Sistemas de Espera: Probabilidades Estacionarias y Estrategias de Equilibrio de Nash

Tesis que presenta

Tania Sarahi Rivera Pérez

Para obtener el grado de

Maestra en Ciencias

(Matemáticas)

Asesor de Tesis: Dr. Raúl Montes de Oca Machorro

Sinodales

Presidenta:	Dra. Patricia Saavedra Barrera	UAM-I
Secretario:	Dr. Raúl Montes de Oca Machorro	UAM-I
Vocal:	Dr. Hugo Adán Cruz Suárez	BUAP

División de Ciencias Básicas e Ingeniería

Departamento de Matemáticas

México, D.F., a 10 de diciembre del 2012.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00086
Matrícula: 210383004

SISTEMAS DE ESPERA: PROBABILIDADES ESTACIONARIAS Y ESTRATEGIAS DE EQUILIBRIO DE NASH

En México, D.F., se presentaron a las 13:00 horas del día 10 del mes de diciembre del año 2012 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DRA. PATRICIA SAAVEDRA BARRERA
DR. HUGO ADAN CRUZ SUAREZ
DR. JOSE RAUL MONTES DE OCA MACHORRO

TANIA SARAHI RIVERA PEREZ
ALUMNA

Bajo la Presidencia de la primera y con carácter de Secretario el último, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRA EN CIENCIAS (MATEMÁTICAS)

DE: TANIA SARAHI RIVERA PEREZ

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

Aprobar

Acto continuo, la presidenta del jurado comunicó a la interesada el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

REVISÓ

LIC. JULIO CESAR DE LARA ISASSI
DIRECTOR DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JOSE ANTONIO DE LOS REYES HEREDIA

PRESIDENTA

Patricia Saavedra

DRA. PATRICIA SAAVEDRA BARRERA

VOCAL

DR. HUGO ADAN CRUZ SUAREZ

SECRETARIO

J. Raúl Montes de Oca M.

DR. JOSE RAUL MONTES DE OCA MACHORRO

Universidad Autónoma Metropolitana
Unidad Iztapalapa

Casa abierta al tiempo

División de Ciencias Básicas e Ingeniería

Departamento de Matemáticas

Sistemas de Espera: Probabilidades Estacionarias y Estrategias de Equilibrio de Nash

T E S I S
DE MAESTRÍA EN CIENCIAS
(MATEMÁTICAS):

P R E S E N T A:

Tania Sarahi Rivera Pérez

Asesor de Tesis: Dr. Raúl Montes de Oca

México, D.F., a 10 de diciembre del 2012.

Agradecimientos

Primeramente agradezco al Dr. Raúl Montes de Oca Machorro por el apoyo que me brindó como asesor de tesis, por su disponibilidad, confianza y porque no, su amistad.

Asimismo agradezco a la Dra. Patricia Saavedra y el Dr. Hugo Adán Suárez por haber revisado y proporcionado valiosos comentarios para la mejora de esta tesis.

A mi familia por brindarme siempre su fortaleza, cariño y confianza.

También agradezco a mi novio Jorge Daniel por su apoyo y consejos.

Finalmente agradezco a CONACyT por el apoyo económico otorgado durante mi estancia.

México, D.F., Diciembre de 2012.
Tania Sarahi Rivera Pérez

Dedicatoria

A mis queridos padres, gracias por protegerme, guiarme, darme siempre la fortaleza para seguir adelante y principalmente gracias por su paciencia en el tiempo que no pude estar con ustedes. A ustedes les debo lo que soy!

Consulta al Señor en todos tus hechos,
y él te dirigirá para bien; sí, cuando te
acuestes por la noche, acuéstate en el
Señor, para que él te cuide en tu sueño;
y cuando te levantes por la mañana,
rebose tu corazón de gratitud a Dios;
y si haces estas cosas, serás enaltecido
en el postrer día.

Índice general

1. Introducción **1**

1.1. Colas simples . . . 1

1.2. Juegos y colas . . . 1

1.3. Motivación y antecedentes a colas paralelas . . . 2

1.4. Objetivos de tesis . . . 3

1.5. Contenido de la tesis . . . 4

1.6. Estructura de la tesis . . . 4

2. Preliminares **7**

2.1. Cadena de Markov a tiempo continuo . . . 7

2.1.1. Procesos de nacimiento y muerte . . . 8

2.1.2. Distribución estacionaria para el PNM . . . 10

2.2. Modelo de colas paralelas basados en el proceso de nacimiento y muerte . . . 11

2.3. Estabilidad en colas paralelas . . . 12

3. Equilibrios de Nash para los sistemas de Colas FCFS con tasa de servicio creciente. **15**

3.1. Introducción . . . 15

3.2. El modelo . . . 16

3.3. Variables aleatorias y procesos estocásticos asociados al modelo . . . 17

3.3.1. Llegadas a Qs . . . 19

3.3.2. Servicio en Qs . . . 19

3.4. Monotonicidad con respecto al tamaño de la cola al entrar . . . 19

3.5. Monotonicidad y continuidad con respecto a políticas umbrales . . . 24

3.6. Estructura y existencia del equilibrio de Nash . . . 31

3.7. Conclusiones . . . 35

4. Cálculo de probabilidades estacionarias para el modelo de dos colas en paralelo **37**

4.1. Introducción . . . 37

4.2. Dos colas en paralelo con "salto" instantáneo y "salto" umbral . . . 38

4.2.1. $N = 2$, $\mu_1 = \mu_2 = \mu$ . . . 38

4.2.2. N arbitraria, $\mu_1 \neq \mu_2$ . . . 43

4.3. Dos colas en paralelo con "salto" y capacidad restringida a L . . . 54

4.3.1. Descripción del modelo . . . 54

4.3.2. Resultados teóricos . . . 58

4.3.3. Casos particulares . . . 63

4.4. Conclusiones . . . 64

5. Valor de la información en un sistema de dos colas con "salto" umbral. **65**

5.1. Introducción . . . 65

5.2. La distribución estacionaria . . . 66

5.2.1. Descripción del modelo . . . 67

5.2.2. Cálculo de la matriz de tasas R . . . 72

5.2.3. Vectores de probabilidades límite . . . 73

5.3. El tiempo de espera previsto . . . 74

5.4. El valor de la información y estrategias de equilibrio de Nash . . . 77

5.5. Las externalidades de comprar información para $N = 3$ . . . 77

5.6. Conclusiones . . . 81

6. Conclusiones y Perspectivas **83**

A. Teoría de colas **85**

A.1. Descripción de un sistema de espera . . . 85

A.2. Tiempo de servicio residual . . . 88

A.3. Notación Kendall . . . 89

B. Teoría de juegos **91**

C. Descomposición espectral de una matriz real **93**

Bibliografía **95**

CAPÍTULO 1

Introducción

1.1. Colas simples

La teoría de colas se inició con el matemático danés Agner Krarup Erlang, trabajador de la compañía telefónica estatal de Dinamarca (a principios del siglo XX); Erlang publicó el primer artículo sobre la teoría de colas en 1909. Específicamente se preocupó del estudio para problemas de dimensionamiento de líneas y centrales de conmutación telefónica para el servicio de llamadas.

La teoría de colas se ve principalmente como una rama de la teoría de la probabilidad aplicada. Sus aplicaciones se encuentran en diferentes campos, por ejemplo, redes de comunicación, sistemas informáticos, industrias maquiladoras, etc. Una cola se produce cuando la demanda de un servicio por parte de los clientes excede la capacidad de éste. Debemos hacer cola y esperar para recibir servicio médico, para comprar una estampilla en el correo o para cobrar un cheque en el banco. Se forma una cola en un centro de computación cuando el número de trabajos a procesar excede la capacidad del sistema de cómputo. En los aeropuertos se forman colas en las ventanillas de recepción de documentos y equipaje. También los aviones tienen que esperar hasta que haya una pista libre y puedan despegar (o aterrizar). En todas estas situaciones y muchas más, se tienen ciertas características, como son la llegada de clientes, el tiempo requerido de servicio, y otras más, que se pueden representar en forma natural como procesos estocásticos y aplicarlos al análisis de un sistema de espera [19], [31].

1.2. Juegos y colas

Por otro lado, existen ejemplos de problemas de decisión que surgen en las colas, y cuya solución requiere de modelos de la teoría de juegos [16], [17]. Algunas preguntas que surgen de la relación entre colas y juegos son:

- ¿Cuándo unirse a una cola? Las personas que tienen que hacer cola para recibir el servicio a menudo se enfrentan a este problema. Pueden tener una idea acerca de la demanda, es decir la distribución de probabilidad del número total de personas que llegan y compiten por el mismo servicio, pero las decisiones de los tiempos de llegada de otras personas son por lo general desconocidas. Es por

eso que existen las reglas de decisión, para que cada cliente pueda elegir entre ellas con una cierta probabilidad.

- Dentro del sistema ¿Qué rutas conviene tomar? Este es uno de los problemas más estudiados en los juegos de redes. En éstos hay varias rutas entre una o varias fuentes y uno o más destinos que implican costos o retrasos. Esto último depende de la congestión en cada parte de la ruta. La congestión es por supuesto una función de las decisiones de enrutamiento.

Así es como, para el análisis de problemas relacionados con colas, nos apoyamos frecuentemente en la teoría de juegos.

El concepto de la solución de *equilibrio de Nash* es comúnmente utilizado en los modelos de sistemas de colas (para ver ejemplos acerca de esto, véase [16]), y en muchos casos existen equilibrios múltiples. Este es un fenómeno común en las colas observables, es decir, sistemas en los cuales un cliente que llega observa la cola antes de tomar una decisión. Tal concepto fue desarrollado por el economista francés Antonie A. Cournot en el que planteó un modelo de varias empresas que compiten por un mismo bien y lo definió como la situación en la que los agentes económicos interactúan entre sí y eligen cada uno su mejor estrategia, dadas las estrategias que han elegido todos los demás; pero para nuestro interés, es visto como en el que cada uno de los clientes (jugadores) compiten por un mismo servicio (un bien común), y en el que cada uno de ellos intenta determinar su mejor decisión óptima que debe elegir para reducir su tiempo de espera (i.e. maximizar sus ganancias) de forma individual.

Una posible mejora del concepto de equilibrio es el de *subjuegos perfectos*. En este concepto, se prescriben respuestas óptimas para cada estado, incluyendo aquellos que están fuera de la trayectoria de equilibrio, es decir, los estados que no se visitan mientras que se utiliza la estrategia en consideración.

1.3. Motivación y antecedentes a colas paralelas

Desde años atrás, la atención se ha centrado en los sistemas de colas que permiten la variación en la estrategia de los servidores [4], [5]; tales variaciones incluyen varios servidores, sistemas de servicios prioritarios, variaciones de la tasa de servicio (en particular, uno de nuestros estudios realizados aquí) y las variaciones en el número de servidores como funciones de la longitud de la cola (o fila) o tiempo de retardo, etc.

También hay interés sobre los efectos de diferentes estrategias abiertas para el cliente. En particular nos enfocaremos a una de las estrategias del comportamiento *cliente impaciente*. En este trabajo, consideramos una estrategia de clientes, específicamente llamado la estrategia *"salto"*. En la literatura de la teoría de colas, el *"salto"* se refiere a los movimientos de los clientes que tienen la opción de cambiar de una fila a otra cuando hay varios servidores, donde cada uno tiene una cola disponible de espera por separado.

Por lo que podemos describir a un *sistema de colas con "salto"* como un sistema que involucra forzosamente múltiples servidores, una cola por cada servidor, donde cada cola es independiente una de la otra.

Este es un sistema con estructura en paralelo donde se permite que los clientes *"salten"* de una cola a otra para disminuir su tiempo de espera en la cola. [28], [1].

Como una motivación para ver las ventajas y necesidades de que podría ofrecernos un sistema de espera con estructura en paralelo, a continuación se presenta un ejemplo [32].

El estudio del rendimiento de un gran número de sistemas de satélites depende básicamente del análisis de cómo los sistemas de colas están relacionados. Las principales medidas interesantes de dicho análisis incluyen el rendimiento del sistema, el retardo medio de los paquetes en el satélite, y la probabilidad de desbordamiento del buffer[1] para el caso de tamaño de buffer finito. Sistemas de satélite multihaz[2] se han estudiado ampliamente (por ejemplo, véase Chlamtac Ganz 1986, y Chang, 1983), y se ha demostrado que proporcionan una mayor flexibilidad del sistema y un mejor rendimiento, pues cuando hay más de un buffer, la introducción de maniobras de los paquetes en espera entre los buffers parece ser una forma prometedora de mejorar el rendimiento de los sistemas. Por ejemplo, si permitimos que un paquete espere en el buffer con muchos paquetes esperando a trasladarse a algún otro buffer con un menor número de paquetes en espera en él, entonces el tiempo de espera promedio del paquete es obviamente reducido. Sin embargo, el análisis de los sistemas con *"salto"* es más difícil, porque no podemos lidiar con ellos mediante el análisis de una sola entrada y salida especificada (cola simple). En su lugar, se debe controlar el sistema como un sistema en paralelo.

Debido a la complejidad de los modelos de colas en paralelo con *"salto"*, los estudios analíticos han ido evolucionando poco a poco, para luego extenderse a dicho modelo. Por ejemplo, Haight [15] ha considerado un sistema que consta de dos colas no acotadas, con un único servidor, en el que un cliente a la llegada, se une a la cola más corta. Kingman [20] y Flatto McKean [12] hacen el supuesto de simetría entre las dos colas, donde tal suposición les permite utilizar funciones generadoras para estudiar el comportamiento de la solución estacionaria. Fayolle y Lasnogorodski [11], Cohen y Boxma [8] muestran cómo el análisis de dos colas en paralelo se puede reducir a partir del acoplamiento de procesos y la solución es un problema de valor límite de Riemann-Hilbert. Sin embargo, el enfoque no conduce a expresiones explícitas para las probabilidades de equilibrio.

1.4. Objetivos de tesis

El objetivo principal de la tesis es analizar cierta clase de sistemas de espera (con colas simples o colas paralelas) considerando técnicas de monotonicidad estocástica, de matrices geométricas y de ecuaciones en diferencias. Además, se utilizará el concepto de equilibrio de Nash para medir, en un sentido apropiado, la eficiencia de los sistemas en consideración.

Para esto, los objetivos particulares son:

[1] Un buffer (o búfer) en informática es un espacio de memoria, en el que se almacenan datos para evitar que el programa o recurso que los requiere, ya sea hardware o software, se quede sin datos durante una transferencia.

[2] Las antenas multihaz se utilizan generalmente en sistemas de satélite. Este tipo de antenas están formadas por "arreglos" de elementos capaces de generar varios haces por unidad de tiempo.

(a) Aplicar en un sistema de espera con colas simples el concepto de equilibrio de Nash (de la Teoría de Juegos) para garantizar la existencia de una estrategia óptima que le permita a un cliente decidir si debe unirse o no al sistema;

(b) Presentar la manera en que se calculan las probabilidades estacionarias para un sistema de dos colas en paralelo, utilizando el método de Neuts [24];

(c) Ser capaces de aplicar el método mencionado en (b) cuando, adicionalmente, existe la opción de adquirir cierta información para disminuir el tiempo de espera del cliente.

1.5. Contenido de la tesis

En este trabajo se presentan diversos modelos de sistemas de espera. En cada uno de ellos se hace el supuesto de la estabilidad, claramente cumpliendo con ciertas condiciones que hacen posible que así sea. Estos supuestos se han estudiado previamente y se expondrán en un primer capítulo.

Las herramientas que se utilizarán para determinar las probabilidades estacionarias son: álgebra lineal para la solución de sistemas de ecuaciones lineales con el método de eliminación de Gauss y la descomposición espectral de una matriz simétrica, acoplamiento de procesos, el método de solución matriz geométrica bajo condiciones apropiadas de estabilidad, y el método de funciones generadoras para la solución de ecuaciones en diferencia.

1.6. Estructura de la tesis

La estructura temática del texto es la siguiente. En el capítulo 2 se hace un pequeño estudio de las cadenas de Markov a tiempo continuo y las condiciones de estabilidad para sistemas de colas simples. Luego se presenta cómo un proceso de nacimiento y muerte puede describir a un modelo en paralelo (donde podemos introducir a la estrategia de *"salto"*) y de igual manera se da la condición de estabilidad para tal proceso el cual es una cadena de Markov a tiempo continuo bidimensional [30].

En el capitulo 3 se considera un sistema que opera de acuerdo a una disciplina FCFS (First Come First Service), en la cual la tasa de servicio es una función de la longitud de la cola. Los clientes arriban de manera secuencial al sistema, y deciden unirse o no usando reglas de decisión en base a la longitud de la cola a la llegada al sistema. Cada cliente está interesado en seleccionar una regla que cumple un cierto criterio de optimización con respecto a su tiempo de permanencia estimado en el sistema; como una consecuencia, las reglas de decisión para un juego con un jugador asociado, la estructura de las políticas de enrutamiento para el equilibrio de Nash están caracterizadas mediante un valor umbral no aleatorio.

