1\,.$$

(iv) Las **colas markovianas** más sencillas, que son las que consideraremos en este artículo, son procesos de nacimiento y muerte en los que los nacimientos son las llegadas de clientes al sistema, y las muertes o defunciones son las salidas de clientes del sistema cuando finaliza su servicio. Veremos ejemplos concretos de estas colas en las secciones siguientes, las conocidas como M/M/1, M/M/c y M/M/c/c.

2.3. Comportamiento asintótico

Ya hemos comentado que salvo algunos casos muy sencillos, en general es difícil o imposible determinar las probabilidades de transición $P_{n\,m}(t)$ para todo $m,\, n \in E$ y todo tiempo $t > 0$, y lo que se hace es considerar el comportamiento asintótico del proceso, el cual se podrá estudiar a partir de las tasas de nacimiento y defunción dadas por las probabilidades de transición infinitesimales.

Suponiendo que exista el siguiente límite, definimos

$$p_n = \lim_{t \to +\infty} P(X_t = n) \quad \text{para todo } n \in E\,,$$

que es la probabilidad de que haya n individuos o clientes en el sistema "cuando pasa mucho tiempo" o "en situación estacionaria".

La *distribución límite* de la cadena, que sería la dada por las probabilidades $\{p_n\}_{n\in E}$, no tiene por qué existir, en general. Sin embargo, se sabe que si todos los estados de la cadena comunican entre sí [7], se pueden plantear unas ecuaciones en las que las incógnitas son las p_n, conocidas como **ecuaciones de balance**, y resulta que: o bien la ecuaciones de balance sólo tienen una solución, la trivial idénticamente cero (y en ese caso no existe distribución límite), o bien las ecuaciones de balance tienen una única solución no trivial y entonces existe distribución límite y coincide con la solución (no trivial) de las ecuaciones de balance.

Para el caso de los procesos de nacimiento y muerte, todos los estados de la cadena comunican entre sí efectivamente si $\lambda_n > 0$ y $\mu_n > 0$ (ya que se puede ir de cada estado n al $n+1$ si $n \geq 0$ y $n+1 \in E$, con probabilidad $\lambda_n\, h + o(h) > 0$ para un $h > 0$ pequeño, y al $n-1$, si $n \geq 1$, con probabilidad $\mu_n\, h + o(h) > 0$), y de esta manera, pasando por todos los estados intermedios, se puede ir de cualquier estado i a cualquier otro estado j con probabilidad estrictamente positiva, en un tiempo finito).

Por tanto, el estudio del comportamiento asintótico del proceso pasa por plantear y resolver, si se puede, dichas ecuaciones de balance, según los casos:

(i) Si $E = \mathbb{N} \cup \{0\}$, las ecuaciones son

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

(ii) Si $E = \{0,\, 1,\, 2,\, \ldots,\, M\}$, con $M \geq 1$, son

$$\begin{cases} \mu_M\, p_M = \lambda_{M-1}\, p_{M-1} \\ (\lambda_n + \mu_n)\, p_n = \lambda_{n-1}\, p_{n-1} + \mu_{n+1}\, p_{n+1} & \text{para } 1 \leq n \leq M-1 \\ \lambda_0\, p_0 = \mu_1\, p_1 \end{cases} \tag{2}$$

La idea intuitiva que subyace tras estas ecuaciones es muy simple: se trata de un balance o equilibrio entre la tasa de entrada y la de salida de cada estado de la cadena. Así, para el estado $n \geq 1$ en el caso (i), (para el

[7] Se dice que dos estados de la cadena, por ejemplo los estados i y j, comunican entre sí, si existen $t_1 \geq 0$ y $t_2 \geq 0$ tales que $P_{i\,j}(t_1) > 0$ y $P_{j\,i}(t_2) > 0$.

estado n con $1 \leq n \leq M-1$ en el caso (ii)), vemos que la tasa de salida es $\lambda_n\, p_n$ (la de que haya una llegada y pasemos al estado $n+1$) más $\mu_n\, p_n$ (la de que haya una salida y pasemos al estado $n-1$), mientras que la tasa de entrada es la suma de la de entrar desde el estado $n-1$, que es $\lambda_{n-1}\, p_{n-1}$, ya que ha de ser con una llegada, y la de entrar desde el $n+1$, que es $\mu_{n+1}\, p_{n+1}$ puesto que ha de ser con una salida. Para el estado 0 el caso es más sencillo, pues sólo se puede salir del estado con una llegada (a tasa $\lambda_0\, p_0$) y entrar con una salida (a tasa $\mu_1\, p_1$). Y en el caso (ii), para el estado M también lo es, pues sólo se puede salir de ese estado con una salida (a tasa $\mu_M\, p_M$), y llegar a él con una llegada (a tasa $\lambda_{M-1}\, p_{M-1}$).

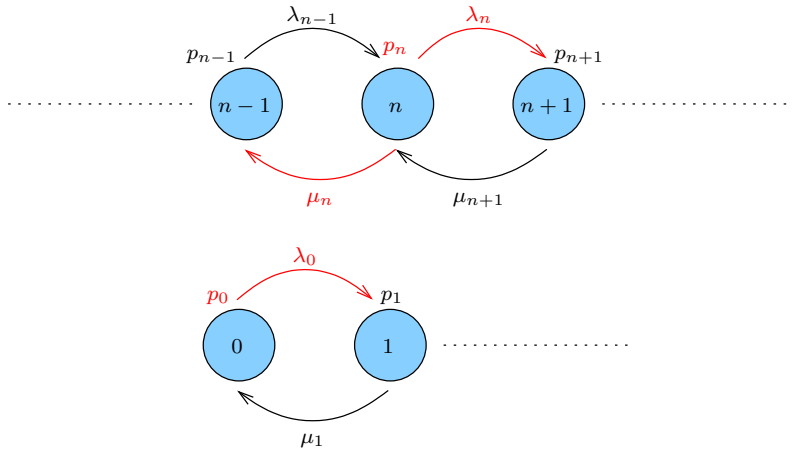


Figura 4: Tasas de entrada (en negro) y de salida (en rojo) para el caso (i)

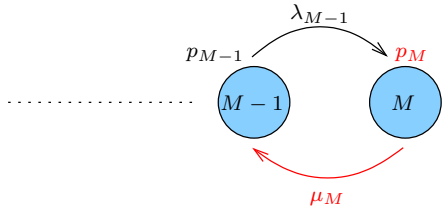


Figura 5: Tasas de entrada (en negro) y de salida (en rojo) para el estado M en el caso (ii)

Las ecuaciones (1) se pueden resolver recurrentemente de la siguiente

manera:

$$\begin{aligned}
p_1 &= \frac{\lambda_0}{\mu_1}\, p_0 \\
p_2 &= \frac{\lambda_1+\mu_1}{\mu_2}\, p_1 - \frac{\lambda_0}{\mu_2}\, p_0 = \frac{\lambda_1\,\lambda_0}{\mu_2\,\mu_1}\, p_0 \\
p_3 &= \frac{\lambda_2+\mu_2}{\mu_3}\, p_2 - \frac{\lambda_1}{\mu_3}\, p_1 = \frac{\lambda_2\,\lambda_1\,\lambda_0}{\mu_3\,\mu_2\,\mu_1}\, p_0 \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
p_n &= \frac{\lambda_{n-1}\,\cdots\,\lambda_1\,\lambda_0}{\mu_n\,\cdots\,\mu_2\,\mu_1}\, p_0 = p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}
\end{aligned}$$

e imponiendo que $\{p_n\}_{n\geq 0}$ sea una distribución de probabilidad, esto es, que $\sum_{n\geq 0} p_n = 1$, tenemos que

$$1 = p_0 + \sum_{n\geq 1} p_n = p_0 + \sum_{n\geq 1} \left(p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right)$$

lo que implica que

$$p_0 = \left(1 + \sum_{n\geq 1} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right)^{-1} \quad \text{y} \quad p_n = p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}\,, \quad n \geq 1 \tag{3}$$

y esta expresión tiene sentido si y sólo si

$$\sum_{n\geq 1} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} < +\infty\,. \tag{4}$$

Las ecuaciones (2) se resuelven de similar forma, obteniendo que

$$p_0 = \left(1 + \sum_{n=1}^{M} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right)^{-1} \quad \text{y} \quad p_n = p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}\,, \quad 1 \leq n \leq M \tag{5}$$

pero con la salvedad de que ahora **no** tenemos que imponer ninguna condición del tipo (4) para que exista la solución (cosa intuitiva, ya que si hay un número máximo permitido de clientes en el sistema, el número de clientes en el sistema **no** puede crecer sin control, aunque no impongamos ninguna restricción adicional).

3. La cola M/M/1 (un único servidor)

El modelo de colas denotado por **M/M/1** es el más simple de todos. Consiste en un único servidor (de ahí el **1** en la notación **M/M/1**) instalado en una estación de trabajo y trabajos o clientes que llegan para ser servidos, forman una cola si el servidor está ocupado a su llegada y cuando finalizan su servicio dejan el sistema.

El espacio de estados es $E = \mathbb{N} \cup \{0\}$ ya que no hay limitación de clientes en el sistema. Los clientes llegan a la estación siguiendo un proceso de Poisson de intensidad $\lambda > 0$ (los tiempos entre llegadas sucesivas son variables aleatorias independientes e idénticamente distribuidas, con ley exponencial de parámetro λ). A esto hace referencia la primera **M** de la notación utilizada para designar al modelo[8].

Los tiempos de servicio de los clientes también se suponen variables aleatorias independientes e idénticamente distribuidas, con ley exponencial de parámetro $\mu > 0$ (así que ésta es la razón de la segunda **M** en la notación). Por tanto, se trata de un *Proceso de Nacimiento y Muerte* con tasas de nacimiento $\lambda_n = \lambda$, y de defunción $\mu_n = \mu$, para todo n .

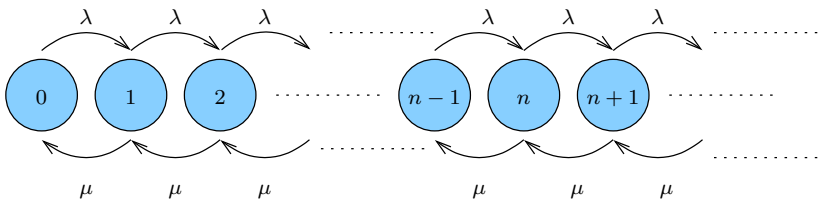


Figura 6: Tasas para la cola M/M/1

3.1. La distribución límite

Se define $\rho = \frac{\lambda}{\mu}$, la **intensidad de tráfico** del sistema o cola. Cuando $\rho \geq 1$, el número medio de llegadas al sistema excede o iguala al número

[8]Se utiliza una **M** y no una **E**, lo que podría parecer más lógico tratándose de la distribución exponencial, pues esta letra está reservada para la distribución de Erlang. Por otra parte, la **M** hace referencia al carácter "Markov" o de falta de memoria de la distribución exponencial. Esta nomenclatura es aceptada y generalizada en la actualidad, y es debida principalmente a Kendall (1953).

medio de salidas y se esperaría que al correr el tiempo la cola fuese creciendo sin parar ya que los clientes no abandonan el sistema hasta que no acaban su servicio (este hecho parece más claro si $\rho > 1$ que si $\rho = 1$, pero en este caso también es cierto ya que debido a la aleatoriedad de las llegadas, de tanto en tanto éstas serán más frecuentes que las salidas y entonces los clientes empezarán a acumularse en la cola; el caso contrario no puede darse ya que sólo pueden salir más clientes del sistema de los que entran mientras haya clientes en cola). Sólo en el caso $\rho < 1$ podemos esperar que el número de clientes en cola no crezca sin parar y podremos hablar del *steady state*, que es el estado de "equilibrio" al que llega el sistema después de haber estado operando un largo período de tiempo. Y resulta que efectivamente esto es así si miramos la condición (4): la condición necesaria y suficiente para que exista distribución límite en este caso es que

$$\sum_{n\geq 1}\prod_{i=1}^{n}\frac{\lambda_{i-1}}{\mu_i}=\sum_{n\geq 1}\rho^n<+\infty \Leftrightarrow \rho<1$$