En el capítulo 4, se presenta un sistema de espera con dos colas en paralelo cuando ocurre una estrategia de *"salto"*. En específico se calculan las probabilidades estacionarias para tres distintos valores umbrales y distintas variaciones en las tasas de servicio. Se muestra un diagrama para cada caso, de manera que hace que podamos visualizar el comportamiento de dicho modelo.

En el capítulo 5, también se estudia un sistema de espera con dos colas en paralelo cuando ocurre una estrategia *"salto"* para un valor umbral. A la llegada, cada cliente decide si compra la información sobre cuál cola es más corta o selecciona una de las colas al azar. En específico se calculan las probabilidades estacionarias para tres distintos valores umbral, con variaciones en las tasas de servicio. También se estudian las externalidades impuestas por un cliente sobre los demás, se obtiene una expresión explicita en el caso en que un *"salto"* se lleva a cabo tan pronto como las colas difieran en tres.

En el capítulo 6 daremos las conclusiones generales de todo el trabajo y propondremos algunos puntos para trabajo futuro.

El apéndice A se muestra de manera muy breve una descripción de un sistema de espera en la teoría de colas, y alguno de sus conceptos importantes que nos serán útiles en este trabajo, así como la notación universal de colas que se utiliza.

El apéndice B presenta algunos elementos básicos de la teoría de juegos.

Por último, en el apéndice C se incluye un poco de álgebra lineal, donde se analiza como una matriz simétrica puede ser representada por su descomposición espectral.

CAPÍTULO 2

Preliminares

2.1. Cadena de Markov a tiempo continuo

Esta sección y la subsección siguiente están basadas en las referencias: [27] y [10].

Suponga que tenemos un proceso estocástico a tiempo continuo $\{X(t), t \geq 0\}$ donde las variables aleatorias toman valores en el conjunto de los enteros no negativos. En analogía con la definición de una cadena de Markov a tiempo discreto, decimos que el proceso $\{X(t), t \geq 0\}$ es una *cadena de Markov a tiempo continuo* si para todo $s, t \geq 0$ y enteros no negativos $i, j, x(u)$, $0 \leq u < s$

$$P\{X(t+s) = j | X(s) = i,\ X(u) = x(u),\ 0 \leq u < s\} = P\{X(t+s) = j | X(s) = i\}. \tag{2.1}$$

En otras palabras, una cadena de Markov a tiempo continuo es un proceso estocástico con la propiedad (conocida como propiedad de Markov) de que la distribución condicional de el futuro $X(s+t)$ dado el presente $X(s)$ y el pasado $X(u)$, $0 \leq u < s$ depende solo del presente y es independiente del pasado. Si en adición $P\{X(t+s) = j | X(s) = i\}$ es independiente de s, entonces la cadena a tiempo continuo se dice que tiene probabilidades de transición estacionarias u homogéneas.

Suponga que una cadena de Markov a tiempo continuo entra al estado i en algún momento, por ejemplo al estado 0, y suponga que el proceso no deja el estado i (es decir, una transición no ocurre) durante los próximos 10 minutos. ¿Cuál es la probabilidad de que el proceso no deje el estado i durante los siguientes 5 minutos? Ya que el proceso esta en el estado i al tiempo 10 se sigue por la propiedad markoviana, que la probabilidad de que se mantenga en ese estado durante el intervalo [10,15] es justamente la probabilidad (incondicional) que se mantenga en el estado i de por lo menos cinco minutos.
Es decir, si denotamos a T_i la cantidad de tiempo que el proceso se mantiene en el estado i antes de hacer una transición a un estado diferente, entonces

$$P\{T_i > 15 | T_i > 10\} = P\{T_i > 5\},$$

o en general, por la misma razón

$$P\{T_i > s+t | T_i > s\} = P\{T_i > t\} \qquad \forall\, s,t \geq 0.$$

Por lo tanto, T_i es una variable aleatoria *sin memoria* y por lo tanto debe estar distribuida de forma exponencial (véase sección 5.2.2 de [27]).
De hecho, lo anterior nos da otra manera de definir una cadena de Markov a tiempo continuo. A saber, es un proceso estocástico que tiene la propiedad que cada vez que entra en el estado i:

(a) la cantidad de tiempo que pasa en ese estado antes de hacer una transición a un estado diferente tiene una distribución exponencial con media $1/\nu_i$, $\nu_i > 0$ (por ejemplo), y

(b) cuando el proceso deja el estado i, entra al próximo estado j con cierta probabilidad, por ejemplo $P_{i,j}$. Por supuesto la $P_{i,j}$ debe satisfacer

$$\begin{aligned} P_{i,i} &= 0 \qquad \forall i\ ; \\ \sum_j P_{i,j} &= 1 \qquad \forall i. \end{aligned}$$

En otras palabras, una cadena de Markov a tiempo continuo es un proceso estocástico que se mueve de un estado a otro de acuerdo a una cadena (a tiempo discreto) de Markov, pero es tal que la cantidad de tiempo que pasa en cada estado, antes de proceder al siguiente estado, es una distribución exponencial. Además la cantidad de tiempo que el proceso pasa en el estado i y visita al siguiente estado, deben ser variables aleatorias independientes. Porque si el siguiente estado fuera dependiente de T_i, entonces la información del tiempo en que el proceso ha permanecido en el estado i sería relevante para la predicción del siguiente estado, y esto contradice el supuesto markoviano.

2.1.1. Procesos de nacimiento y muerte

Definición 2.1.1 *Un proceso de nacimiento y muerte (PNM) es una cadena de Markov (homogénea) en tiempo continuo* $\{X(t), t \geq 0\}$ *con espacio de estados* $S = \{0, 1, 2, \cdots\}$ *cuyas probabilidades infinitesimales son de la forma*

$$q_{i,j} = \begin{cases} \lambda_i & si \quad j = i+1, \quad i \geq 0 \\ \mu_i & si \quad j = i-1, \quad i \geq 1 \\ -(\lambda_i + \mu_i) & si \quad i = j, \qquad i \geq 0 \\ 0 & si \quad |j-i| \geq 2 \end{cases} \tag{2.2}$$

para λ_i, μ_i *valores no negativos; y con* $\mu_0 = 0$ *por definición. Los parámetros* $\lambda_i = q_{i,i+1}$, $i = 0, 1, 2, \cdots$ *se llaman los índices (o intensidades) de nacimiento y los parámetros* $\mu_i = q_{i.i-1}$, $i = 1, 2, \cdots$ *se llaman los índices (o intensidades) de muerte.*

La última de las relaciones de (2.2), o sea

$$q_{i,j} = 0 \quad si \quad |j-i| \geq 2,$$

es una característica importante de los PNM; nos dice que un PNM es una cadena de Markov a tiempo continuo cuyas transiciones infinitesimales desde un estado i sólo pueden ocurrir a sus vecinos inmediatos $(i+1, i-1)$.

Así, estamos listos para darle una interpretación de los PNM a teoría de colas. Denotemos al proceso $X(t)$ como el número de clientes en un sistema de espera al tiempo t. Supongamos que cada vez que hay i personas en el sistema, entonces

1. las llegadas entran en el sistema a una tasa exponencial λ_i y
2. las personas abandonan el sistema a una tasa exponencial μ_i.

Es decir cada vez que hay i personas en el sistema, entonces el tiempo hasta la próxima llegada tiene una distribución exponencial con media $1/\lambda_i$ y este tiempo es independiente de la próxima salida la cual es en sí una distribución exponencial con media $1/\mu_i$.

De (2.1) (junto con (7.11) en [27]) , vemos que cuando $t \longrightarrow 0^+$, las probabilidades infinitesimales del PNM están dadas por:

$$\begin{aligned} p_{i,j}(t) &= \lambda_i t + o(t) && si \quad j = i+1 \quad (i \geq 0), \\ &= \mu_i t + o(t) && si \quad j = i-1 \quad (i \geq 1), \\ &= 1 - (\lambda_i + \mu_i)t + o(t) && si \quad i = j \quad (i \geq 0), \\ &= o(t) && si \quad |j-i| \geq 2 \quad (i \geq 0). \end{aligned}$$

Por otra parte, la ecuación de Kolmogorov hacia atrás para el PNM es

$$p'_{i,j}(t) = -(\lambda_i + \mu_i)p_{i,j}(t) + \lambda_i p_{i+1,j}(t) + \mu_i p_{i-1,j}(t), \tag{2.3}$$

para $i = 0, 1, \cdots$ $(\mu_0 = 0)$, así como la ecuación de Kolmogorov hacia adelante resulta

$$p'_{i,j}(t) = -(\lambda_j + \mu_j)p_{i,j}(t) + \lambda_{j-1} p_{i,j-1}(t) + \mu_{j+1} p_{i,j+1}(t). \tag{2.4}$$

Así, la ecuación diferencial para la distribución $p_j(t) = P\{X(t) = j\}, \quad j = 0, 1, ...$ del proceso en el tiempo t está dada por:

$$p'_j(t) = -(\lambda_j + \mu_j)p_{j-1}(t) + \lambda_{j-1}p_{j-1}(t) + \mu_{j+1}p_{j+1}(t), \tag{2.5}$$

para $j = 0, 1, ...$, con $p_{-1}(t) \equiv 0$. Así podemos expresar estas probabilidades de transición por medio de la siguiente matriz $Q = (q_{i,j})$

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & 0 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3 + \mu_3) & \lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Donde se define la matriz Q como la *matriz generadora infinitesimal* (o matriz de transición de tasas) ya que sus elementos son las tasas instantáneas de salir de un estado para pasar a otro [10].

2.1.2. Distribución estacionaria para el PNM

La distribución estacionaria $\pi_j, j = 0, 1, ...$ del PNM se puede calcular directamente de la ecuación (2.5), tomando $p_j(t) \equiv \pi_j$ (constante). Puesto que la derivada de una constante es cero, obtenemos de (2.5) que para $j = 1, 2, ...$, las π_j satisfacen

$$0 = -\pi_j(\lambda_j + \mu_j) + \pi_{j-1}\lambda_{j-1} + \pi_{j+1}\mu_{j+1}, \tag{2.6}$$

$$0 = -\pi_0\lambda_0 + \pi_1\mu_1 \ . \tag{2.7}$$

En teoría de colas, la ecuación (2.8), la cual se puede escribir como

$$\pi_j(\lambda_j + \mu_j) = \pi_{j-1}\lambda_{j-1} + \pi_{j+1}\mu_{j+1}, \tag{2.8}$$

se llama *ecuación de balance* porque describe el hecho de que, en estado estacionario, el "flujo de entrada" al sistema es igual al "flujo de salida".

Para determinar las distribución estacionaria π_j debemos resolver el sistema de ecuaciones (2.6) y (2.7). Supongamos que $\mu_j > 0$ para todo $j \geq 1$. Luego, de (2.7), se tiene que $\pi_1 = (\lambda_0/\mu)\pi_0$, y de (2.6),

$$\pi_2\mu_2 = \pi_1(\lambda_1 + \mu_1) - \pi_0\lambda_0 = \pi_0\frac{\lambda_0\lambda_1}{\mu_1},$$

o bien

$$\pi_2 = \pi_0 \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2}.$$

Análogamente se pueden obtener $\pi_3, \pi_4, \cdots$. En general, se puede probar fácilmente, por inducción, que

$$\pi_j = \pi_0 \frac{\lambda_0 \lambda_2 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j}, \qquad j = 1, 2, \cdots \tag{2.9}$$

Para que $\{\pi_j, j \geq 0\}$ sea, efectivamente, una distribución de probabilidad debe satisfacer que $\sum_{j=0}^{\infty} \pi_j = 1$, es decir,

$$\pi_0 \left(1 + \sum_{j=0}^{\infty} \frac{\lambda_0 \lambda_2 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} \right) = 1,$$

en cuyo caso

$$\pi_0 = \left(1 + \sum_{j=0}^{\infty} \frac{\lambda_0 \lambda_2 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} \right)^{-1}. \tag{2.10}$$

Obviamente ésto se cumple si y sólo si la serie entre paréntesis converge. Así podemos concluir que el PNM con probabilidades infinitesimales (A.2) tiene una distribución estacionaria si, y sólo si, la serie

$$\sum_{j=1}^{\infty} \frac{\lambda_0 \lambda_2 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} \tag{2.11}$$

converge.

2.2. Modelo de colas paralelas basados en el proceso de nacimiento y muerte

El modelo de colas $M/M/s$ (véase apéndice A) es el más utilizado en el análisis de las estaciones de servicio con más de un servidor, como los bancos, las cajas registradoras en las tiendas, mostradores de facturación, en los aeropuertos y similares. Como ya hemos mencionamos en la sección 2.1, existen distintos comportamientos de clientes que van llegando al sistema, en particular, saltar entre colas para reducir

el tiempo de espera, que denotaremos como "salto". Este "salto" describe forzosamente un sistema de espera en paralelo, el cual se puede analizar como un proceso de Markov bidimensional de nacimiento y muerte.

Sea $N_k(t)$ la longitud de la cola k, $(k = 1, 2)$ en el tiempo t. Ya que las llegadas son procesos de Poisson y los tiempos de servicio son exponenciales, entonces $\{(N_1(t), N_2(t)), t \geq 0\}$ describen un proceso de Markov. El espacio de estados del proceso de Markov es bidimensional $\{(i, j) : i \geq 0,\ 0 \leq j \leq H\}$, para un H dado entero positivo y se hace referencia por el *nivel* n al conjunto de estados $\{(n, 0), (n, 1), ..., (n, H)\}$. Tal proceso de Markov es llamado un PNM homogéneo donde las transiciones de un paso están restringidas a estados en el mismo *nivel* o a dos *niveles* adyacentes y suponemos que las tasas de transiciones son independientes en cada *nivel* .

En este proceso de Markov de nacimiento y muerte se encuentra descrito (en términos probabilísticos) cómo cambia $N_k(t)$ al aumentar t. Donde supondremos que para un cierto estado, una llegada a la cola se interpreta como un nacimiento y una salida (o término de servicio) como una muerte. Sea un estado (i, j) en el cual se encuentra el proceso, entonces se describe que

- desde (i, j) a $(i + 1, j)$ hay un nacimiento con tasa $\lambda/2$,
- desde (i, j) a $(i, j + 1)$ hay un nacimiento con tasa $\lambda/2$,
- desde $(i + 1, j)$ a $(i + 1, j + 1)$ hay un nacimiento con tasa λ,
- desde $(i, j + 1)$ a $(i + 1, j + 1)$ hay un nacimientos con tasa λ,
- desde $(i + 1, j)$ a (i, j) hay una muerte con tasa μ_1,
- desde $(i, j + 1)$ a (i, j) hay una muerte con tasa μ_2,
- desde $(i + 1, j + 1)$ a $(i + 1, j)$ y $(i, j + 1)$ hay dos muertes con tasas respectivamente μ_1 y μ_2.

2.3. Estabilidad en colas paralelas

Para un PNM como se describe en la sección 2.2, ordenamos los estados lexicograficamente, es decir

$$\{(0, 0), ..., (0, H), (1, 0), ..., (1, H), ..., (n, 0), ..., (n, H), ...\},$$

y suponemos que la *matriz generadora infinitesimal* Q tiene la siguiente estructura de tridiagonal en bloques:

$$Q = \begin{pmatrix} B_1 & B_0 & & & & & \\ B_2 & A_1 & A_0 & & & & \\ & A_2 & A_1 & A_0 & & & \\ & & A_2 & A_1 & A_0 & & \\ & & & A_2 & A_1 & A_0 & \\ & & & & \ddots & \ddots & \ddots \end{pmatrix},$$

donde A_0, A_1 y A_2 son matrices cuadradas de orden $H+1$. Las matrices A_0, A_2, B_0 y B_2 son no negativas y las matrices B_1 y A_1 tienen elementos no negativos fuera de la diagonal y estrictamente negativos en la

diagonal. Los renglones de Q suman cero.

El PNM dado por Q es ergódico si y solo si satisface el método de "mean drift condition" (ver [24], Teorema 1.7.1)

$$\omega A_0 e < \omega A_2 e, \tag{2.12}$$

donde $\omega = (\omega_0, ..., \omega_H)$ es la distribución de equilibrio del generador $A_0 + A_1 + A_2$ y e es el vector unitario. Cuando (2.12) se satisface, la distribución estacionaria de el PNM existe. Denotando por $\pi(i, j)$ la probabilidad estacionaria del proceso iniciando en el estado (i, j), y usando la notación del vector $\pi_n = (\pi(n, 0), ..., \pi(n, H))$, las ecuaciones de balance de el PNM están dadas por

$$\pi_{n-1} A_0 + \pi_n A_1 + \pi_{n+1} A_2 = \mathbf{0}, \qquad n \geq 2, \tag{2.13}$$

y

$$\pi_0 B_1 + \pi_1 B_2 = 0, \tag{2.14}$$

$$\pi_0 B_0 + \pi A_1 + \pi A_2 = 0. \tag{2.15}$$

Introduciendo *la matriz de tasas* R como la solución mínima no negativa de la ecuación de matriz no líneal

$$A_0 + R A_1 + R^2 A_2 = 0, \tag{2.16}$$

se puede probar que las probabilidades de equilibrio satisfacen (véase por ejemplo [24], pp. 80-83)

$$\pi_{n+1} = \pi_n R, \qquad n \geq 1. \tag{2.17}$$

Los vectores π_0 y π_1 se determinan de las condiciones límite (2.14)-(2.15) y la condición de normalización

$$\sum_{i=0}^{\infty} \sum_{j=0}^{H} \pi(i, j) = \pi_0 e + \pi_1 (I - R)^{-1} e = 1, \tag{2.18}$$

donde I representa la matriz de identidad. Con el fin de determinar la distribución estacionaria, uno debería de poder determinar *la matriz de tasas* R. Varios procedimientos iterativos existen para resolver (2.18). Por ejemplo, el método modificado SS (véase [26] y [1]) utiliza el siguiente esquema

$$R^{(k+1)} = -[A_0 + {R^{(k)}}^2 A_2] A_1^{-1}, \qquad k = 0, 1, ..., \tag{2.19}$$

iniciando con $R^{(0)}$ la matriz cuyas entradas son ceros.