(es la serie geométrica), y la suma de la serie si se da esta condición es:

$$\sum_{n\geq 1}\rho^n=\frac{1}{1-\rho}-1\,.$$

Entonces, a partir de (3) tenemos que

$$p_0=\left(\frac{1}{1-\rho}\right)^{-1}=1-\rho \quad \text{y} \quad p_n=p_0\,\rho^n=(1-\rho)\,\rho^n\,,\ n\geq 1\,,$$

es decir, que

$$\boxed{p_n=(1-\rho)\,\rho^n \quad \text{para todo } n\geq 0\,, \quad \text{si } \rho=\frac{\lambda}{\mu}<1}$$

y obtenemos la conocida fórmula de la función de probabilidad de la distribución geométrica de parámetro $\rho\,(\in (0,\,1))$.

3.2. Algunas medidas de efectividad

Vamos a definir y a ver cómo se pueden calcular algunas medidas de efectividad de la cola "a largo plazo", es decir, medidas sobre su comportamiento, a partir de la distribución límite

$$p_n=(1-\rho)\,\rho^n \quad \text{para todo } n\geq 0\,, \quad \text{con } \ \rho<1\,,$$

en función de las tasas de llegadas y salidas (o servicios), λ y μ respectivamente. Definimos en primer lugar la variable aleatoria

$$N = \text{“número de clientes en el sistema” (a largo plazo).}$$

Vemos que N toma los valores $n \in \{0, 1, 2, \ldots\}$ con probabilidades p_n. Sea también

$$L = \text{“número medio de clientes en el sistema” (a largo plazo).}$$

Entonces, L es una medida de efectividad del sistema o cola y se obtiene así (denotando por $(\ )'$ la derivada respecto de ρ y por $E(\ )$ la esperanza, o valor medio, de una variable aleatoria[9]):

$$\begin{aligned} L = E(N) &= \sum_{n\geq 0} n\, p_n = (1-\rho) \sum_{n\geq 1} n\, \rho^n = (1-\rho)\,\rho \sum_{n\geq 1} n\, \rho^{n-1} \\ &= (1-\rho)\,\rho \sum_{n\geq 1} (\rho^n)' = (1-\rho)\,\rho \left(\sum_{n\geq 1} \rho^n \right)' = (1-\rho)\,\rho \left(\frac{1}{1-\rho} \right)' \\ &= (1-\rho)\,\rho\, \frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho}\,, \end{aligned}$$

Es decir,

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

Notemos que cuando $\rho \to 1$ (es decir, cuando el tráfico tiende a ser “intenso”), el número medio de clientes en el sistema L tiende a $+\infty$ lo que parece bastante natural.

[9] Si Y es una variable aleatoria discreta (esto es, que toma valores en conjunto S finito o numerable) la esperanza de Y se define como:

$$E(Y) = \sum_{y\in S} y\, P(Y = y)$$

Si Y es una variable aleatoria (absolutamente) continua con función de densidad f, se define la esperanza de Y como:

$$E(Y) = \int_{-\infty}^{\infty} y\, f(y)\, dy$$

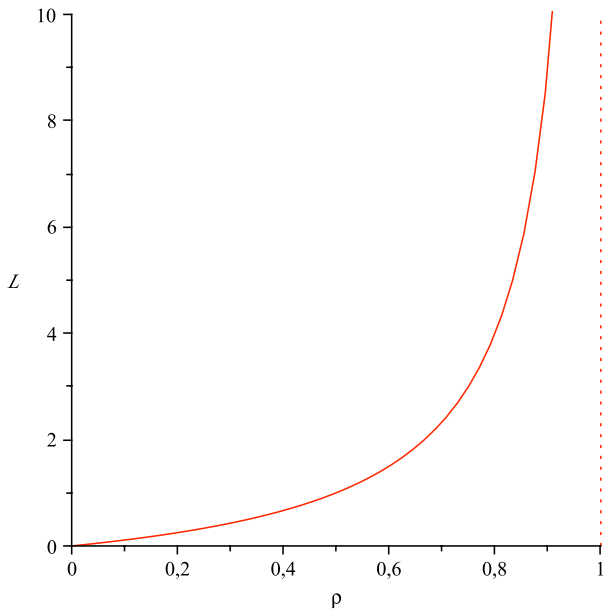


Figura 7: Gràfica de la función $L = \dfrac{\rho}{1-\rho}$ para $\rho \in (0,1)$

También podemos definir la variable aleatoria

$N_q =$ "número de clientes haciendo cola" (a largo plazo).

Entonces,

$$N_q = \begin{cases} N-1 & \text{si } N \geq 1 \\ 0 & \text{si } N = 0 \end{cases}$$

Así, también introducimos

$L_q = E(N_q) =$ "número medio de clientes haciendo cola" (a largo plazo).

Por tanto, resulta que

$$L_q = \sum_{n\geq 1}(n-1)\,p_n = \sum_{n\geq 1} n\,p_n - \sum_{n\geq 1} p_n = L - (1-p_0)$$
$$= \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}$$

Esto es,

$$\boxed{L_q = L - (1 - p_0) = \frac{\rho^2}{1 - \rho}}$$

(notemos que la relación $L_q = L - (1 - p_0)$ es cierta en general, es decir, no depende de la suposición de que los tiempos entre llegadas y los tiempos de servicio sean exponenciales).

También podemos definir la probabilidad de que haya al menos n clientes en el sistema a largo plazo, que es

$$\begin{aligned} P(N \geq n) &= \sum_{k \geq n} p_k = \sum_{k \geq n} (1 - \rho)\, \rho^k = (1 - \rho)\, \rho^n \sum_{k \geq n} \rho^{k-n} \\ &= (1 - \rho)\, \rho^n \sum_{\ell \geq 0} \rho^\ell = \frac{(1 - \rho)\, \rho^n}{1 - \rho} = \rho^n \end{aligned}$$

para todo $n \geq 0$. Es decir,

$$\boxed{P(N \geq n) = \rho^n}$$

En particular, con $n = 1$ tenemos que

$$P(N \geq 1) = \rho$$

es la proporción de tiempo, a largo plazo, que en el sistema habrá al menos un cliente (es decir, que el único servidor estará ocupado). Por tanto, $\rho \times 100\,\%$ es el porcentaje de ocupación del servidor.

Aunque se podrían introducir otras medidas de efectividad, la última que consideraremos es el valor numérico W definido así:

$W =$ "tiempo medio de estancia en el sistema"
(=tiempo medio de espera en cola
+ tiempo medio de servicio (que siempre es $1/\mu$)).

Para poder calcular su valor hemos de introducir la variable aleatoria

$\tau =$ "tiempo de estancia en el sistema" (de un cliente, a largo plazo).

Resulta interesante observar, como veremos, que τ *también* es una variable exponencial, concretamente, τ es una exponencial de parámetro $\mu - \lambda$. Entonces,

$$\boxed{W = E(\tau) = \frac{1}{\mu - \lambda}}$$

Además, si denotamos por W_q el tiempo medio de espera en cola, tendremos que

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

En efecto, τ es una exponencial de parámetro $(\mu - \lambda)$ puesto que su función generatriz de momentos[10] es:

$$E\left(e^{t\,\tau}\right) = \begin{cases} \dfrac{\alpha}{\alpha - t} & \text{si } t < \alpha \\ +\infty & \text{en caso contrario} \end{cases} \tag{6}$$

con $\alpha = \mu - \lambda\,(> 0$ en el caso $\rho < 1)$. La expresión (6) caracteriza a las variables aleatorias exponenciales de parámetro α (como se puede ver en el Apéndice), y su justificación es como sigue:

Demostración de (6): Por la *Fórmula de la Esperanza Total*, análoga a la Fórmula de la Probabilidad Total pero con esperanzas condicionadas[11] en vez de probabilidades condicionadas,

$$E\left(e^{t\,\tau}\right) = \sum_{k\geq 0} E\left(e^{t\,\tau} \,/\, N = k\right) P(N = k) \tag{7}$$

donde N es el número de clientes en el sistema a largo plazo (luego $P(N = k) = p_k = (1 - \rho)\,\rho^k$). Utilizamos esta fórmula ya que podemos calcular $E\left(e^{t\,\tau} \,/\, N = k\right)$ teniendo en cuenta que condicionado a que haya k clientes

[10] véase el Apéndice para la definición y propiedades básicas de esta función.

[11] Para el caso discreto la definición de esperanza condicionada es la siguiente: si Y toma valores en S, finito o numerable, se define la esperanza de Y condicionada al suceso $\{N = k\}$ como

$$E(Y \,/\, N = k) = \sum_{y \in S} y\, P(Y = y \,/\, N = k)$$

Para el caso continuo, que es el de la fórmula, la definición es similar pero más complicada y por ello no se explicita.

en el sistema, la variable τ, que es el tiempo que pasa en el sistema un cliente que llegue a él, será la suma de los tiempos de servicio de los k clientes más el del que llega. Denotemos por S_i, con $i = 1, \ldots, k, k+1$ estos tiempos de servicios (son variables aleatorias independientes e idénticamente distribuidas, exponenciales de parámetro μ), así que

$$E\left(e^{t\,\tau} \,/\, N = k\right) = E\left(e^{t\,(S_1+\cdots+S_k+S_{k+1})} \,/\, N = k\right) = E\left(e^{t\,(S_1+\cdots+S_k+S_{k+1})}\right)$$

(la última igualdad es debida a la independencia entre las S_i y la variable N). Además, por las propiedades de la exponencial tenemos que

$$e^{t(S_1+\cdots+S_k+S_{k+1})} = \prod_{i=1}^{k+1} e^{tS_i}$$

y debido a la independencia entre las S_i, tenemos que

$$\begin{aligned} E\left(\prod_{i=1}^{k+1} e^{tS_i}\right) = \prod_{i=1}^{k+1} E\left(e^{t\,S_i}\right) = \prod_{i=1}^{k+1} \int_0^{+\infty} e^{t\,y}\,\mu\,e^{-\mu\,y}\,dy \\ = \begin{cases} \left(\dfrac{\mu}{\mu-t}\right)^{k+1} & \text{si } t < \mu \\ +\infty & \text{si } t \geq \mu \end{cases} \end{aligned}$$

Sustituyendo en (7),

$$E\left(e^{t\,\tau}\right) = \begin{cases} \displaystyle\sum_{k\geq 0} \left(\frac{\mu}{\mu-t}\right)^{k+1} (1-\rho)\,\rho^k & \text{si } t < \mu \\ +\infty & \text{si } t \geq \mu \end{cases}$$

Para acabar la demostración de (6) vamos a simplificar la suma de la serie que aparece en la expresión anterior usando que $\frac{\mu}{\mu-t}\,\rho = \frac{\mu}{\mu-t}\,\frac{\lambda}{\mu} = \frac{\lambda}{\mu-t}$. En

efecto,

$$\sum_{k\geq 0}\left(\frac{\mu}{\mu-t}\right)^{k+1}(1-\rho)\,\rho^k = (1-\rho)\,\frac{\mu}{\mu-t}\sum_{k\geq 0}\left(\frac{\mu}{\mu-t}\right)^k\rho^k$$

$$= (1-\rho)\,\frac{\mu}{\mu-t}\sum_{k\geq 0}\left(\frac{\lambda}{\mu-t}\right)^k$$

$$= \begin{cases} \left(1-\dfrac{\lambda}{\mu}\right)\dfrac{\mu}{\mu-t}\,\dfrac{1}{1-\dfrac{\lambda}{\mu-t}} = \dfrac{\mu-\lambda}{\mu-\lambda-t} & (\text{ si } \dfrac{\lambda}{\mu-t}<1) \\ +\infty & (\text{en caso contrario}) \end{cases}$$