CAPÍTULO 3

Equilibrios de Nash para los sistemas de Colas FCFS con tasa de servicio creciente.

3.1. Introducción

Uno de los trabajos primarios y más relevante en el análisis de teoría de juegos dentro de la clase de sistemas de colas se llevó a cabo por Altman y Shimkin [3], en el que se investigó un sistema de procesador compartido, el cual que será el ejemplo de motivación para este capítulo.

Altman y Shimkin consideran a los usuarios potenciales de un ordenador, donde cada uno requiere el uso de un ordenador para ejecutar un cierto trabajo, estos llegan secuencialmente a una instalación informática. Los usuario, a su llegada, pueden elegir entre las siguientes dos opciones: o bien conectarse a una *computadora central* (MF), que normalmente sirve a muchos usuarios en paralelo, o utilizar una *computadora personal* (PC), donde esta alternativa es una opción que tiene costos constantes. Cada usuario es el único interesado en minimizar su tiempo de servicio, que coincide con el tiempo de permanencia.

El servicio en el ordenador MF se realiza de acuerdo a la disciplina de *procesador compartido*. Un usuario que llega puede observar la intensidad de la corriente de llegadas, a saber, el número de usuarios que ya están en MF. Sin embargo, para evaluar su tiempo de servicio esperado en MF se debe tener en cuenta la posible intensidad en este ordenador a lo largo de su tiempo de servicio, la cual se ve afectada por las decisiones de los usuarios subsiguientes. Esto le llevó a considerar un problema en un marco de teoría de juegos donde se exploró la solución de equilibrio de Nash para un juego dinámico.

Ahora bien, en este capítulo, consideremos un sistema de *servicio compartido* denotado por Qs, donde cada cliente (usuario) a su llegada, podrá elegir unirse a Qs (interpretado como el ordenador MF), o negarse a ésto (interpretado como unirse al ordenador PC).

En este capítulo se analiza el comportamiento del cliente al unirse a un sistema de colas de un solo servidor con disciplina First Come First Service (FCFS) y cupo finito o infinito. El proceso de salida es un proceso de Poisson con tasa $\mu(x)$ si x clientes están presentes, donde la tasa de servicio responde a los cambios en el tamaño de la cola. Los clientes a su llegada deciden unirse o no en base a la longitud de la cola y están dispuestos a unirse al sistema solo si su tiempo estimado no es demasiado alto, es decir, cada cliente está interesado en seleccionar una regla que cumpla con un criterio de optimalidad con respecto a su tiempo estimado de permanencia en el sistema.

Esto nos lleva a analizar el problema como un juego *no cooperativo*, con jugador infinito (estacionario), donde los tiempos de permanencia estimados, en particular, los estados iniciales en que entran son los considerados. El objetivo es caracterizar las condiciones bajo las cuales las políticas de unión de equilibrio de Nash existen y explorar la estructura de dichas políticas.

En la siguiente sección se específica en detalle el modelo, que incluye una descripción minuciosa de las reglas de decisión utilizadas por los clientes que llegan al sistema. En la sección 3.3 se presentan las variables aleatorias que se generan en el modelo y los diversos procesos definidos con respecto a éstas. En la sección 3.4 se utilizan argumentos de acoplamiento para establecer resultados de orden estocástico para el tiempo de permanencia en Qs con respecto al estado de entrada. A esto, le sigue la sección 3.5 un debate sobre la monotonicidad y la continuidad del tiempo de permanencia en lo que respecta a las políticas de umbral simétrico. Las propiedades establecidas en estas dos últimas secciones se reunieron en la sección 3.6 para caracterizar la existencia de *políticas de equilibrio de Nash simétricas (SNEPs).*

3.2. El modelo

Sea $\mathbb{Z}^+ = \{1, 2, ...\}$, $\mathbb{N} = \mathbb{Z}^+ \cup \{0\}$ y $\mathbb{R}^+ = \{x \in \mathbb{R} \,|x > 0\}$ en todo el capítulo. Consideremos un sistema de *servicio compartido* Qs. Un cliente que llega al sistema tiene que elegir entre unirse a Qs, o negarse a ésto.

Se supone que Qs tiene un cupo de tamaño B, que puede ser finito o infinito. Cualquier cliente que llega cuando el cupo está lleno no se le permite entrar al sistema.

El proceso de salida en Qs de longitud de cola x, forma un proceso de Poisson a una tasa $\mu(x)$, donde $\mu(x)$, $x \in \mathbb{N}$ es una función estrictamente creciente y acotada en $\mathbb{N}$, con $\mu(0)$=0. El conjunto $\overline{\mu} = sup\{\mu(x) : x = 1, 2, ...\}$.

Sea θ un número no negativo, donde éste se define como el valor *umbral* de servicio, es decir la tasa de servicio máxima a la cual los clientes estarían dispuestos a ser atendidos en el sistema Qs. Si un cliente que llega percibe que el tiempo de permanencia esperado en Qs es mayor que este valor, entonces se mostrará renuente a entrar al sistema. Se supone que $\mu(1)^{-1} < \theta$. Esta condición garantiza que siempre vale la pena para un cliente acceder a Qs si el sistema está vacío a su llegada.

Sea el número de clientes en Qs al tiempo t denotado por $\mathrm{X}(t)$ con el estado inicial $\mathrm{X}(0) = x_0$. Sea A_k el tiempo de llegada del k-ésimo cliente al sistema (aunque no necesariamente dentro), donde $0 = A_0 < A_1 < A_2 < ...$; denotamos a este k-ésimo cliente por la etiqueta C_k, $k \in \mathbb{N}$ donde se supone que C_0 llega al tiempo A_0 (i.e., al tiempo 0). Llamamos a la sucesión de clientes $C_0, C_1, ..., C_k, ...$ *corriente de arribos* (total). La decisión de si C_k entra o no, se toma sobre la base de $\mathrm{X}(A_k)$; la longitud de la cola en Qs justo *antes* de la llegada de C_k.

Una *regla de decisión*, $u(\cdot) : \{0, 1, ..., B-1\} \longrightarrow [0, 1]$, se define como una función que específica la probabilidad con la que un cliente entra a Qs, la cual es igual a $u(x)$ si el número de clientes en Qs es igual a x, justo antes de su llegada. La *regla de decisión* para C_k está representada por $u(\cdot)$, y la colección de reglas de decisión usadas por cada cliente, una *política*, esta denotada por el vector $\delta = (u_0(\cdot), u_1(\cdot), u_2(\cdot), ..., u_k(\cdot), ...)$. Sea $v_k(x, \delta)$, $x \in \{0, 1, ..., B-1\}$ el tiempo de permanencia de C_k en Qs, dado que x clientes están presentes en Qs justo antes de su llegada, y que cualquier cliente que llega en el futuro se adhiere a esta regla de decisión inferida por δ. Además se define $V_k(x, \delta)$ por el valor de la esperanza de $v_k(x, \delta)$.

Se supone que no hay colaboración entre clientes y cada cliente busca elegir una regla óptima para unirse con respecto a alguna medida de su tiempo de permanencia previsto en Qs y la calidad de servicio estipulado. Teniendo en cuenta estos puntos, estamos preparados para analizar este sistema en el paradigma de un juego *no cooperativo* con jugador infinito.

Una regla de decisión $u_k(\cdot)$ para el k-ésimo cliente de la corriente de llegada total se dice que es *óptima* respecto a la política δ si

$$u_k(x) = \begin{cases} 1 & V_k(x, \delta) < \theta, \\ 0 & V_k(x, \delta) > \theta, \\ q & V_k(x, \delta) = \theta, \end{cases} \tag{3.1}$$

para $0 \leq q \leq 1$, $x \in \{0, 1, ...B-1\}$. Así, la colección de todas las posibles *reglas de decisión* para C_k que son óptimas respecto a δ es denotado por $\mathbb{U}_k(\delta)$.

Una política $\delta = (u_0(\cdot), u_1(\cdot), u_2(\cdot), ..., u_k(\cdot), ...)$ se dice que es una *política de equilibrio de Nash* si, para todo $k \in \mathbb{N}$, la regla de decisión de el k-ésimo cliente, $u_k(\cdot)$, es óptima respecto a δ.

3.3. Variables aleatorias y procesos estocásticos asociados al modelo

Sea $\{M_i : i \in \mathbb{Z}^+\}$, $\{N_j : j \in \mathbb{Z}^+\}$, $\{U_k : k \in \mathbb{Z}^+\}$ y $\{U'_l : l \in \mathbb{Z}^+\}$ sucesiones mutuamente independientes de variables aleatorias, donde:

- $\{M_i : i \in \mathbb{Z}^+\}$ es una sucesión de variables aleatorias continuas independientes idénticamente distribuidas ($i.i.d$) con media $0 < \lambda^{-1} < \infty$;

- $\{A_k : k \in \mathbb{N}\}$ es una sucesión de tiempos de llegada al sistema donde $A_0 := 0$ y $A_k := \sum_{i=1}^{k} M_i$, $k \in \mathbb{Z}^+$.
- $\{U_k : k \in \mathbb{N}\}$ es una sucesión de variables aleatorias *i.i.d* las cuales están distribuidas uniformemente en el intervalo (0,1]. La variable aleatoria U_k se utiliza para decidir si el cliente C_k entra o no a Qs;
- $\{N_j : j \in \mathbb{Z}^+\}$ es una sucesión de variables aleatorias *i.i.d* exponenciales con media $\overline{\mu}^{-1} < \infty$;
- $\{S_l : l \in \mathbb{Z}^+\}$ es una sucesión de tiempos de complementación de servicio *potencial* para los clientes en Qs, donde $S_l := \sum_{j=1}^{l} N_j$;
- $\{U'_l : l \in \mathbb{Z}^+\}$ es una sucesión de variables aleatorias *i.i.d* uniformes en (0,1] utilizadas para determinar si un tiempo de salida potencial corresponde a una *salida efectiva* o a un *evento ficticio*.

Además definimos $\{t_n : n \in \mathbb{Z}^+\}$ siendo las *estadísticas de orden* para el conjunto $\{A_k : k \in \mathbb{N}\} \cup \{S_l : l \in \mathbb{Z}^+\}$, donde $t_i < t_j$ para $i < j$.
Las especificaciones de las decisiones de llegada y salida de Qs se representan al final de la sección. Esto proporcionará una mayor motivación para las definiciones formales de los procesos estocásticos que se dan a continuación.

Definición 3.3.1 ***(Proceso de longitud de cola)*** *Para un determinado estado inicial* $\mathrm{X}(0) = x_0$ *y política* δ*, sea el proceso* $\{\mathrm{X}(t) : t \geq 0\}$ *la longitud de la cola, donde* $\mathrm{X}(t)$ *representa el número de clientes en el sistema al tiempo* t*. Este proceso está definido de forma tal que es continuo por la izquierda, constante a trozos y con saltos potenciales descritos por las siguientes relaciones*

$$\begin{aligned} \mathrm{X}(A_k^+) &= \mathrm{X}(A_k) + \mathbf{1}\{\mathrm{U}_k < u_k(\mathrm{X}(A_k))\}, \qquad k \in \mathbb{N} \\ \mathrm{X}(S_l^+) &= x_l - \mathbf{1}\{\mathrm{U}'_l < \mu(\mathrm{X}(S_l))/\overline{\mu}\}, \qquad l \in \mathbb{Z}^+ \end{aligned} \tag{3.2}$$

donde $\mathbf{1}$ *es la función indicadora.*

Observemos que si U'_l fuera elegida uniforme en el intervalo [0,1] en lugar de (0,1], entonces tendríamos que incluir $\mathrm{X}(S_l) > 0$ dentro del indicador de la segunda relación de funciones en (3.2).

Definición 3.3.2 ***(Proceso de "transiciones de servicios residual" (RST))*** *Sea* $\{\mathrm{Z}(t) : t \geq 0\}$ *el proceso-RST. Este proceso está definido y es continuo por la izquierda, constante por pedazos y no creciente, tal que* $\mathrm{Z}(0) = \mathrm{X}(0) = x_0$ *y con saltos potenciales (coincidiendo con tiempos de salida), satisfaciendo la siguiente relación:*

$$\mathrm{Z}(S_l^+) = \mathrm{Z}(S_l) - \mathbf{1}\{\mathrm{Z}(S_l) > 0,\ \mathrm{U}'_l < \mu(\mathrm{X}(S_l))/\overline{\mu}\}, \qquad l \in \mathbb{Z}^+. \tag{3.3}$$

Cuando C_0 está en la cola, entonces $\mathrm{Z}(t)$ representa el número de clientes presentes menos los que se encuentran detrás de C_0 (o el número de transiciones de servicio actual que todavía tiene que ocurrir antes de las salidas de C_0) al tiempo t, y $\mathrm{Z}(t) = 0$ si el cliente C_0 no está presente en el tiempo t (se sugiere ver el apéndice A.2).

Definición 3.3.3 ***(Tiempo de permanencia/estancia)*** *Si C_0 realmente entra a Qs, entonces su tiempo de permanencia sera igual a*

$$v_0 = min\{t : \mathrm{Z}(t) = 0\}.$$

3.3.1. Llegadas a Qs

Al tiempo A_k, el cliente C_k llega a Qs y entra al sistema con probabilidad γ, donde este valor depende de su regla de decisión y el valor de $\mathrm{X}(A_k)$. Así su decisión actual está basada en el valor de la variable aleatoria U_k y γ de manera que C_k entra a Qs si y solo si $\mathrm{U}_k < \gamma$.

3.3.2. Servicio en Qs

Para facilitar la exposición, definimos x_l a ser igual a $\mathrm{X}(S_l)$, la longitud de la cola en Qs justo antes de una salida potencial en el tiempo S_l. Si $\mathrm{U}_1' \in \left(\frac{\mu(x_l)}{\overline{\mu}}, 1\right]$, entonces S_l es considerado un instante de complementación de servicio ficticio; de lo contrario cualquier cliente en el servidor completa su servicio y se aparta del sistema.

El procedimiento anterior invoca una técnica de uniformización [22]. El hecho de que este procedimiento genera los tiempo reales de salida con la correcta distribución puede considerarse como sigue. Siempre y cuando la longitud de la colas permanezca en $x \in \mathbb{Z}^+$, la salida potencial siguiente se genera a partir de un proceso de Poisson con tasa $\mu(x)$. Consideremos ahora un proceso de Poisson en el cual los eventos ocurren con la tasa uniforme $\overline{\mu}$, la tasa más rápida a la cual una salida podría ocurrir. Siempre que la longitud de la cola es x, y un evento del proceso de Poisson con tasa μ se produce, entonces corresponde a una salida real con una probabilidad $\mu(x)/\overline{\mu}$, independientemente de todos los demás eventos. Pero como este evento corresponde a un muestreo Bernoulli de un proceso de Poisson, entonces las salidas en la longitud x de la cola son de Poisson con tasa $\overline{\mu} \times \mu(x)/\overline{\mu} = \mu(x)$, como se había previsto.

3.4. Monotonicidad con respecto al tamaño de la cola al entrar

Se presenta, en un sentido de dominación estocástica, que $v_0(x, \delta)$ en una función creciente de x para cualquier δ que es miembro de una determinada clase de políticas, está clase está definida como sigue:

Definición 3.4.1 *Sea $\mathbb{T}^\infty$ la clase de políticas en las que la regla de decisión para cada cliente es una función no creciente de $x \in \{0, 1, ..., B-1\}$.*

Evaluar la distribución de $v_0(x,\delta)$ no parece ser tan sencillo. Esta dificultad se evita mediante la utilización de la técnica de acoplamiento [29] y las técnicas de inducción hacia adelante. Las colecciones de variables aleatorias y procesos estocásticos que entre comillas se refiere como "sistemas" en las que estas comparaciones estocásticas se basarán, se introducen a continuación.

Definición 3.4.2 *El **sistema** $\mathcal{X}$ está caracterizado por los conjuntos de variables aleatorias, reglas de decisión y los procesos estocásticos que se mencionan a continuación.*
(I): $\mathcal{M}$=$\{M_i\}$, $\mathcal{N}$=$\{N_j\}$, $\mathcal{U}$=$\{U_k\}$ y $\mathcal{U}'$=$\{U'_l\}$;
(II): La sucesión de tiempos de arribos $\mathcal{A}=\{A_k\}$, y sucesión de tiempos potenciales de salida $\mathcal{S}=\{S_l\}$;
(III): el cliente C_0 entra a Qs al tiempo $A_0=0$ con todos los clientes adheridos a la política $\delta\in\mathbb{T}^\infty$;
(IV): el proceso de longitud de cola $\{\mathrm{X}(t):t\geq 0\}$ con $\mathrm{X}(0)=x$;
(V): el proceso-RST $\{\mathrm{Z}(t):t\geq 0\}$ con $\mathrm{Z}(0)=x$;
(VI): v_o, el tiempo de permanencia de C_0 en Qs.