Con esto acabamos la demostración de (6). □

Vemos que existe una relación entre L y W, que es:

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

Esta relación, que es cierta para modelos más generales que los considerados aquí, se conoce como *fórmula de Little* y se suele expresar de esta manera:

Fórmula de Little: $L = \lambda W$

3.3. Un ejemplo

Vamos a ver con un ejemplo muy sencillo cómo el uso del modelo de cola M/M/1 y de las medidas de efectividad que acabamos de introducir puede ayudarnos a la toma de decisiones.

Ana Poodle es una peluquera canina que tiene una pequeña peluquería en un barrio de la ciudad. Ella es la única peluquera que trabaja en su negocio, pero los sábados por la tarde tiene mucho trabajo y está considerando la posibilidad de ampliarlo (tomar algún ayudante y/o hacer reformas para mejorar la sala de espera).

Como su principal problema radica en los sábados por la tarde, vamos a restringirnos a considerar solo esos días, pero antes de poder realizar ningún

estudio hemos de tomar observaciones. Ana recoge información durante unos cuantos sábados por la tarde y llega a las siguientes conclusiones:

(a) Los clientes llegan los sábados por la tarde de manera independiente y a razón de unos 2 por hora.

(b) Ella tarda un promedio de 20 minutos por cliente, y los tiempos de servicios de los clientes son independientes entre sí y no dependen tampoco ni de la hora ni del trabajo que tenga acumulado (el éxito de su negocio radica en servir a los clientes siempre bien, aunque tenga otros esperando). Naturalmente, los atiende por estricto orden de llegada.

(c) Como tiene muy buena fama como peluquera y trata muy bien a los clientes, éstos esperan en cola a ser servidos todo el tiempo que haga falta (es decir, no se van aunque tengan que esperar). La sala de espera es pequeña y sólo tiene espacio para dos clientes; cuando hay más clientes esperando, lo hacen en el parque que hay frente a su negocio.

A partir de esta información suministrada por Ana vemos que el modelo M/M/1 se adapta bien a su negocio y que los parámetros del modelo se pueden estimar por:

$$\lambda = 2 \text{ clientes/hora}$$

$$\mu = \frac{1}{20} \text{ clientes/minuto} = 3 \text{ clientes/hora} \Rightarrow \rho = \frac{\lambda}{\mu} = \frac{2}{3} < 1$$

Como medidas de efectividad podemos calcular:

- El número esperado de clientes en la peluquería a largo plazo un sábado por la tarde:

$$L = \frac{\rho}{1-\rho} = \frac{2/3}{1-2/3} = 2 \text{ clientes.}$$

- El número esperado de clientes haciendo cola en la peluquería a largo plazo un sábado por la tarde:

$$L_q = \frac{\rho^2}{1-\rho} = \frac{(2/3)^2}{1-2/3} = \frac{4}{3} \text{ clientes.}$$

- El porcentaje de ocupación a largo plazo un sábado por la tarde es:

$$\rho \times 100\,\% = 66.\widehat{6}\,\%\,,$$

es decir, el $66.\widehat{6}\,\%$ del tiempo hay algún cliente en la peluquería, por lo que si llega un nuevo cliente tendrá que esperar forzosamente, y el $33.\widehat{3}\,\%$ restante no habrá ninguno y si llega un nuevo cliente pasa a ser atendido sin tener que esperar.

- El tiempo medio de estancia en la peluquería un sábado por la tarde para un cliente es:

$$W = \frac{1}{\mu - \lambda} = \frac{1}{3-2} = 1 \text{ hora}\,,$$

es decir, en promedio cada cliente se pasa 40 minutos esperando a ser servido (esto es W_q), y 20 minutos más siendo atendido por la peluquera.

- La probabilidad de que haya al menos 2 clientes esperando (y si llega uno nuevo tenga que esperar en el parque frente a la peluquería) es:

$$P(N \geq 3) = \rho^3 = \left(\frac{2}{3}\right)^3 \cong 0{,}3$$

esto es, un 30 % del tiempo la sala de espera está llena y si llega un nuevo cliente habrá de esperar fuera.

A la vista de estas medidas sobre la efectividad del servicio ofrecido por Ana, ésta deberá decidir si tomar un ayudante y/o hacer reformas en la sala de espera para ampliarla y reducir las esperas de clientes fuera. Para tomar esta decisión deberá tener en cuenta los gastos asociados a la contratación de un ayudante y a la realización de las reformas, así como valorar el perjuicio que le suponen para su negocio las esperas de los clientes.

4. La cola M/M/c (múltiples servidores)

Supongamos ahora que no tenemos uno sino $c > 1$ servidores idénticos que dan servicio a los clientes que llegan al sistema y que si han de esperar

forman una única cola que es atendida por todos los servidores (véase la figura 1).

Por lo demás, el modelo de cola M/M/c es análogo al de un sólo servidor M/M/1, esto es, las llegadas al sistema se siguen produciendo según un proceso de Poisson de intensidad $\lambda > 0$, y cada uno de los c servidores idénticos tiene una distribución del tiempo de servicio que es exponencial de parámetro $\mu > 0$, todas ellas independientes. Por tanto, se trata de nuevo de un Proceso de Nacimiento y Muerte con espacio de estados $E = \mathbb{N} \cup \{0\}$, con tasa de llegadas constante $\lambda_n = \lambda$ para todo $n \geq 0$, y tasa de servicio:

$$\mu_n = \begin{cases} n\,\mu & \text{si } 1 \leq n \leq c \\ c\,\mu & \text{si } n > c \end{cases}$$

(esto es así porque cada servidor tiene tasa de servicio μ, luego la tasa de servicio del sistema completo será la tasa combinada de los servidores que estén ocupados, que es el número de dichos servidores multiplicado por μ).

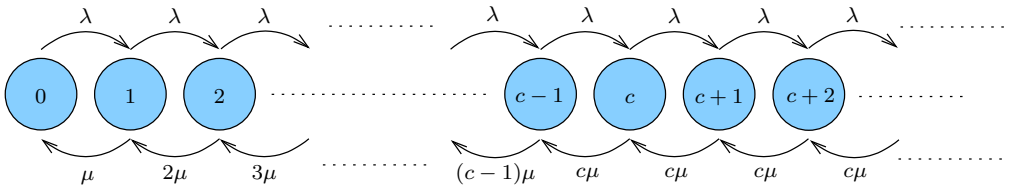


Figura 8: Tasas para la cola M/M/c

4.1. La distribución límite

Utilizando que

$$\sum_{n\geq 1} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} = \sum_{n=1}^{c-1} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} + \sum_{n\geq c} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{c!\, c^{n-c}}$$

tendremos por (4) que existe distribución límite si y sólo si la siguiente serie es convergente:

$$\sum_{n\geq c} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{c!\, c^{n-c}} = \frac{c^c}{c!} \sum_{n\geq c} \left(\frac{\lambda}{c\,\mu}\right)^n,$$

y esta serie converge si y sólo si

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

(serie geométrica de razón $\dfrac{\lambda}{c\,\mu}$). Para este modelo es habitual utilizar la letra ρ para denotar la razón de la serie geométrica, es decir,

$$\rho = \frac{\lambda}{c\,\mu}\,, \quad \text{y por ello también se define } r = \frac{\lambda}{\mu}\,,$$
$$\text{con lo que podemos expresar } \rho = \frac{r}{c}\,.$$

En el caso particular de un servidor ($c = 1$), $\rho = r$. Con estas notaciones, tenemos que la serie converge $\Leftrightarrow \rho < 1$ y en ese caso,

$$\sum_{n\geq 1}\prod_{i=1}^{n}\frac{\lambda_{i-1}}{\mu_i} = \sum_{n=1}^{c-1}\frac{r^n}{n!} + \frac{c^c}{c!}\sum_{n\geq c}\rho^n = \sum_{n=1}^{c-1}\frac{r^n}{n!} + \frac{c^c}{c!}\,\rho^c\sum_{\ell\geq 0}\rho^\ell$$
$$= \sum_{n=1}^{c-1}\frac{r^n}{n!} + \frac{c^c}{c!}\,\rho^c\,\frac{1}{1-\rho} = \sum_{n=1}^{c-1}\frac{r^n}{n!} + \frac{r^c}{c!\,(1-\rho)}\,.$$

A partir de la expresión (3) para la distribución límite, utilizando lo anterior tenemos que

$$p_0 = \left(\sum_{n=0}^{c-1}\frac{r^n}{n!} + \frac{r^c}{c!\,(1-\rho)}\right)^{-1}$$

$$\text{y} \quad p_n = \begin{cases} p_0\,r^n/n! & \text{si } 1 \leq n \leq c-1 \\ p_0\,\rho^n\,c^c/c! & \text{si } n \geq c \end{cases}, \ \text{si } \rho = \frac{\lambda}{c\,\mu} < 1$$

y utilizando esta distribución límite podríamos obtener medidas de efectividad del sistema a largo plazo análogamente a lo que vimos para la cola M/M/1. De todas esta medidas sólo veremos una, sin embargo, la conocida como *fórmula Erlang C*, que no es más que la probabilidad de que un cliente que llega al sistema tenga que esperar.

4.2. La fórmula (de probabilidad de espera) Erlang C

En primer lugar vemos la expresión para la probabilidad de que si llega un cliente al sistema **no** tenga que esperar que, si N denota el número de clientes en el sistema "a largo plazo", es:

$$P(N \leq c-1) = \sum_{n=0}^{c-1} p_n = p_0 \sum_{n=0}^{c-1} \frac{r^n}{n!}\,.$$

Utilizando que

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!\,(1-\rho)} \right)^{-1},$$

podemos despejar de esta expresión el término $\displaystyle\sum_{n=0}^{c-1} \frac{r^n}{n!}$ obteniendo que

$$\sum_{n=0}^{c-1} \frac{r^n}{n!} = p_0^{-1} - \frac{r^c}{c!\,(1-\rho)}\,.$$

Si sustituimos esta expresión en la probabilidad de no tener que esperar, tenemos que

$$P(N \leq c-1) = p_0 \left(p_0^{-1} - \frac{r^c}{c!\,(1-\rho)} \right) = 1 - p_0 \frac{r^c}{c!\,(1-\rho)}\,.$$

De esta manera, la conocida **función Erlang C**, que es *la probabilidad de que al llegar un cliente al sistema* **sí** *tenga que esperar*, es 1 menos la probabilidad anterior, esto es,

Probabilidad de espera: $$C(c,r) = p_0 \frac{r^c}{c!\,(1-\rho)} = \frac{\frac{r^c}{c!\,(1-\rho)}}{\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!\,(1-\rho)}}$$

(la notación habitual, que es la que hemos introducido, expresa esta probabilidad como función de c y de $r = \lambda/\mu$). En la práctica el cómputo del valor de esta función presenta problemas debido a los factoriales; por ello se suele calcular no utilizando la expresión anterior sino la dada por la fórmula (8) que aparece en la sección 5.2.