Definición 3.4.3 *El **sistema** $\widetilde{\mathcal{X}}$ está caracterizado de manera similar a la del sistema $\mathcal{X}$, excepto que las cantidades de $(\widetilde{II})$-$(\widetilde{VI})$ se definen en términos de $(\widetilde{I})$ de manera obvia.*

$(\widetilde{I})$: $\widetilde{\mathcal{M}}$=$\{\widetilde{M}_i\}$, $\widetilde{\mathcal{N}}$=$\{\widetilde{N}_j\}$, $\widetilde{\mathcal{U}}$=$\{\widetilde{U}_k\}$ y $\widetilde{\mathcal{U}}'$=$\{\widetilde{U}'_l\}$ tienen las mismas distribuciones como en $\mathcal{M}$, $\mathcal{N}$, $\mathcal{U}$, $\mathcal{U}^\phi$ y $\mathcal{U}'$;
$(\widetilde{II})$: La sucesión de tiempos de arribos $\widetilde{\mathcal{A}}=\{\widetilde{A}_k\}$ y sucesión de tiempos potenciales de salida $\widetilde{\mathcal{S}}=\{\widetilde{S}_l\}$;
$(\widetilde{III})$: como en (II) para $\widetilde{A}_0=0$;
$(\widetilde{IV})$: el proceso de longitud de cola $\{\widetilde{\mathrm{X}}(t):t\geq 0\}$ con $\widetilde{\mathrm{X}}(0)=x+1$;
$(\widetilde{V})$: el proceso-RST $\{\widetilde{\mathrm{Z}}(t):t\geq 0\}$ con $\widetilde{\mathrm{Z}}(0)=x+1$;
$(\widetilde{VI})$: $\tilde{v}_0$, el tiempo de permanencia de C_0 en Qs.

Tenemos la intención de relacionar $\mathcal{X}$ con $\widetilde{\mathcal{X}}$ entre sí con el siguiente acoplamiento. Este acoplamiento se puede definir ya que existe un proceso de Markov a tiempo continuo, un estado inicial en el proceso y un espacio de probabilidad común [29], Teorema 2.10.

Definición 3.4.4 *(**Acoplamiento** $\mathcal{C}$) Sea*

$$\begin{aligned} M_i &= \widetilde{M}_i,\ i\in\mathbb{Z}^+ \\ N_j &= \widetilde{M}_j,\ j\in\mathbb{Z}^+ \\ U_k &= \widetilde{U}_k,\ k\in\mathbb{N} \\ U'_l &= \widetilde{U}'_k,\ l\in\mathbb{Z}^+ \end{aligned}$$

El efecto de este procedimiento es que los instantes de llegada $\{A_k\}$, los tiempos potenciales de salida $\{S_l\}$, las $\{U_k\}$, posiciones de los clientes controlados en la sucesión de llegada total, y la $\{U'_l\}$ toman los mismos valores bajo $\mathcal{X}$ que sus contra partes en $\widetilde{\mathcal{X}}$ en cada realización.

El siguiente resultado nos permite inferir que bajo el acoplamiento descrito arriba, $v_0\leq\tilde{v}_0$.

Lema 3.4.1 *Para el sistema $(\mathcal{X}, \widetilde{\mathcal{X}})$ bajo el acoplamiento $\mathcal{C}$, en el que $\delta \in \mathbb{T}^{\infty}$, uno de los siguientes conjuntos de relaciones se llevará a cabo en cada tiempo $t \in \mathbb{R}^{+}$:*

$$\begin{aligned} \mathrm{X}(t) + 1 &= \widetilde{\mathrm{X}}(t), \\ Z(t) + 1 &= \widetilde{\mathrm{Z}}(t), \end{aligned} \tag{3.4}$$

$$\begin{aligned} \mathrm{X}(t) &= \widetilde{\mathrm{X}}(t), \\ \mathrm{Z}(t) &= \widetilde{\mathrm{Z}}(t), \end{aligned} \tag{3.5}$$

$$\begin{aligned} \mathrm{X}(t) &= \widetilde{\mathrm{X}}(t), \\ \mathrm{Z}(t) + 1 &= \widetilde{\mathrm{Z}}(t). \end{aligned} \tag{3.6}$$

Demostración: Por definición de $\mathcal{X}$ y $\widetilde{\mathcal{X}}$,

$$\mathrm{Z}(0^{+}) = \mathrm{X}(0^{+}) = x + 1 < x + 2 = \widetilde{\mathrm{X}}(0^{+}) = \widetilde{\mathrm{Z}}(0^{+}). \tag{3.7}$$

Supongamos que t_{n+1} corresponde a una llegada, con $t_{n+1} = A_r$ o una salida con $t_{n+1} = S_m$, tal que C_0 está todavía presente en Qs bajo $\mathcal{X}$. Antes de continuar, observemos que si t_{n+1} efectivamente corresponde a un tiempo de llegada, entonces debido a la clase de políticas y el acoplamiento, se considera que uno de los siguientes tres escenarios se debe cumplir:

(i) C_r entra a Qs bajo ambos $\mathcal{X}$ y $\widetilde{\mathcal{X}}$,
(ii) C_r entra a Qs bajo solo $\mathcal{X}$,
(iii) C_r no entra a Qs bajo $\mathcal{X}$ y $\widetilde{\mathcal{X}}$.

Caso 1: Suponga que (3.4) se cumple al tiempo t_n^{+}.

$\underline{t_{n+1} \in \{A_k\}}$.
Bajo (i),

$$\mathrm{X}(t_{n+1}^{+}) = \mathrm{X}(t_n^{+}) + 1 = \widetilde{\mathrm{X}}(t_n^{+}) = \widetilde{\mathrm{X}}(t_{n+1}^{+}) - 1 \, . \tag{3.8}$$

Bajo (ii),

$$\mathrm{X}(t_{n+1}^{+}) = \mathrm{X}(t_n^{+}) + 1 = \widetilde{\mathrm{X}}(t_n^{+}) = \widetilde{\mathrm{X}}(t_{n+1}^{+}) \, . \tag{3.9}$$

Bajo (iii), los estados del proceso de la longitud de cola para $\mathcal{X}$ y $\widetilde{\mathcal{X}}$ al tiempo t_{n+1}^{+} son el mismo como lo fueron en el tiempo t_n^{+}.

También dado que C_r se colocaría detrás de C_0 si fuera a entrar a Qs, entonces no habría ningún cambio en los procesos de Z(t)-RST en cada uno de los escenarios anteriores, es decir,

$$\mathrm{Z}(t_{n+1}^+) = \mathrm{Z}(t_n^+) = \widetilde{\mathrm{Z}}(t_n^+) - 1 = \widetilde{\mathrm{Z}}(t_{n+1}^+) - 1 \ . \tag{3.10}$$

Así al tiempo t_{n+1}, (3.4) se mantiene bajo escenarios (i) y (iii), mientras (3.6) se mantiene bajo el escenario (ii).

$\underline{t_{n+1} \in \{S_l\}}$
ya que $x_m < \widetilde{x}_m$ por hipótesis, entonces $\mu(x_m) < \mu(\widetilde{x}_m)$.
Si $\mathrm{U}'_m \leq \mu(x_m)/\overline{\mu}$, entonces una actual salida ocurre bajo ambos $\mathcal{X}$ y $\widetilde{\mathcal{X}}$. Por lo tanto

$$\mathrm{X}(t_{n+1}^+) = \mathrm{X}(t_n^+) - 1 = (\widetilde{\mathrm{X}}(t_n^+) - 1) - 1 = \widetilde{\mathrm{X}}(t_{n+1}^+) - 1 \ , \tag{3.11}$$
$$\mathrm{Z}(t_{n+1}^+) = \mathrm{Z}(t_n^+) - 1 = (\widetilde{\mathrm{Z}}(t_n^+) - 1) - 1 = \widetilde{\mathrm{Z}}(t_{n+1}^+) - 1 \ . \tag{3.12}$$

Así (3.4) se cumple al tiempo t_{n+1}^+.
Si $\mathrm{U}'_m \in (\mu(x_m)/\overline{\mu}, \mu(\widetilde{x}_m)/\overline{\mu}]$, entonces una salida real ocurre bajo $\widetilde{\mathcal{X}}$, pero no bajo $\mathcal{X}$.

$$\mathrm{X}(t_{n+1}^+) = \mathrm{X}(t_n^+) = \widetilde{\mathrm{X}}(t_n^+) - 1 = \widetilde{\mathrm{X}}(t_{n+1}^+) \ , \tag{3.13}$$
$$\mathrm{Z}(t_{n+1}^+) = \mathrm{Z}(t_n^+) = \widetilde{\mathrm{Z}}(t_n^+) - 1 = \widetilde{\mathrm{Z}}(t_{n+1}^+) \ . \tag{3.14}$$

es decir, (3.5) se cumple al tiempo t_{n+1}.

Si $\mathrm{U}'_m > \mu(\widetilde{x}_m)/\overline{\mu}$, entonces los estados de los procesos permanecen inalterados,

$$\mathrm{X}(t_{n+1}^+) = \mathrm{X}(t_n^+) = \widetilde{\mathrm{X}}(t_n^+) - 1 = \widetilde{\mathrm{X}}(t_{n+1}^+) - 1 \ , \tag{3.15}$$
$$\mathrm{Z}(t_{n+1}^+) = \mathrm{Z}(t_n^+) = \widetilde{\mathrm{Z}}(t_n^+) - 1 = \widetilde{\mathrm{Z}}(t_{n+1}^+) - 1 \ . \tag{3.16}$$

es decir, (3.4) se cumple al tiempo t_{n+1}^+.

Caso 2: Supongamos que (3.5) se cumple al tiempo t_n^+.

Se sigue que (3.5) también se cumple al tiempo t_{n+1}^+, como se muestra en los siguientes argumentos. Otra vez, supongamos primero que t_{n+1} corresponde a una llegada,

$\underline{t_{n+1} \in \{A_k\}}$.
Dado que las longitudes de colas son idénticas en ambos $\mathcal{X}$ y $\widetilde{\mathcal{X}}$, como son las reglas de decisión para C_r, la decisión de entrar o no entrar a Qs será la misma en ambos sistemas. Por lo tanto

$$\mathrm{X}(t_{n+1}^+) = \widetilde{\mathrm{X}}(t_{n+1}^+) \; . \tag{3.17}$$

Una vez más, ya que C_r se forma detrás de C_0 en caso de que efectivamente ingrese a Qs, entonces los estados de los procesos-RST permanecen invariantes.

$\underline{t_{n+1}^+ \in \{S_l\}}$
Ya que $x_m = \widetilde{x}_m$, tenemos que $\mu(x_m) = \mu(\widetilde{x}_m)$. Si $\mathrm{U}'_m \leq \mu(x_m)/\overline{\mu}$, entonces ocurre una salida real bajo ambos $\mathcal{X}$ y $\widetilde{\mathcal{X}}$. Por lo tanto

$$\mathrm{X}(t_{n+1}^+) = \mathrm{X}(t_n^+) - 1 = \widetilde{\mathrm{X}}(t_n^+) - 1 = \widetilde{\mathrm{X}}(t_{n+1}^+) \; , \tag{3.18}$$
$$\mathrm{Z}(t_{n+1}^+) = \mathrm{Z}(t_n^+) - 1 = \widetilde{\mathrm{Z}}(t_n^+) - 1 = \widetilde{\mathrm{Z}}(t_{n+1}^+) \; . \tag{3.19}$$

Si $\mathrm{U}'_m > \mu(x_m)/\overline{\mu}$, entonces no hay salidas reales bajo ambos $\mathcal{X}$ y $\widetilde{\mathcal{X}}$, por lo que no hay cambios ya sea en la longitud de la cola o en el proceso-RST, es decir,

$$\mathrm{X}(t_{n+1}^+) = \mathrm{X}(t_n^+) = \widetilde{\mathrm{X}}(t_n^+) = \widetilde{\mathrm{X}}(t_{n+1}^+) \; , \tag{3.20}$$
$$\mathrm{Z}(t_{n+1}^+) = \mathrm{Z}(t_n^+) = \widetilde{\mathrm{Z}}(t_n^+) = \widetilde{\mathrm{Z}}(t_{n+1}^+) \; . \tag{3.21}$$

Caso 3: Supongamos que (3.6) se cumple al tiempo t_n^+.

Se sigue que (3.6) también se cumple al tiempo t_{n+1}, con los argumentos que se presentan a continuación. Una vez más, supongamos primero que t_{n+1} corresponde a una llegada,

$\underline{t_{n+1} \in \{A_k\}}$.
Por la misma razón como en el caso anterior, las longitudes de las colas permanecen iguales, es decir

$$\mathrm{X}(t_{n+1}^+) = \widetilde{\mathrm{X}}(t_{n+1}^+) \; , \tag{3.22}$$

y no hay cambios en los procesos-RST.

Finalmente, supongamos que t_{n+1} corresponde a una salida,

$\underline{t_{n+1}^+ \in \{S_l\}}$.
Ya que $x_m = \widetilde{x}_m$, tenemos que $\mu(x_m) = \mu(\widetilde{x}_m)$. Otra vez, como en los casos previos, si $\mathrm{U}'_m \leq \mu(x_m)/\overline{\mu}$, entonces

$$\mathrm{X}(t_{n+1}^+) = \mathrm{X}(t_n^+) - 1 = \widetilde{\mathrm{X}}(t_n^+) - 1 = \widetilde{\mathrm{X}}(t_{n+1}^+) \ , \tag{3.23}$$

$$\mathrm{Z}(t_{n+1}^+) = \mathrm{Z}(t_n^+) - 1 = (\widetilde{\mathrm{Z}}(t_n^+) - 1) - 1 = \widetilde{\mathrm{Z}}(t_{n+1}^+) - 1 \ . \tag{3.24}$$

Por otro lado, si $\mathrm{U}'_m > \mu(x_m)/\overline{\mu}$, entonces no hay cambios ya sea en la longitud de la cola o en el proceso-RST. □

Recordando la definición del tiempo de permanencia de C_0 en Qs, el siguiente lema se tiene ahora.

Lema 3.4.2 *Para todo $\delta \in \mathbb{T}^\infty$ y $k \in \mathbb{N}$, $V_k(x, \delta)$ es estrictamente creciente en x, en el sentido de que para $x \in \{1, 2, ..., B-2\}$, existe una constante $\{\varepsilon_x\}$ tal que*

$$V_k(x+1, \delta) - V_k(x, \delta) \geq \varepsilon_x > 0 \ ,$$

independientemente de δ.

Demostración: Sin perdida de generalidad, y para ser específicos, consideremos al cliente C_0 y a los sistemas $\mathcal{X}$ y $\widetilde{\mathcal{X}}$ bajo el acoplamiento $\mathcal{C}$. A partir de las definiciones de v_0 y $\widetilde{v}_0$, y el lema 3.4.1, $v_0 \leq \widetilde{v}_0$ lo cual implica que $E[v_0] \leq E[\widetilde{v}_0]$. Para establecer la desigualdad, definimos el evento I_x:

$$I_x = \{S_{x+1} < A_1, \mathrm{U}'_m \leq \mu(x_m)/\overline{\mu}, m = 1, ..., x+1\} \ .$$

Este es el caso en el que el cliente C_1 llega después que la $(x+1)$-énesima salida bajo ambos sistemas, donde C_0 deja bajo $\mathcal{X}$ al tiempo S_{x+1}, pero se convierte en el único cliente que queda en Qs bajo $\widetilde{\mathcal{X}}$ en ese momento. Al condicionar en este evento, note que $v_0 < \widetilde{v}_0$ en I_x y que $P(I_x) > 0$, entonces el resultado se sigue. □

3.5. Monotonicidad y continuidad con respecto a políticas umbrales

En esta sección se examina el comportamiento del tiempo de permanencia en Qs con respecto a un cierto tipo de regla umbral, que se introduce a continuación.

Definición 3.5.1 *Para $L \in \mathbb{N}$ y $q \in [0, 1]$, una $[L, q]$-regla de decisión umbral $u(\cdot)$ está definida como sigue:*

$$u(x) = \begin{cases} 1 & x < L \ , \\ q & x = L \ , \\ 0 & x > L \ . \end{cases} \tag{3.25}$$

Esto puede ser representado de manera más compacta por $[L, q]$ o en efecto $[g]$, donde $g = L + q$. Cuando $B < \infty$ cualquier $[g]$, con $g > B$ es equivalente a $[B]$.

Estamos interesados en este caso, en la caracterización de las políticas *simétricas*: éstas son políticas en las que cada cliente adopta la misma regla de decisión. Así, si tenemos una política δ en la cual la regla de decisión para cada cliente está dada por $[g]$, esta se denota por $[g]^\infty$: llamamos a ésta, una *política umbral simétrica*.