5. La cola M/M/c/c y la función de pérdida Erlang B

El modelo de colas denotado por **M/M/c/c** es análogo al M/M/c salvo por el hecho de que ahora el número máximo de clientes permitidos en el sistema es c (esto es lo que indica el segundo c en la notación de Kendall que utilizamos). Es decir, que tenemos como antes c servidores instalados en una estación de trabajo, y clientes que llegan para ser servidos. Si cuando un cliente llega algún servidor está libre, pasa a ocuparlo y cuando finaliza deja el sistema. Si cuando llega, todos los servidores están ocupados, se va (se produce una "pérdida"). En este modelo no se permite que se queden clientes en espera en el sistema.

El espacio de estados es ahora $E = \{0, 1, \ldots, c\}$. Los clientes van llegando según un proceso de Poisson de intensidad $\lambda > 0$, pero en realidad al sistema no llegan con esa tasa, ya que cuando todos los servidores están ocupados, no llegan a entrar en el sistema sino que se van. En cambio, los tiempos de servicio de los clientes siguen siendo, como en el modelo M/M/c, variables aleatorias independientes e idénticamente distribuidas, con ley exponencial de parámetro $\mu > 0$ para cada uno de los servidores, independientes entre sí. Por tanto, se trata de un *Proceso de Nacimiento y Muerte* con tasas de nacimiento y de defunción:

$$\lambda_n = \lambda \quad \text{si } 0 \leq n \leq c-1 \quad \text{y} \quad \mu_n = n\,\mu \quad \text{si } 1 \leq n \leq c\,.$$

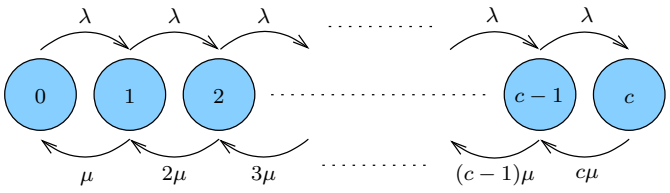


Figura 9: Tasas para la cola M/M/c/c

5.1. La distribución límite

Las ecuaciones de balance vienen dadas por la expresión (2) con $M = c$ y las λ_n y las μ_n son como acabamos de comentar. La solución de las ecuaciones

de balance es la dada por la expresión (5) con $M = c$:

$$p_0 = \left(1 + \sum_{n=1}^{c} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right)^{-1} \quad \text{y} \quad p_n = p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}, \quad 1 \leq n \leq c\,.$$

Teniendo en cuenta que en este caso

$$\prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} = \frac{\lambda^n}{\mu^n} \frac{1}{n!} = \frac{r^n}{n!}\,,$$

tenemos que

$$p_0 = \left(\sum_{\ell=0}^{c} \frac{r^\ell}{\ell!} \right)^{-1} \quad \text{y} \quad p_n = \frac{\frac{r^n}{n!}}{\sum_{\ell=0}^{c} \frac{r^\ell}{\ell!}} \quad \text{para } 1 \leq n \leq c,$$

es decir,

$$p_n = \frac{\frac{r^n}{n!}}{\sum_{\ell=0}^{c} \frac{r^\ell}{\ell!}} \quad \text{para } 0 \leq n \leq c$$

5.2. La fórmula (función de pérdida) Erlang B

Para este modelo, la medida de efectividad que nos interesa básicamente es lo que se conoce como *función de pérdida de Erlang* o *fórmula Erlang-B* (la "B" viene de "blocked", ya que las llegadas que no pueden ser atendidas inmediatamente son bloqueadas), que no es más que la probabilidad de que al llegar un nuevo cliente encuentre todos los servidores ocupados (y, por tanto, se produzca una pérdida):

Probabilidad de pérdida: $$B(c, r) = p_c = \frac{\frac{r^c}{c!}}{\sum_{\ell=0}^{c} \frac{r^\ell}{\ell!}}$$

(aquí también la notación habitual expresa la probabilidad como función de c y de $r = \lambda/\mu$).

Como la expresión de $B(c,r)$ causa problemas de computación debido a los factoriales (que son números muy altos si c es relativamente grande), se suele usar esta otra fórmula iterativa que resulta inmediato comprobar:

$$B(c=0,r)=1\,,\ B(c,r)=\frac{r\,B(c-1,r)}{c+r\,B(c-1,r)}\ \text{si}\ c\geq 1$$

Nota: Además, como el cálculo de la función Erlang C también tiene problemas de computación, para obtenerla primero se calcula la fórmula Erlang B de forma iterativa, y después se utiliza esta fórmula que relaciona ambas:

$$C(c,r)=\frac{c\,B(c,r)}{c-r+r\,B(c,r)} \tag{8}$$

5.3. Un ejemplo

Volviendo al ejemplo de la peluquería canina de la sección anterior, imaginemos ahora que muy cerca de la peluquería de Ana acaban de inaugurar un nuevo centro comercial en el que hay otra peluquería que le hace la competencia. Sus clientes son bastante fieles, así que cuando lo necesitan van al local de Ana en primer lugar, pero si no pueden ser atendidos inmediatamente, en vez de esperar como hacían antes, deciden ir a probar en la peluquería del centro comercial, que es mucho mayor que la de Ana, y allí se quedan.

A la vista de estas muy malas noticias para su negocio, Ana está decidida a contratar uno o varios ayudantes, todos los que necesite para asegurar que un sábado por la tarde, el día que tiene más trabajo, la probabilidad de que un cliente llegue y no pueda ser atendido por ninguno de los peluqueros no supere el 5 % (es decir, 0,05).