A continuación presentamos otros dos "sistemas" que facilitarán las pruebas de los resultados de esta sección.

Definición 3.5.2 *(**Sistema** $\mathcal{G}$) (I), (II), (IV), (V) y (VI) son exactamente como en el sistema $\mathcal{X}$, sin embargo (III) se convierte en:*
(III): El cliente C_0 entra a Qs al tiempo $A_0 = 0$, con todos los clientes adheridos a la política $[g]^\infty$, donde $g \in [0, B)$.

Definición 3.5.3 *(**Sistema** $\widetilde{\mathcal{G}}$) $(\widetilde{I})$, $(\widetilde{II})$ y $(\widetilde{VI})$ son precisamente los mismos que para $\widetilde{\mathcal{X}}$;*
$(\widetilde{III})$: el cliente C_0 entra a Qs al tiempo $A_0 = 0$, con todos los clientes adheridos a la política $[\widetilde{g}]^\infty$, con $g \leq \widetilde{g} \leq B$ (donde la última desigualdad es estricta si $B = \infty$);
$(\widetilde{IV})$: el proceso de longitud de cola $\{\widetilde{\mathrm{X}}(t) : t \geq 0\}$ con $\widetilde{\mathrm{X}}(0) = x$;
$(\widetilde{V})$: el proceso-RST $\{\widetilde{\mathrm{Z}}(t) : t \geq 0\}$ con $\widetilde{\mathrm{Z}}(0) = x$.

Esto implica de la definición de g y $\tilde{g}$ que $L < B$. De aquí en adelante, $g \in [0, B]$ se entenderá como: $0 \leq g \leq B$ cuando B es finito y $0 \leq g < B$ cuando B es infinito, a menos que se especifique lo contrario.

Los siguientes dos resultados serán utilizados para deducir algunas consecuencias sobre $V_k(\cdot, [g]^\infty)$ en los intervalos $[0, 1]$ y $[1, B]$.

Lema 3.5.1 *Para el sistema $(\mathcal{G}, \widetilde{\mathcal{G}})$ bajo el acoplamiento $\mathcal{C}$, el siguiente conjunto de relaciones se mantiene en cada momento $t \in \mathbb{R}^+$,*

$$\mathrm{X}(t) \leq \widetilde{\mathrm{X}}(t), \tag{3.26}$$

$$Z(t) \geq \widetilde{\mathrm{Z}}(t). \tag{3.27}$$

Demostración: Suponga que t_{n+1} corresponde a un tiempo de arribo, donde $t_{n+1} = A_r$, o un tiempo potencial de salida, donde $t_{n+1} = S_m$, tal que C_0 todavía está presente en Qs bajo $\mathcal{G}$.
Primero note que

$$Z(0^+) = X(0^+) = x + 1 = \widetilde{\mathrm{X}}(0^+) = \widetilde{\mathrm{Z}}(0^+) \, .$$

Ahora suponga que

$$\mathrm{X}(t_n^+) \leq \widetilde{\mathrm{X}}(t_n^+), \tag{3.28}$$
$$Z(t_n^+) \geq \widetilde{\mathrm{Z}}(t_n^+). \tag{3.29}$$

Caso 1: La relación (3.26) es estricta
$\underline{t_{n+1} \in \{A_k\}}$
Si $\widetilde{\mathrm{X}}(t_n^+) < B$ entonces C_r entra a Qs bajo ninguna, una, o ambos sistemas $\mathcal{G}$ y $\widetilde{\mathcal{G}}$; resulta que

$$\mathrm{X}(t_{n+1}^+) \leq \widetilde{\mathrm{X}}(t_{n+1}^+)\ , \tag{3.30}$$

(la cual se cumple con igualdad cuando C_r entra a Qs bajo $\mathcal{G}$ solamente y $\mathrm{X}(t_n^+) = \widetilde{\mathrm{X}}(t_n^+) - 1$).

Si $\widetilde{\mathrm{X}}(t_n^+) = B$ entonces C_r puede solo entrar a Qs bajo $\mathcal{G}$. Aquí

$$\mathrm{X}(t_{n+1}^+) \leq \widetilde{\mathrm{X}}(t_{n+1}^+) = B.$$

Dado que C_r nunca puede colocarse por delante de C_0 en cualquiera de estos escenarios, entonces no puede haber ningún cambio en los procesos-RST, de modo que

$$\mathrm{Z}(t_{n+1}^+) = \mathrm{Z}(t_n^+) \geq \widetilde{\mathrm{Z}}(t_n^+) = \widetilde{\mathrm{Z}}(t_{n+1}^+). \tag{3.31}$$

$\underline{t_{n+1}^+ \in \{S_l\}}$
Ya que $x_m < \widetilde{x}_m$, entonces $\mu(x_m) < \mu(\widetilde{x}_m)$.
Si $\mathrm{U}'_m \leq \mu(x_m)/\overline{\mu}$, entonces una salida real ocurre bajo ambos $\mathcal{G}$ y $\widetilde{\mathcal{G}}$ y también

$$\mathrm{X}(t_{n+1}^+) = \mathrm{X}(t_n^+) - 1 < \widetilde{\mathrm{X}}(t_n^+) - 1 = \widetilde{\mathrm{X}}(t_{n+1}^+), \tag{3.32}$$
$$\mathrm{Z}(t_{n+1}^+) = \mathrm{Z}(t_n^+) - 1 \geq \widetilde{\mathrm{Z}}(t_n^+) - 1 = \widetilde{\mathrm{Z}}(t_{n+1}^+). \tag{3.33}$$

Si $\mathrm{U}'_m \in (\mu(x_m)/\overline{\mu}, \mu(\widetilde{x}_m)/\overline{\mu}]$, entonces una salida real ocurre bajo $\widetilde{\mathcal{G}}$ pero no bajo $\mathcal{G}$, y también

$$\mathrm{X}(t_{n+1}^+) = \mathrm{X}(t_n^+) \leq \widetilde{\mathrm{X}}(t_n^+) - 1 = \widetilde{\mathrm{X}}(t_{n+1}^+), \tag{3.34}$$
$$\mathrm{Z}(t_{n+1}^+) = \mathrm{Z}(t_n^+) \geq \widetilde{\mathrm{Z}}(t_n^+) - 1 = \widetilde{\mathrm{Z}}(t_{n+1}^+). \tag{3.35}$$

Si $\mathrm{U}'_m > \mu(\tilde{x}_m)/\overline{\mu}$, entonces no hay cambios, i.e., (3.25) se mantiene estrictamente y (3.26) también mantiene al tiempo t_{n+1}^+.

Caso 2: La relación (3.27) se cumple con la igualdad
$\underline{t_{n+1} \in \{A_k\}}$
Aquí, o C_r entra a Qs bajo ninguno, o bajo ambos sistemas $\mathcal{G}$ y $\widetilde{\mathcal{G}}$, o bien sólo por $\widetilde{\mathcal{G}}$. Por lo tanto

$$\mathrm{X}(t_{n+1}^+) \leq \widetilde{\mathrm{X}}(t_{n+1}^+). \tag{3.36}$$

Como en el caso anterior, C_r no puede se residir delante de C_0 y así una vez más, (3.26) tiene al tiempo t_{n+1}^+.
$\underline{t_{n+1} \in \{S_l\}}$
ya que $x_m = \widetilde{x}_m$, entonces $\mu(x_m) = \mu(\widetilde{x}_m)$.
Si $\mathrm{U}'_m \leq \mu(x_m)/\overline{\mu}$, entonces una salida real ocurre bajo ambos $\mathcal{G}$ y $\widetilde{\mathcal{G}}$, y también

$$\mathrm{X}(t_{n+1}^+) = \mathrm{X}(t_n^+) - 1 = \widetilde{\mathrm{X}}(t_n^+) - 1 = \widetilde{\mathrm{X}}(t_{n+1}^+), \tag{3.37}$$
$$\mathrm{Z}(t_{n+1}^+) = \mathrm{Z}(t_n^+) - 1 \geq \widetilde{\mathrm{Z}}(t_n^+) - 1 = \widetilde{\mathrm{Z}}(t_{n+1}^+). \tag{3.38}$$

Si $\mathrm{U}'_m > \mu(x_m)/\overline{\mu}$, entonces no hay cambios, es decir (3.25) se cumple con la igualdad y (3.26) tiene al tiempo t_{n+1}^+. □

Lema 3.5.2 *Para el sistema ($\mathcal{G}$, $\widetilde{\mathcal{G}}$) bajo el acoplamiento $\mathcal{C}$, con $0 < g < \tilde{g} \leq 1$, el siguiente conjunto de relaciones se mantiene en cada momento $t \in \mathbb{R}^+$ durante toda la permanencia de C_0 (en cualquier sistema)*

$$\mathrm{X}(t) = \widetilde{\mathrm{X}}(t), \tag{3.39}$$
$$\mathrm{Z}(t) = \widetilde{\mathrm{Z}}(t). \tag{3.40}$$

Demostración: Ya que $\mathrm{X}(0) = x = \widetilde{\mathrm{X}}(0)$, llegan a Qs, salidas reales y ficticias coincidirán en los dos sistemas durante la estancia de C_0 por lo menos hasta el momento en que ocurra una desigualdad en la decisión de llegada. Sin embargo, ya que $0 < g < \tilde{g} \leq 1$, entonces la primera oportunidad de que un cliente entre a Qs bajo $\widetilde{\mathcal{G}}$ pero no bajo $\mathcal{G}$ es cuando las colas están completamente vacías, pero C_0 obviamente ha abandonado por ese momento al sistema. Por lo tanto, como consecuencia las relaciones (3.38) y (3.39) se llevarán a cabo. □

A continuación, se define la siguiente cantidad para el análisis posterior, donde $[\widetilde{g}] = [\widetilde{L}, \widetilde{q}]$:

$$\widehat{q} := 1 - (1 - \widetilde{q})\mathbf{1}_{\{L=\widetilde{L}\}}. \tag{3.41}$$

Lema 3.5.3 *Para cada $k \in \mathbb{N}$, y $x \in \{0, 1, ..., B-1\}$*
(i) $V_k(x, [g]^\infty)$ es constante en g sobre [0,1] y
(ii) $V_k(x, [g]^\infty)$ es estrictamente decreciente en g sobre [1,B].

Demostración: Sin perdida de generalidad, y siendo específicos consideremos el cliente C_0 y los sistemas $\mathcal{G}$ y $\widetilde{\mathcal{G}}$ bajo el acoplamiento $\mathcal{C}$. De las definiciones de v_0 y $\tilde{v}_0$, e invocando el lema 3.5.1, tenemos que $v_0 \geq \tilde{v}_0$, por lo que $E[v_0] \geq E[\tilde{v}_0]$. Luego $E[v_0] = E[\tilde{v}_0]$ cada vez que $0 < g < \tilde{g} \leq 1$ por el lema 3.5.2, estableciendo así (i).

Ahora, supongamos que $1 \leq g < \tilde{g}$ y definiendo los siguientes eventos:

Para $x < L$:

$$\begin{aligned} F_\alpha = \{&A_{L-x} < S_1; A_{L-x+1} > S_{x+1}; U_{L-x} \in (q, \widehat{q}\,]; \\ &U_1' \in (\mu(X(S_1))/\mu, \mu(\tilde{X}(S_1))/\overline{\mu}\,]; \\ &U_m' \geq \mu(X(S_m))/\overline{\mu} : m = 2, , ..., x+1\}. \end{aligned}$$

Cualquier realización de los sistemas en F_α bajo el acoplamiento de $\mathcal{C}$ da como resultado los siguientes acontecimientos, en el orden indicado a continuación:

1. $L - x - 1$ clientes entran a Qs bajo ambos sistemas $\mathcal{G}$ y $\widetilde{\mathcal{G}}$, dando como resultado el aumento del tamaño de cola hasta L y C_0 queda en la posición $x + 1$ en ambos casos.
2. Un cliente entra a Qs bajo $\widetilde{\mathcal{G}}$, pero no bajo $\mathcal{G}$. Esto da como resultado que el tamaño de la cola bajo $\widetilde{\mathcal{G}}$ aumenta desde L a $L + 1$, pero permanece en L bajo $\mathcal{G}$, con C_0 todavía en la posición $x + 1$ en ambos casos.
3. Una salida ocurre bajo $\widetilde{\mathcal{G}}$, pero no bajo $\mathcal{G}$. Esto da como resultado que el tamaño de la cola es igual a L en ambos casos, C_0 permanece en la posición $x + 1$ bajo $\mathcal{G}$, y C_0 se mueve a la posición x (o de hecho sale del sistema si $x = 0$) bajo $\widetilde{\mathcal{G}}$.
4. Otras x salidas ocurren bajo ambos sistemas (antes del próximo arribo), dando como resultado que C_0 sale bajo $\widetilde{\mathcal{G}}$ (si todavía está presente), pero residiendo al inicio de las colas bajo $\mathcal{G}$, al tiempo S_{x+1}.

Para $L \leq x < B$:

$$\begin{aligned} F_\beta = \{&S_{x-L+1} < A_1 < S_{x-L+2}; S_{x+1} < A_2; U_1 \in (q, \widehat{q}\,]; \\ &U_m' \in (\mu(X(S_m))/\overline{\mu} : m = 1, ..., x+1, m \neq x-L+2; \\ &U_m' \in (\mu(X(S_1))/\mu, \mu(\tilde{X}(S_1))/\overline{\mu}] : m = x-L+2\}, \end{aligned}$$

Cualquier realización de los sistemas en F_β bajo el acoplamiento de $\mathcal{C}$ da como resultado los siguientes acontecimientos, en el orden en que se presentan.

1. El tamaño de la cola bajo ambos sistemas $\mathcal{G}$ y $\widetilde{\mathcal{G}}$ decrece desde $x + 1$ hasta L, y C_0 se mueve a la posición L en ambos casos.

2. Un cliente entra a Qs bajo $\widetilde{\mathcal{G}}$, pero no bajo $\mathcal{G}$; dando como resultado que el tamaño de la cola sea igual a L y $L+1$ bajo $\mathcal{G}$ y $\widetilde{\mathcal{G}}$ respectivamente; C_0 permanece en la posición L en ambos casos.

3. Una salida ocurre bajo $\widetilde{\mathcal{G}}$ pero no bajo $\mathcal{G}$. Esto da como resultado que el tamaño de la cola sea igual a L en ambos casos, C_0 permanece en la posición $x+1$ bajo $\mathcal{G}$, y C_0 se mueve a la posición x (o incluso sale del sistema si $x=0$) bajo $\widetilde{\mathcal{G}}$.

4. Otras $L-1$ salidas ocurren bajo ambos sistemas $\mathcal{G}$ y $\widetilde{\mathcal{G}}$ después del segundo arribo, resultando que C_0 sale bajo $\widetilde{\mathcal{G}}$ (si todavía está presente) pero residiendo al inicio de la cola bajo $\mathcal{G}$, al tiempo S_{x+1}.

Defina

$$F_\zeta = F_\alpha \, \mathbf{1}_{\{x<L\}} + F_\beta \, \mathbf{1}_{\{x\geq L\}}.$$

Por el condicionamiento de F_ζ, se establece que $E[v_0] > E[\widetilde{v}_0]$ y se sigue (ii). □

Observación
En la prueba anterior, para mayor claridad de la exposición, $\mu(x_m)$ y $\mu(\tilde{x}_m)$ se han escrito más explícitamente como $\mu(X(S_m))$ y $\mu(\widetilde{X}(S_m))$ con el fin de evitar confusiones con x que aparece en la indicación de las variables aleatorias.

A continuación se muestra que el valor esperado del tiempo de permanencia de un cliente en Qs, cuando otros clientes se adhieren a la regla de decisión $[g]$, en particular para x estados de entrada, es una función continua de g. Este resultado está establecido bajo la siguiente condición no estricta, la cual se presenta a continuación.

Condición de Estabilidad (CE): Bajo el sistema $\mathcal{G}$, existe una cota D_x, tal que

$$E\left[\sum_{k=1}^{\infty} \mathbf{1}_{\{A_k \leq v_0\}}\right] \leq D_x,$$

uniformemente en $g \in [0, B]$.

Esta condición dice que el número esperado de las llegadas que se producen durante la estancia de C_0 en Qs cuando los clientes adoptan la regla de decisión umbral $[g]$, está acotada superiormente por D_x, sobre todo $g \in [0, B]$. Esta condición ciertamente se cumple cuando los **tiempos de inter-arribos**, los cuales están dados por $\{M_i\}$, **son exponenciales**. Para ver esto, considere un sistema similar a $\mathcal{G}$, excepto que la tasa de servicio es siempre $\mu(1)$. Indicaremos por un subíndice "$*$" a los valores asociados a este sistema $\mathcal{G}$ con tasa de servicio fija $\mu(1)$. Bajo un acoplamiento similar a $\mathcal{C}$ entre este sistema y $\mathcal{G}$ se puede demostrar que $v_0 \leq v_0^*$ para todo $g \in [0, B]$. Sin embargo, es también claro que v_0^* solo depende de las $\{N_j\}$ y de las $\{U_l'\}$ y por lo tanto es independiente de los instantes de llegada $\{A_k\}$.

Por lo tanto

$$E\left[\sum_{k=1}^{\infty}\mathbf{1}_{\{A_k\leq v_0\}}\right]\leq E\left[\sum_{k=1}^{\infty}\mathbf{1}_{\{A_k\leq v_0^*\}}\right] = \int_0^{\infty}E\left[\sum_{k=1}^{\infty}\mathbf{1}_{\{A_k\leq t\}}\right]f_{v_0^*}(t)dt$$
$$= \lambda\int_0^{\infty}tf_{v_0^*}(t)dt$$
$$= \lambda\, E[v_0^*],$$

donde $f_{v_0^*}(\cdot)$ es la función de densidad de probabilidad de v_0^*. Sin embargo $E[v_0^*]$ es justamente el valor esperado de la suma de $x+1$ variables aleatorias *i.i.d.* exponenciales cada una con media $\mu(1)^{-1}$. Así,

$$E\left[\sum_{k=1}^{\infty}\mathbf{1}\{A_k\leq v_0\}\right]\leq\frac{\lambda(x+1)}{\mu(1)}.$$

Lema 3.5.4 *Supongamos que la condición de estabilidad CE (los tiempos de inter-arribos son exponenciales) se mantiene, entonces para todo $k\in\mathbb{N}$, $x\in\{0,1,...,B-1\}$, $V_k(x,[g]^{\infty})$ es continua en $g\in[0,B]$.*

Demostración: Sin perdida de generalidad, y siendo específicos, considere al cliente C_0 y los sistemas $\mathcal{G}$ y $\widetilde{\mathcal{G}}$ bajo el acoplamiento $\mathcal{C}$, con la restricción adicional que $g=L+q$ y $\widetilde{g}=L+\widetilde{q}$ tal que $0\leq q<\widetilde{q}\leq 1$. Defina

$$k_0=\text{ínf}\{k\in\mathbb{Z}^+ : A_k<v_0,\ X(A_k)=L,\ U_k\in(q,\tilde{q}]$$

donde $\text{ínf}\,\emptyset:=\infty$. En efecto si $k_0=\infty$, entonces $v_0=\tilde{v}_0$.
Para $k_0=k<\infty$, entonces $\tilde{v}_0>A_k$ y

$$E\left[\tilde{v}_0-v_0|k_0=k\right]=E\left[\tilde{v}_0-A_k|k_0=k\right]-E\left[v_0-A_k|k_0=k\right]\leq\frac{L+1}{\mu(1)},$$

esto se deriva del hecho de que el primer término está acotado por arriba por la esperanza de tiempo de servicio a $L+1$ clientes a la tasa más lenta posible $\mu(1)$, y que el segundo término está acotado por cero.

Es fácil deducir que $\{A_k<v_0\}$ es independiente de $\{U_k\in(q,\tilde{q}]\}$ (señalando que el primero es equivalente a $\{Z(A_k)>0\}$). Para ello,

$$\begin{aligned} P(k_0 = k) &\leq P(A_k < v_0, U_k \in (q, \widetilde{q}]) \\ &= P(A_k < v_0) P(U_k \in (q, \widetilde{q}]) \\ &= (\widetilde{q} - q) P(A_k < v_0) \end{aligned}$$

y por lo tanto,

$$\begin{aligned} E\left[\tilde{v}_0 - v_0\right] &\leq \frac{L+1}{\mu(1)} \sum_{k=1}^{\infty} \mathbb{P}(k_0 = k) \\ &\leq (q - \tilde{q}) \frac{L+1}{\mu(1)} \sum_{k=1}^{\infty} \mathbb{P}(A_k < \tilde{v}_0). \end{aligned}$$

Por el teorema de convergencia monótona, $\sum_{k=1}^{\infty} \mathbb{P}(A_k < \tilde{v}_0)$ puede ser expresada como

$$E\left[\sum_{k=1}^{\infty} \mathbf{1}\{A_k \leq v_0\}\right].$$

Además observe que $\widetilde{q} - q = \widetilde{g} - g$. Así

$$E\left[\tilde{v}_0 - v_0\right] \leq (\widetilde{g} - g) \frac{L+1}{\mu(1)} D_x,$$

como es requerido. □

3.6. Estructura y existencia del equilibrio de Nash

En esta sección en primer lugar exploraremos la estructura necesaria para ser candidato a una *política de equilibrio de Nash simétrica* (SNEP) dentro de la clase $\mathbb{T}^{\infty}$. A continuación se demuestra que existe un número finito de SNEPs dentro de esta clase, y que al menos uno de estos se caracteriza por un umbral no aleatorio.

Lema 3.6.1 *Supongamos que $\delta \in \mathbb{T}^{\infty}$, entonces*
(a) cualquier regla de decisión óptima de C_k respecto a δ es una regla de decisión umbral;

(b) el conjunto de reglas de decisión óptimas de C_k respecto a δ, es decir $\mathbb{U}_k(\delta)$, se puede encontrar de la siguiente manera: donde $\widehat{L} := \min\{L \in \mathbb{Z}^+ : V_k(L, \delta) \geq \theta\}$:
(i) Si $V_k(\widehat{L}, \delta) = \theta$, entonces $\mathbb{U}_k(\delta) = \{[\widehat{L}, q] : 0 \leq q \leq 1\}$;
(ii) en otro caso, $\mathbb{U}_k(\delta) = \{[\widehat{L}, 0]\}$.

Demostración: De la definición de una regla de decisión óptima y el resultado de la monotonicidad del lema 3.4.2, ya que $V_k(L,\delta) \geq (L+1)/\overline{\mu} \longrightarrow \infty$ cuando $(L \longrightarrow \infty)$, entonces $\widehat{L}$ está bien definido.

En la siguiente discusión, el mejor plano de respuesta de un cliente arbitrario, por ejemplo C_k se construye en el contexto de otros clientes que se adhieren a la política $[g]^\infty$. Por simplicidad, sin perdida de generalidad, construimos este plano para el cliente C_0; esto es como decir que C_k corresponde a C_0 en la corriente de llegada total. A continuación se define el plano $l(\cdot)$ como sigue:
Para $g \in [0,B]$, sea

$$l(g) = \text{mín}\{n \in \mathbb{N} : n < B, V_0(n,[g]^\infty) \geq \theta\}, \tag{3.42}$$

con $\text{mín}\,\emptyset = B$. Ya que $V_0(x,\cdot) \geq x + 1/\overline{\mu} \longrightarrow \infty$ cuando $x \longrightarrow \infty$, entonces $l(\cdot)$ está bien definido.

Además, sean $\{g_1, g_2, ..., g_J\}$ los puntos de discontinuidad de $l(g)$ para $g \in [0,B]$, donde

$$0 := g_0 \leq g_1 < g_2 < \cdots < g_{J-1} < g_J \leq g_{J+1} := B.$$

Note que $g_1 = 0$ si hay un punto de discontinuidad en el origen y $g_J = B$ si hay uno en B cuando B es finito.

Siguiendo la metodología similar a la de [3]. Definimos la función del punto al conjunto potencia $G^*(g) : [0,B] \longrightarrow 2^{[0,B]}$, que se asocia con cada umbral $g \in [0,B]$ el conjunto de "umbrales óptimos" contra $[g]^\infty$, a saber

$$G^*(g) = \{g\prime \in [0,B] : [g'] \textit{ optima para } C_0 \textit{ contra } [g]^\infty\}. \tag{3.43}$$

Ya que $[g]^\infty$ es un miembro de $\mathbb{T}^\infty$, entonces podemos invocar el lema 3.6.1 y deducir que $G^*(\cdot)$ está dado por

$$G^*(g) = \begin{cases} \{l(g) + q : 0 \leq q \leq 1\} & V_0(l(g),[g]^\infty) = \theta,\ l(g) < B\ , \\ l(g) & V_0(l(g),[g]^\infty) > \theta,\ l(g) < B\ , \\ B & l(g) = B. \end{cases} \tag{3.44}$$

Por las propiedades antes mencionadas de $V_0(\cdot,[g]^\infty)$, y siempre que CE se mantiene, puede deducirse que

$$l(g) = l(0), \qquad 0 \leq g \leq 1; \tag{3.45}$$

$$l(g) = l(0) + j, \qquad g > 1,\ g_j < g \leq g_{j+1}. \tag{3.46}$$

Así $G^*(g)$ puede ser re-escrito como:

$$G^*(g) = \begin{cases} l(0) & g < g_1 \\ [l(0) + j - 1,\ l(0) + j] & g = g_j,\ j = 1, ..., J \\ [l(0),\ l(0) + j] & g_0 = g_1 \text{ y } g \leq 1 \\ l(0) + j & g_j < g < g_{j+1},\ j = 2, ..., J \\ l(0) + 1 & g_0 < g_1 < g < g_2,\ o\ g_0 = g_1 \text{ y } 1 < g < g_2. \end{cases} \tag{3.47}$$

Observe que la gráfica de $G^*(\cdot)$ es una escalera conectada y no decreciente cuando $g_0 < g_1$; una estructura similar cuando $g_0 = g_1$, excepto que se produce una región rectangular con coordenadas inferior izquierda y superior derecha dadas por (0, $l(0)$) y (1, $l(0)$+1), respectivamente.

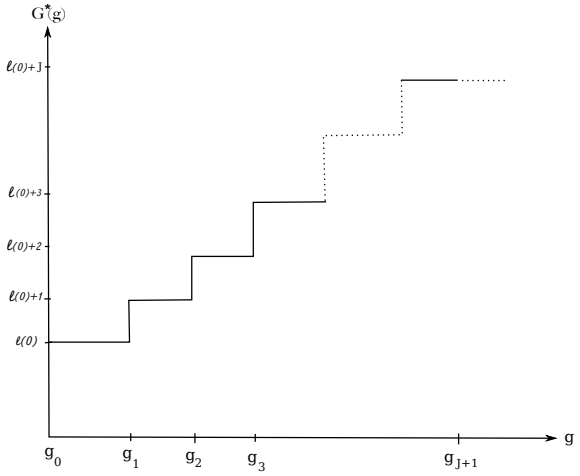
Defina el mapeo $H(g) := G^*(g) - g$ con el mismo dominio como $G(\cdot)$. Aquí $G^*(g) - g$ se entiende por $[\text{mín}\{G^*(g)\} - g, \text{máx}\{G^*(g) - g\}]$. La gráfica de $H(\cdot)$ tiene un "diente de sierra" como estructura, excepto que en el caso cuando $g_0 = g_1$, este se modifica para incluir un rombo en el intervalo [0,1] con coordenadas $(0, l(0))$, $(0, l(0) + 1)$, $(1, l(0) - 1)$ y $(1, l(0))$. Veamos esta construcción en la demostración del siguiente teorema.

Teorema: *Supongamos que la **CE** (los tiempos de inter-arribos son exponenciales) se cumple, entonces en la clase de políticas* $\mathbb{T}^\infty$*,*

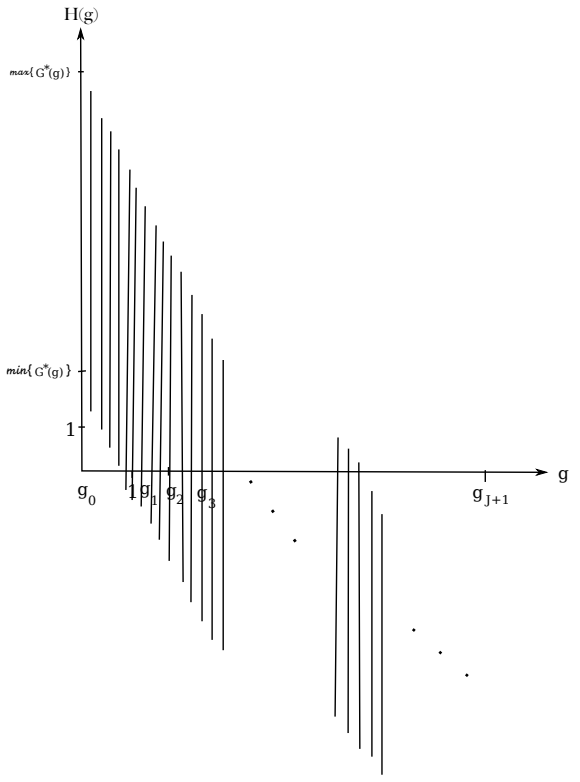
(i) existe un número finito de SNEPs;
(ii) al menos uno de los SNEPs es no aleatorio.

Demostración: Los umbrales asociados con los SNEPs corresponden a los ceros del mapeo $H(\cdot)$. Por el lema 3.5.3 (i), y usando el hecho de que $l(0) \geq 1$, tenemos que $\text{mín}\{H(g)\} > 0$ para todo $g \in [0, 1)$. Así, restringimos muestra discusión a $g \in [1, B]$.

Supongamos que B es finito. Entonces, ya que $\text{mín}\{G^*(1)\} \geq 1$ y $\text{máx}\{G^*(B)\} \leq B$, tenemos $\text{mín}\{H(1)\} \geq 0$ y $\text{máx}\{H(B)\} \leq 0$. Así, por la estructura de la gráfica de $H(\cdot)$, y el Teorema de valor intermedio, debe existir un $g^* \in [1, B]$ para el cual $0 \in H(g^*)$.



Ahora supongamos que B es infinito. Notemos que $V_0(x,[g]^\infty)$ esta acotada por abajo de $(x+1)/\overline{\mu}$, el cual es independiente de g, y que $(x+1)/\overline{\mu} \longrightarrow \infty$ cuando $x \longrightarrow \infty$. Así, existe algún $n \in \mathbb{Z}^+$ para el cual $V_0(n,[g]^\infty) > \theta$ para todo $g \geq 0$. Esto implica que $l(g) \leq n$ para todo $g \geq 0$ y, así el número de saltos de $G^*(\cdot)$ (y $H(\cdot)$), el cual está dado por J debe ser finito. Por lo tanto, $G^*(g) = l(g_J) + 1 < \infty$ para todo $g > g_J$, el cuál implica que $H(g) < 0$ para todo g suficientemente grande. A partir de esto, junto con el hecho de que mín$\{H(1)\} \geq 0$, se deduce que existe un $g^{**} \in [1,\infty)$ tal que $0 \in H(g^{**})$.



Como se ha comentado anteriormente, ya que el número J de saltos verticales es finito, el número de ceros de $H(\cdot)$ está acotado (para B finito e infinito) y, así existe un número finito de SNEPs, estableciéndose la parte (i).

Supongamos ahora, en contradicción con la parte (ii), que no hay SNEPs no aleatorios. Por la parte (i), al menos un SNEP aleatorio existe. Tal punto debe corresponder a un punto de salto de $H(\cdot)$. En efecto, un SNEP ocurre en $g = g_m$ si y sólo si un cero que es estrictamente interior a la parte vertical de la gráfica en g_m, es decir mín$\{H(g_m)\} < 0 <$ máx$\{H(g_m)\}$, también ocurre. Sin embargo, sabemos que el mín$\{H(1)\} \geq 0$; así por el Teorema de valor intermedio, existe al menos un punto $g' \in [1, g_m)$, en una sección diagonal de la gráfica (que se toma para incluir esquinas), para el cual $0 \in H(g')$. Todo punto tal debe corresponder a un umbral no aleatorio, proporcionando la contradicción necesaria para la parte (ii). □

3.7. Conclusiones

En este capítulo, fuimos capaces de aplicar los conceptos y definiciones de la teoría de juegos a cierta clase de colas simples, pues gracias a ello se estableció la existencia de una política con cierto *equilibrio simétrico de Nash* para los clientes que se unen a una cola con tasa de servicio creciente basada en el tamaño de la cola observable, al llegar al sistema. Se examinó el comportamiento del tiempo de permanencia en el sistema Qs con respecto a un cierto tipo de regla umbral.

Además, hemos demostrado que al menos uno de los *equilibrios de Nash* es no aleatorio. Este es un fenómeno algo diferente a la exhibido en otros trabajos, (por ejemplo [3]). Más aún, el valor constante del tiempo de permanencia estimado en g $(V_k(x, [g]^\infty))$ en la región [0,1], y la monotonicidad estricta en el intervalo [1, B], se estableció como un escalón fundamental para resultados en la teoría de juegos.

Por último, es importante hacer mención que cuando un único (no aleatorio) *SNEP* del juego estacionario existe en la clase $\mathbb{T}^\infty$, sería interesante adquirir información de cantidades tales como las medias empíricas y probabilidades con ayuda de la simulación (esto se deja como problema a futuro).