Tenemos que el valor de $r=\lambda/\mu$ es 2/3. Veamos cómo podemos ayudar a Ana diciéndole cuantos ayudantes debería contratar: hemos de encontrar

el primer valor de c en la fórmula iterativa o de recurrencia, que hace que $B(c,r) \leq 0{,}05$.

$$B(0,r) = 1$$

$$B(1,r) = \frac{r \times 1}{1 + r \times 1} = \frac{2/3}{1 + (2/3)} = \frac{2}{5} = 0{,}4 \ (> 0{,}05)$$

$$B(2,r) = \frac{(2/3) \times (2/5)}{2 + (2/3) \times (2/5)} = \frac{2}{17} = 0{,}117647\ldots \ (> 0{,}05)$$

$$B(3,r) = \frac{(2/3) \times (2/17)}{3 + (2/3) \times (2/17)} = \frac{4}{157} = 0{,}025477\ldots \ (\leq 0{,}05)$$

Por tanto, como en total deberá haber al menos 3 servidores o peluqueros, Ana debería contratar a 2 ayudantes para satisfacer su requerimiento (de hecho, con 2 ayudantes la probabilidad de pérdida es, en realidad, sensiblemente inferior a lo requerido, ya que es aproximadamente la mitad, el 2,5 %).

Si el centro comercial se ve obligado a cerrar y Ana recupera la "fidelidad" de sus clientes (es decir, éstos vuelven a esperar el tiempo necesario hasta ser atendidos por Ana o sus ayudantes), nos preguntamos cómo habrá mejorado su negocio al contratar los dos ayudantes, en cuanto a la disminución en la probabilidad que tiene un cliente que llega a la peluquería un sábado por la tarde de tener que esperar.

Cuando Ana trabajaba sola vimos que esta probabilidad de tener que esperar era de $0.\widehat{6}$, es decir, un $66.\widehat{6}\,\%$ (sección 3.3). Ahora, con un total de $c = 3$ servidores, la probabilidad de tener que esperar se reduce hasta $C(c = 3, r = 2/3)$. Para calcular este valor usamos la fórmula (8):

$$C(c=3,r) = \frac{3\,B(c=3,r)}{3 - \frac{2}{3} + \frac{2}{3}\,B(c=3,r)} = \frac{3\,\frac{4}{157}}{3 - \frac{2}{3} + \frac{2}{3}\,\frac{4}{157}} = \frac{36}{1107} = 0{,}03252\ldots$$

esto es, aproximadamente un 3,25 %, frente a un $66.\widehat{6}\,\%$!!!

Apéndice: la función generatriz de momentos

Definición: Sea X una variable aleatoria real cualquiera. La **función generatriz de momentos** de X es la función $\psi_X : \mathbb{R} \longrightarrow [0, +\infty]$, definida

por

$$\psi_X(t) = E(e^{t\,X})\,.$$

Notemos que $\psi_X(0) = 1 < +\infty$ para toda X (aunque éste podría ser el único valor de t para el que $\psi_X(t)$ sea finita).

Ejemplo: una variable Exponencial.

Sea X una variable exponencial de parámetro λ, con $\lambda > 0$, es decir, X es una variable aleatoria con función de densidad dada por

$$f(x) = \begin{cases} \lambda\, e^{-\lambda\, x} & \text{si} \quad x > 0 \\ 0 & \text{en caso contrario} \end{cases}$$

Entonces, su función generatriz de momentos es finita para todo $t < \lambda$, y su expresión es:

$$\psi_X(t) = E(e^{t\,X}) = \int_0^{+\infty} e^{t\,x}\,\lambda\, e^{-\lambda\, x}\,dx = \lambda \int_0^{+\infty} e^{x\,(t-\lambda)}\,dx$$

$$= \begin{cases} \dfrac{\lambda}{\lambda - t} & \text{si } t < \lambda \\ +\infty & \text{en caso contrario.} \end{cases}$$

El siguiente resultado (que no demostraremos) es bien conocido y nos dice que la función generatriz de momentos caracteriza la distribución, siempre que esté definida (sea finita) en un entorno del cero:

Teorema A1: *(caracterización de la distribución)*

Si X e Y son dos variables aleatorias reales que tienen funciones generatrices de momentos finitas en un entorno del cero y que coinciden en dicho entorno, entonces X e Y tienen la misma distribución.

Veamos (también sin demostrar) otros dos resultados importantes sobre esta función:

Teorema A2: *(origen del nombre)*

Si X es una variable aleatoria real con función generatriz de momentos ψ_X finita en un entorno del cero, entonces X tiene momentos de todos los órdenes (es decir, existe y es finita $E(X^k)$ para todo $k \geq 1$), y se obtienen así:

$$E\,(X^k) = \psi_X^{(k)}(0)$$

donde $\psi^{(k)}(0)$ indica la derivada de orden k de la función ψ_X evaluada en el cero.

Teorema A3: *(relación con la independencia)*

Si X e Y son dos variables aleatorias reales independientes, con funciones generatrices de momentos respectivas ψ_X y ψ_Y, se cumple que:

para todo t tal que $\psi_X(t)$ y $\psi_Y(t)$ sean finitas, existirá $\psi_{X+Y}(t)$ y se cumplirá que

$$\psi_{X+Y}(t) = \psi_X(t)\,\psi_Y(t)$$

Bibliografía

[1] Brockmeyer, E.; Halstrom, H. L.; Jensen, A. *"The life and works of A. K. Erlang"*, The Copenhagen Telephone Company. Transactions of the Danish Academy of Technical Sciences, nº 2, 1848.

http://oldwww.com.dtu.dk/teletraffic/Erlang.htm

[2] Gross, D.; Shortle, J. F., Thompson, J. M.; Harris, C. M. *"Fundamentals of Queueing Theory"* (fourth edition), Wiley Series in Probability and Statistics, 2008.

[3] Kleinrock, L. *"Queueing Systems, Volume I: Theory*, Wiley-Interscience, 1975.

[4] Kleinrock, L.; Gail, R. *"Queueing Systems: Problems and Solutions*, Wiley-Interscience, 1996.

[5] Medhi, J. *"Stochastic Models in Queing Theory"*, second edition, Academic Press, 2003.

[6] Sobre A. K. Erlang en Wikipedia:

http://en.wikipedia.org/wiki/Agner_Krarup_Erlang

Departament de matemàtiques
Universitat Autònoma de Barcelona
delgado@mat.uab.cat
http://www.mat.uab.cat/~delgado

Publicat el 29 de setembre de 2009