CAPÍTULO 4

Cálculo de probabilidades estacionarias para el modelo de dos colas en paralelo

4.1. Introducción

Un sistema de espera con dos colas (o filas) en paralelo puede ser analizado extensamente. Una cola (o fila) se produce cuando la demanda de un servicio por parte de los clientes excede la capacidad del servicio. En tales casos, debe ser posible formular una estrategia a seguir para los clientes con el fin de reducir el tiempo de permanencia en el sistema. La estrategia estudiada en esta parte del trabajo son las maniobras. Este tipo de maniobras puede ser descrito como el *"salto"* que un cliente en espera realiza, de una cola a otra. Por lo que en este capítulo y el siguiente, nos referimos a dicha estrategia como la estrategia *"salto"*. Por un lado, se encuentra el modelo con *"salto" instantáneo*, en el que cada vez que la diferencia entre la longitud de las dos colas es mayor que uno, el cliente en la parte de atrás de la cola brinca hasta atrás de la otra cola. En este modelo, las llegadas siempre se unen a la cola más corta. Por otro lado se encuentra el modelo de *no "salto"*, donde las llegadas todavía se unen a la cola más corta, pero no se permite brincar más tarde a la otra cola, incluso si es mucho más corta. Así un modelo intermedio sería el de estrategia con *"salto umbral"*, éste ocurre sólo cuando la diferencia entre la longitud de las dos colas alcanza un valor umbral, digamos N. Este valor *umbral* está definido como el valor máximo en el cual los clientes de la cola más larga ya no están a gusto con la diferencia de las longitudes de las colas.

El estudio de las propiedades de dos colas paralelas ha recibido mucha atención. El problema original se planteó por Haight [15]. El supone tiempos de arribos y tiempos de servicio exponenciales. Haight encuentra las propiedades de las probabilidades límite cuando el *"salto"* no es permitido. Gertsbakh[14] discutió algunos aspectos y propiedades del modelo *"salto" umbral*, el cual llamó el *sistema de colas pequeño modificado (MSQS)*. El supone que las tasas de servicio de los dos servidores son iguales. En este trabajo lo resolvemos a tasas distintas a través del *método de solución geométrica* descrito por Neuts [24].

En la sección 4.2 se analizan dos tipos de *"salto"*: el *"salto" instántaneo* y el *"salto" umbral*. Este último se describirá cuando un cliente se mueve desde la cola más larga a la cola más corta tan pronto como la diferencia en las longitudes de las dos colas sea mayor o igual que $N + 1$ (o estrictamente mayor que N)

para algún valor prescrito N. Para ambos tipos de *"salto"*, se supone que los arribos son un proceso de Poisson con tasa de arribo λ y los tiempo de servicio tienen distribución exponencial con tasas de servicio μ_1 y μ_2 (donde $\lambda < \mu_1 + \mu_2$). Si la longitud de las colas no son iguales, entonces al llegar un cliente se une a la cola más corta. Si la longitud de las colas son iguales, el cliente se une a cualquiera de las dos con probabilidad 0.5.

De manera similar, en la sección 4.3 se analiza un modelo de dos colas en paralelo con la estrategia *"salto"*, pero con la restricción de que hay una capacidad limitada a un valor finito L, incluyendo a los clientes que están siendo servidos. Los clientes llegan de acuerdo a un proceso de Poisson con tasa λ con servidores independientes que ofrecen servicio con distribución exponencial. Un cliente que llega se une a la cola más corta; si la dos colas son iguales, él elige la primera cola con probabilidad α o la segunda con probabilidad β, donde $\alpha + \beta = 1$.

4.2. Dos colas en paralelo con "salto" instantáneo y "salto" umbral

4.2.1. $N = 2$, $\mu_1 = \mu_2 = \mu$

Descripción del modelo

- El modelo que aquí se considera, es el sistema de colas más corto mejor llamado como el *"salto" instantáneo* cuando $N = 2$.
- Suponemos que los arribos son un proceso de Poisson con tasa de arribo λ y los tiempo de servicio tienen distribución exponencial con tasas de servicio $\mu_1 = \mu_2 = \mu$ (donde $\lambda < 2\mu$).
- Si las longitudes de las colas no son iguales, entonces al llegar un cliente se une a la fila más corta.
- Si las longitudes de las colas son iguales, al llegar un cliente se une a cualquiera de las dos con probabilidad 0.5.
- Sea $N_1(t)$ y $N_2(t)$ las variables aleatorias que describen el número de clientes al tiempo t en la cola 1 y en la cola 2, respectivamente.

Ya que las llegadas son procesos de Poisson y los tiempos de servicio son exponenciales, entonces tenemos que $\{(N_1(t), N_2(t)), t \geq 0\}$ describen un PNM de Markov. Sea (i, j) el estado del sistema que denota que hay i clientes en la colas 1 y j clientes en la cola 2 (incluyendo los clientes que están siendo servidos). Luego, como la estrategia *"salto instantáneo "* ocurre siempre que $|i - j| > 2$, entonces el espacio de estados (bidimensional) está conformado por los únicos estados en el sistema $\{(i, j) : \ |i - j| \leq 2\}$. Sea $\pi_{i,j}$ la probabilidad de equilibrio del sistema, en el estado (i, j), esto es $\pi_{i,j} = P\{N_1 = i, N_2 = j\}$ y como $\mu_1 = \mu_2$, entonces tenemos simetría en nuestras probabilidades, pues $\pi_{i,j} = \pi_{j,i}$ para todo i, j. Por lo que solo necesitamos considerar $\pi_{i,j}$ para $i \geq j$. Este modelo está representado por el diagrama de la figura 4.1

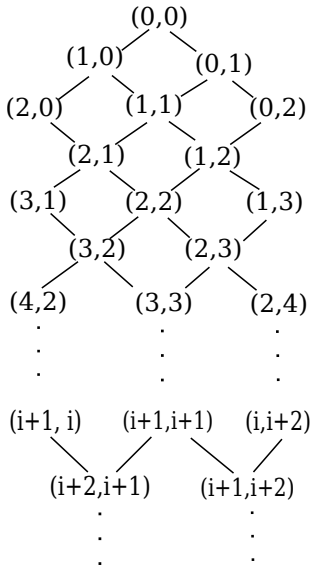


Figura 4.1: Diagrama para umbral N=2

Del diagrama, podemos observar que los *niveles* quedan descritos por:

$$\begin{aligned} n = 1 \quad & \{(0,0)\}, \\ n = 2 \quad & \{(1,0), (0,1)\}, \\ n = 3 \quad & \{(2,0), (1,1), (0,2)\}, \\ n = 4 \quad & \{(2,1), (1,2)\}, \\ n = 5 \quad & \{(3,1), (2,2), (1,3)\}, \\ \ldots, etc. & \end{aligned}$$

Así, obtenemos un proceso de Markov llamado un PNM (proceso de nacimiento y muerte) homogéneo donde las transiciones de un paso están restringidas a estados en el mismo nivel o a dos niveles adyacentes y las tasas de transiciones se supone están en niveles independientes. Es importante observar que todo nivel (después del primer nivel) tiene o bien, dos o tres estados. Cada estado da lugar a una ecuación de equilibrio. Si se cuenta a partir del estado $(0,0)$ hasta el final de un nivel de tres estados, solo contando estados (i,j) con $i \geq j$, entonces el número de ecuaciones que se obtienen es justamente uno menor que el número de variables conocidas $\pi_{i,j}$ involucradas (una variable a partir del siguiente nivel aparece en las dos últimas ecuaciones). Por lo tanto se pueden resolver para todas las $\pi_{i,j}$ en términos de $\pi_{0,0}$. Cuando se visualizan todos los estados y usando el hecho de que todas las probabilidades suman uno, obtenemos a $\pi_{0,0}$ y por lo tanto todas las $\pi_{i,j}$.

La técnica anterior tiene dificultad, sin embargo, para $N > 2$ o para $\mu_1 \neq \mu_2$ la dificultad se debe a que el número de ecuaciones obtenidas al final de cualquier nivel no es lo suficientemente grande como para que coincida con el número de variables desconocidas (estas incluyen variables de los niveles posteriores).

La distribución estacionaria

Ahora procedemos a resolver para $\pi_{i,j}$ en el caso $N = 2$, para $\mu_1 = \mu_2 = \mu$. De las tasas de salida y tasas de entrada en equilibrio para los estados (0,0), (1,0), (1,1) y (2,0), obtenemos las siguientes ecuaciones:

$$\begin{array}{lrcl} (0,0) & \lambda\pi_{0,0} & = & 2\mu\pi_{1,0} \ , \\ (1,1) & (2\mu+\lambda)\pi_{1,1} & = & 2\lambda\pi_{1,0} + 2\mu\pi_{2,1} \ , \\ (2,0) & (\mu+\lambda)\pi_{2,0} & = & \mu\pi_{2,1} \ , \\ (1,0) & (\mu+\lambda)\pi_{1,0} & = & \mu\pi_{1,1} + \mu\pi_{2,0} + \frac{\lambda}{2}\pi_{0,0} \ , \end{array} \tag{4.1}$$

o bien, el sistema homogéneo

$$\begin{cases} \lambda\pi_{0,0} - 2\mu\pi_{1,0} = 0 \\ (2\mu+\lambda)\pi_{1,1} - 2\lambda\pi_{1,0} - 2\mu\pi_{2,1} = 0 \\ (\mu+\lambda)\pi_{2,0} - \mu\pi_{2,1} = 0 \\ (\mu+\lambda)\pi_{1,0} - \mu\pi_{1,1} - \mu\pi_{2,0} - \frac{\lambda}{2}\pi_{0,0} = 0 \end{cases}$$

$$\begin{pmatrix} 0 & 0 & 0 & -2\mu & \lambda \\ -2\mu & 0 & 2\mu+\lambda & -2\mu & 0 \\ -\mu & \mu+\lambda & 0 & 0 & 0 \\ 0 & -\mu & -\mu & \mu+\lambda & -\lambda/2 \end{pmatrix} \begin{pmatrix} \pi_{2,1} \\ \pi_{2,0} \\ \pi_{1,1} \\ \pi_{1,0} \\ \pi_{0,0} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \ .$$

Para resolver este sistema de ecuaciones donde tenemos más variables que ecuaciones, procederemos a utilizar el método de eliminación de Gauss, donde la matriz asociada al sistema de ecuaciones homogéneo es:

$$\begin{pmatrix} 0 & 0 & 0 & -2\mu & \lambda \\ -2\mu & 0 & 2\mu+\lambda & -2\mu & 0 \\ -\mu & \mu+\lambda & 0 & 0 & 0 \\ 0 & -\mu & -\mu & \mu+\lambda & -\lambda/2 \end{pmatrix} \overrightarrow{E_1 \leftrightarrow E_3}$$

$$\begin{pmatrix} -\mu & \mu+\lambda & 0 & 0 & 0 \\ 0 & -\mu & -\mu & \mu+\lambda & -\lambda/2 \\ -2\mu & 0 & 2\mu+\lambda & -2\mu & 0 \\ 0 & 0 & 0 & -2\mu & \lambda \end{pmatrix} \overrightarrow{E_3 - 2E_1}$$

$$\begin{pmatrix} -\mu & \mu+\lambda & 0 & 0 & 0 \\ 0 & -\mu & -\mu & \mu+\lambda & -\lambda/2 \\ 0 & -2(\mu+\lambda) & 2\mu+\lambda & -2\mu & 0 \\ 0 & 0 & 0 & -2\mu & \lambda \end{pmatrix} \xrightarrow[E_3 - 2(1+\frac{\lambda}{\mu})E_2]{}$$

$$\begin{pmatrix} -\mu & \mu+\lambda & 0 & 0 & 0 \\ 0 & -\mu & -\mu & \mu+\lambda & -\lambda/2 \\ 0 & 0 & 4\mu^2+3\lambda\mu & -(6\lambda+2\mu+2\lambda^2) & \mu(\lambda+\mu) \\ 0 & 0 & 0 & -2\mu & \lambda \end{pmatrix} .$$

Esta matriz se encuentra de forma triangular. Despejando y sustituyendo recursivamente y haciendo $\rho = \lambda/2\mu$, obtenemos las siguientes ecuaciones:

$$\pi_{1,0} = \rho\pi_{0,0} \; ; \qquad \pi_{1,1} = \frac{4\rho^2(1+\rho)}{(2+3\rho)}\pi_{0,0} \; ;$$

$$\pi_{2,0} = \frac{2\rho^3}{(2+3\rho)}\pi_{0,0} \; ; \qquad \pi_{2,1} = \frac{2\rho^3(1+2\rho)}{(2+3\rho)}\pi_{0,0} \; .$$

Similarmente, si consideramos estados (i,i), $(i+1,i)$, $(i+2,i)$ y $(i+1,i+1)$ para $i \geq 1$, obtenemos:

$$(2\mu+\lambda)\pi_{i+1,i+1} = 2\lambda_{i+1,i} + 2\mu\pi_{i+2,i+1} \; , \tag{4.2}$$

$$(2\mu+\lambda)\pi_{i+1,i} = 2\mu\pi_{i+2,i} + \frac{\lambda}{2}\pi_{i,i} + \mu\pi_{i+1,i+1} + \lambda\pi_{i+1,i-1} \; , \tag{4.3}$$

$$(2\mu+\lambda)\pi_{i+2,i} = \mu\pi_{i+2,i+1} \; . \tag{4.4}$$

De (4.2):

$$\pi_{i+1,i+1} = \frac{2\rho\pi_{i+1,i} + \pi_{i+2,i+1}}{1+\rho} \; . \tag{4.5}$$

De (4.4):

$$\pi_{i+2,i} = \frac{\pi_{i+2,i+1}}{2(1+\rho)} \; . \tag{4.6}$$

Así sustituyendo (4.5) y (4.6) en (4.3) logramos tener:

$$(2\mu+\lambda)\pi_{i+1,i} = 2\mu\left[\frac{\pi_{i+2,i+1}}{2(1+\rho)}\right] + \frac{\lambda}{2}\,\pi_{i,i} + \mu\left[\frac{2\rho\pi_{i+1,i} + \pi_{i+2,i+1}}{1+\rho}\right] + \lambda\,\pi_{i+1,i-1} \; ,$$

$$(2\mu+\lambda)\pi_{i+1,i} = \frac{\lambda}{\mu}\left[\frac{\pi_{i+2,i+1}}{2(1+\rho)}\right] + \frac{\lambda}{2}\,\pi_{i,i} + \frac{\lambda}{2\rho}\left[\frac{2\rho\pi_{i+1,i}+\pi_{i+2,i+1}}{1+\rho}\right] + \lambda\,\pi_{i+1,i-1}\ ,$$

$$\pi_{i+2,i+1} = \frac{\rho(1+\rho)}{\lambda}\left[\frac{\lambda(1+\rho+\rho^2)}{\rho(1+\rho)}\pi_{i+1,i} - \frac{\lambda}{2}\pi_{i,i} - \lambda\pi_{i+1,i-1}\right].$$

Por lo tanto

$$\begin{aligned}
\pi_{i+2,i+1} &= (1+\rho+\rho^2)\pi_{i+1,i} - \tfrac{\rho}{2}(1+\rho)\pi_{i,i} - \rho(1+\rho)\pi_{i+1,i-1}\ , \\
\pi_{i+1,i+1} &= ((1+\rho)^2+\rho)/(1+\rho)\pi_{i+1,i} - (\rho/2)\pi_{i,i} - \rho\pi_{i+1,i-1}\ , \\
\pi_{i+2,i} &= (1+\rho+\rho^2)/(2(1+\rho))\pi_{i+1,i} - (\rho/4)\pi_{i,i} - (\rho/2)\pi_{i+1,i-1}\ .
\end{aligned} \tag{4.7}$$

Para las ecuaciones (4.1) obtenemos, para $i \geq 1$,

$$\begin{bmatrix}\pi_{i+2,i+1}\\ \pi_{i+1,i+1}\\ \pi_{i+2,i}\end{bmatrix} = X\begin{bmatrix}\pi_{i+1,i}\\ \pi_{i,i}\\ \pi_{i+1,i-1}\end{bmatrix} = X^2\begin{bmatrix}\pi_{i,i-1}\\ \pi_{i-1,i-1}\\ \pi_{i,i-2}\end{bmatrix} = \cdots = X^{i-1}\begin{bmatrix}\pi_{3,2}\\ \pi_{2,2}\\ \pi_{3,1}\end{bmatrix} = X^i\begin{bmatrix}\pi_{2,1}\\ \pi_{1,1}\\ \pi_{2,0}\end{bmatrix}\ .$$

Y este último por la primera solución obtenida anteriormente la cuál está en términos de $P_{0,0}$, se sigue que:

$$\begin{bmatrix}\pi_{i+2,i+1}\\ \pi_{i+1,i+1}\\ \pi_{i+2,i}\end{bmatrix} = X^i\begin{bmatrix}\pi_{2,1}\\ \pi_{1,1}\\ \pi_{2,0}\end{bmatrix} = X^i\begin{bmatrix}\rho+2\rho^2\\ 2+2\rho\\ \rho\end{bmatrix}\frac{2\rho^2}{(2+3\rho)}\,\pi_{0,0}\ ,$$

Donde

$$X = \begin{pmatrix} 1+\rho+\rho^2 & -\frac{\rho}{2}(1+\rho) & -\rho(1+\rho) \\ ((1+\rho)^2+\rho)/1+\rho & -\rho/2 & -\rho \\ (1+\rho+\rho^2)/2(1+\rho) & -\rho/4 & -\rho/2 \end{pmatrix}.$$

Sea $M = X\begin{bmatrix}\rho+2\rho^2\\ 2+2\rho\\ \rho\end{bmatrix}$. Se puede verificar que $XM = \rho^2 M$. Por consiguiente

$$\begin{bmatrix}\pi_{i+2,i+1}\\ \pi_{i+1,i+1}\\ \pi_{i+2,i}\end{bmatrix} = X^{i-1}M\frac{2\rho^2}{(2+3\rho)}\pi_{0,0} \quad = X^{i-2}M\frac{2\rho^4}{(2+3\rho)} \quad = X^{i-3}M\frac{2\rho^6}{(2+3\rho)}$$

$$= \quad . \quad . \quad . \quad = X^{i-i}M\rho^i\frac{2\rho^2}{(2+3\rho)}\pi_{0,0} \quad = M\rho^{2i}\frac{2}{(2+3\rho)}\pi_{0,0}$$

$$= \begin{bmatrix}2\rho(1+2\rho)/(2+3\rho)\\ 2(2+\rho)(1+2\rho)/(1+\rho)(2+3\rho)\\ \rho(1+2\rho)/(1+\rho)(2+3\rho)\end{bmatrix}\rho^{2i+2}\pi_{0,0} \qquad (i \geq 1).$$

Finalmente, ya que $\sum_{i,j}\pi_{i,j} = 1$, tenemos que

$$\pi_{0,0} + 2\pi_{1,0} + 2\pi_{2,0} + 2\pi_{2,1} + \sum_{i=1}^{\infty}(\pi_{i+1,i+1} + 2\pi_{i+2,i+1} + 2\pi_{i+2,i}) = 1\ .$$

Así

$$\pi_{0,0}\left\{1 + \frac{4\rho^2(1+\rho)}{2+3\rho} + 2\rho + \frac{4\rho^3(1+2\rho)}{2+3\rho} + \sum_{i=1}^{\infty}\left\{\frac{2(2+\rho)(1+2\rho)}{(1+\rho)(2+3\rho)} + \frac{4\rho(1+2\rho)}{2+3\rho} + \frac{2\rho(1+2\rho)}{(1+\rho)(2+3\rho)}\right\}\rho^{2i+2}\right\} = 1.$$

Resolviendo y simplificando tenemos

$$\pi_{0,0} = \frac{(2+\rho-3\rho^2)}{2+5\rho+3\rho^2+2\rho^3}. \tag{4.8}$$

Para las ecuaciones (4.1), (4.7) y (4.8), obtuvimos las probabilidades de equilibrio para el sistema en términos de la intensidad de trafico $\rho = \lambda/2\mu$.

4.2.2. N arbitraria, $\mu_1 \neq \mu_2$

Descripción del modelo

- El modelo que aquí se considera, es el sistema de colas con *"salto umbral"*, con N arbitrario ($3 \leq N < \infty$).
- Suponemos que los arribos son un proceso de Poisson con tasa de arribo λ y los tiempo de servicio están distribuidos exponencialmente con diferentes tasas de servicio μ_1 y μ_2.

- Si las longitudes de las colas no son iguales, entonces al llegar un cliente se une a la fila más corta.
- Si las longitudes de las colas son iguales, entonces al llegar un cliente se une a cualquiera de las dos con probabilidad 0.5.
- Sea $N_1(t)$ y $N_2(t)$ las variables aleatorias que describen el número de clientes al tiempo t en la cola 1 y en la cola 2, respectivamente.

De igual manera, como en en la sección 4.2.1, ya que las llegadas son procesos de Poisson y los tiempos de servicio son exponenciales, entonces tenemos que $\{(N_1(t), N_2(t)), t \geq 0\}$ describen un PNM de Markov. Sea (i, j) el estado del sistema que denota que hay i clientes en la colas 1 y j clientes en la cola 2 (incluyendo los clientes que están siendo servidos). Aún más, como la estrategia *"salto umbral"* ocurre siempre que $|i - j| > N$ $(3 \leq N < \infty)$, entonces el espacio de estados (bidimensional) está conformado por los únicos estados en el sistema $\{(i, j) : |i - j| \leq N\}$. Sea $\pi_{i,j}$ la probabilidad de equilibrio del sistema, en el estado (i, j), y como esta vez $\mu_1 \neq \mu_2$, entonces tenemos que las probabilidades $\pi_{i,j} \neq \pi_{j,i}$ para todo i, j. El diagrama de estados está dado como sigue, en la figura 4.2:

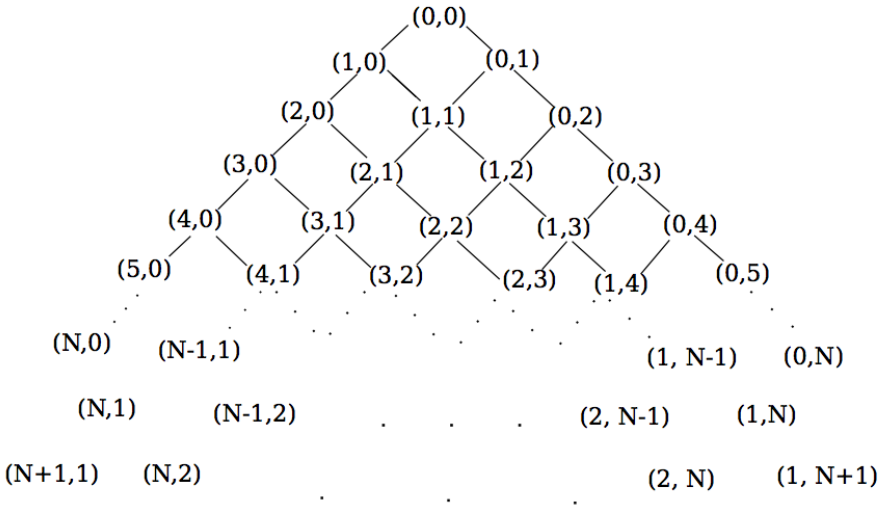


Figura 4.2: Diagrama con umbral N, para N arbitrario

En el diagrama podemos observar que en éste, a diferencia del anterior, todo nivel a partir del primero va creciendo el número de estados; más aún, hay n estados en cada *nivel* n (ver Figura 4.3) por lo que la técnica anterior ya no funciona para este caso. Es así como desearíamos encontrar una técnica que nos permita encontrar una solución a las ecuaciones de equilibrio que surgen de este PNM.

Análisis del modelo

Para el PNM que describe este modelo, ordenando los estados de manera lexicográfica, se obtiene la *matriz generadora infinitesimal* Q descrita como sigue:

$$Q = \begin{pmatrix}
A_1 & C_1 & & & & & & & & \\
B_1 & A_2 & C_2 & & & & & & & \\
 & B_2 & A_3 & C_3 & & & & & & \\
 & & \ddots & \ddots & \ddots & & & & & \\
 & & & & C_N & & & & & \\
 & & & B_N & A_{N+1} & D_2 & & & & \\
 & & & & E_1 & F_1 & D_1 & & & \\
 & & & & & E_2 & F_2 & D_2 & & \\
 & & & & & & E_1 & F_1 & D_1 & \\
 & & & & & & & E_2 & F_2 & D_2 \\
 & & & & & & & & \ddots & \ddots & \ddots
\end{pmatrix},$$

donde todas las entradas que no se muestran son cero. Esta *matriz generadora infinitesimal* Q, muestra una estructura en forma tridiagonal en bloques, donde cada bloque está formado por submatrices. Las matrices A_i, B_i, C_i y D_i son submatrices definidas de la siguiente manera: para el n-ésimo renglón de Q que se muestra, consiste de submatrices que corresponde al conjunto de estados $\{(i,j) : i + j = n - 1\}$. El n-ésimo renglón de Q también corresponde al n-ésimo *nivel* horizontal (véase sección 2.2) como se muestra en la Figura 4.3. con lineas punteadas horizontales.

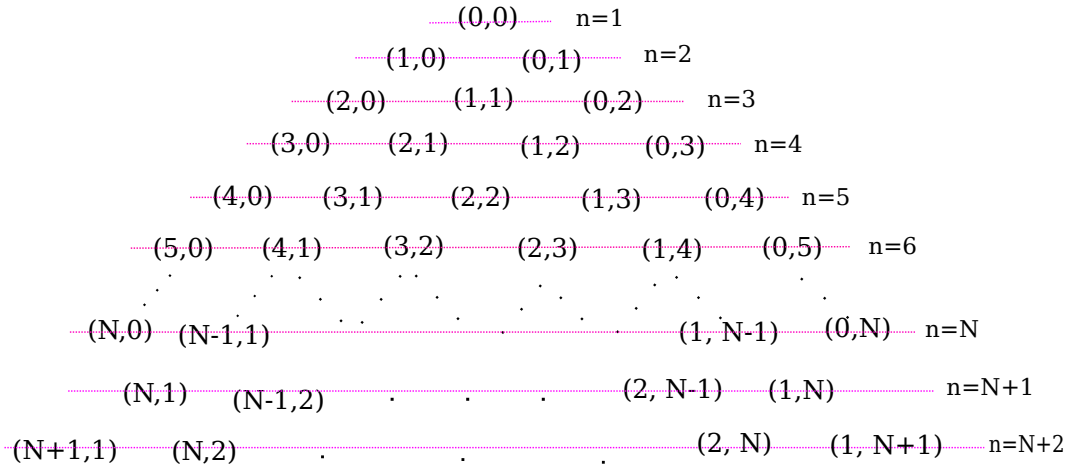


Figura 4.3: Diagrama para umbral N arbitrario, con los niveles horizontales.

Por lo que las submatrices son como se muestran a continuación: sea $a = -(\lambda + \mu_1)$, $b = -(\lambda + \mu_2)$, $c = -(\lambda + \mu_1 + \mu_2)$. Entonces

$$\mathrm{A}_1 = \begin{pmatrix} -\lambda \end{pmatrix}, \qquad \mathrm{A}_2 = \begin{pmatrix} b & 0 \\ 0 & a \end{pmatrix}, \qquad \mathrm{A}_3 = \begin{pmatrix} b & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & a \end{pmatrix},$$

$$\mathrm{A}_4 = \begin{pmatrix} b & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & a \end{pmatrix}, \quad \cdots \quad , \mathrm{A}_{N+1} = \begin{pmatrix} b & 0 & 0 & \ldots & 0 & 0 \\ 0 & c & 0 & \ldots & 0 & 0 \\ 0 & 0 & c & \ldots & 0 & 0 \\ & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \ldots & c & 0 \\ 0 & 0 & 0 & \ldots & 0 & a \end{pmatrix}_{(N+1)\times(N+1)} .$$

$$B_1 = \begin{pmatrix} \mu_2 \\ \mu_1 \end{pmatrix}, \qquad B_2 = \begin{pmatrix} \mu_2 & 0 \\ \mu_1 & \mu_2 \\ 0 & \mu_1 \end{pmatrix}, \qquad B_3 = \begin{pmatrix} \mu_2 & 0 & 0 \\ \mu_1 & \mu_2 & 0 \\ 0 & \mu_1 & \mu_2 \\ 0 & 0 & \mu_1 \end{pmatrix},$$

$$B_N = \begin{pmatrix} \mu_2 & 0 & 0 & 0 & \ldots & 0 & 0 \\ \mu_1 & \mu_2 & 0 & 0 & \ldots & 0 & 0 \\ 0 & \mu_1 & \mu_2 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \mu_1 & \mu_2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & \mu_1 & \mu_2 \\ 0 & 0 & 0 & 0 & \ldots & 0 & \mu_1 \end{pmatrix}_{(N+1)\times N} .$$

$$C_1 = \begin{pmatrix} \lambda/2 & \lambda/2 \end{pmatrix}, \qquad C_2 = \begin{pmatrix} 0 & \lambda & 0 \\ 0 & \lambda & 0 \end{pmatrix}, \qquad C_3 = \begin{pmatrix} 0 & \lambda & 0 & 0 \\ 0 & \lambda/2 & \lambda/2 & 0 \\ 0 & 0 & \lambda & 0 \end{pmatrix},$$

$$C_4 = \begin{pmatrix} 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & \lambda & 0 \end{pmatrix},$$

$$C_N = \begin{pmatrix} 0 & \lambda & 0 & & & & & & & & \\ 0 & 0 & \lambda & & & & & & & & \\ 0 & 0 & 0 & \lambda & & & & & & & \\ \vdots & & & & \ddots & & & & & & \\ & & & & & \lambda & & & & & \\ & & & & & \lambda/2 & \lambda/2 & & & & \\ & & & & & & & \lambda & & & \\ & & & & & & & & \ddots & & \vdots \\ & & & & & & & & & \lambda & 0 & 0 & 0 \\ & & & & & & & & & & \lambda & 0 & 0 \\ & & & & & & & & & & & \lambda & 0 \end{pmatrix}_{N\times(N+1)} \quad \textit{si } N \textit{ es impar ,}$$

$$C_N = \begin{pmatrix} 0 & \lambda & 0 & & & & & & & & \\ 0 & 0 & \lambda & & & & & & & & \\ 0 & 0 & 0 & \lambda & & & & & & & \\ \vdots & & & & \ddots & & & & & & \\ & & & & & \lambda & & & & & \\ & & & & & & \lambda & & & & \\ & & & & & & \lambda & & & & \\ & & & & & & & \ddots & & \vdots & \\ & & & & & & & & \lambda & 0 & 0 & 0 \\ & & & & & & & & & \lambda & 0 & 0 \\ & & & & & & & & & & \lambda & 0 \end{pmatrix}_{N\times(N+1)} \quad \textit{si } N \textit{ es par .}$$

$$D_2 = \begin{pmatrix} \lambda & 0 & & & & & & & \\ 0 & \lambda & & & & & & & \\ 0 & 0 & \lambda & & & & & & \\ & & & \ddots & & & & & \\ & & & & \lambda & & & & \\ & & & & \lambda/2 & \lambda/2 & & & \\ & & & & & & \lambda & & \\ & & & & & & & \ddots & \\ & & & & & & & \lambda & 0 & 0 \\ & & & & & & & & \lambda & 0 \\ & & & & & & & & & \lambda \end{pmatrix}_{N\times(N+1)} \quad \textit{si } N \textit{ es impar},$$

$$D_2 = \begin{pmatrix} \lambda & 0 & & & & & & & \\ 0 & \lambda & & & & & & & \\ 0 & 0 & \lambda & & & & & & \\ & & & \ddots & & & & & \\ & & & \lambda & & & & & \\ & & & & \lambda & & & & \\ & & & & \lambda & & & & \\ & & & & & \ddots & & & \\ & & & & & & \lambda & 0 & 0 \\ & & & & & & & \lambda & 0 \\ & & & & & & & & \lambda \end{pmatrix}_{N\times(N+1)} \quad \textit{si } N \textit{ es par}.$$

$$\mathrm{E}_1 = \begin{pmatrix} \mu_1 & \mu_2 & & & & \\ & \mu_1 & \mu_2 & & & \\ & & \mu_1 & \mu_2 & & \\ & & & \ddots & \ddots & \\ & & & & \mu_1 & \mu_2 \\ & & & & & \mu_1 & \mu_2 \end{pmatrix}_{N\times(N+1)}, E_1 = \begin{pmatrix} \mu_1+\mu_2 & 0 & & & & \\ & \mu_1 & \mu_2 & & & \\ & & \mu_1 & \mu_2 & & \\ & & & \ddots & \ddots & \\ & & & & \mu_1 & \mu_2 \\ & & & & & 0 & \mu_1+\mu_2 \end{pmatrix}_{N\times(N+1)}.$$

$$F_1 = \begin{pmatrix} c & & & \\ & c & & \\ & & \ddots & \\ & & & c \\ & & & & c \end{pmatrix}_{N\times N}, \qquad \mathrm{F}_2 = \begin{pmatrix} c & & & \\ & c & & \\ & & \ddots & \\ & & & c \\ & & & & c \end{pmatrix}_{(N+1)\times(N+1)}.$$

Como podemos observar, la forma en que Q esta representada, a pesar de que tiene la forma tridiagonal

por bloques, no es la que necesitamos para aplicar el método de la solución geométrica ([24] y sección 2.4) pues los bloques no se comportan en la forma estándar que necesitamos.

Otra forma de obtener *la matriz generadora infinitesimal* es la descrita a continuación, donde todas las entradas que no aparecen son ceros.

$$\widetilde{Q} = \begin{pmatrix} X_0 & Y_0 & & & & & \\ Y_2 & X_1 & Y_1 & & & & \\ & Y_2 & X_1 & Y_1 & & & \\ & & Y_2 & X_1 & Y_1 & & \\ & & & Y_2 & X_1 & Y_1 & \\ & & & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Para la $(n-1)$-ésima fila que se muestra en $\widetilde{Q}$ corresponde al conjunto $\{(n, N+n), (n, N+n-1), \ldots, (n, n), \ldots, (N+n-1, n), (N+n, n)\}$ o equivalentemente, el $(n+1)$-ésimo renglón involucra todos los estados $\{(i, j) : \text{mín}\{i, j\} = n\}$. La figura 4.4 puede ayudar a entender como esta descrito este espacio de estados, donde podría considerarse que los *niveles* (véase sección 2.2) tienen la forma de un gorro.

$$\begin{aligned} n = 1 \quad & \{(0, N), (0, N-1), ..., (0, 0), ..., (N-1, 0), (N, 0)\}, \\ n = 2 \quad & \{(1, N+1), (1, N), ..., (1, 1), ..., (N, 1), (N+1, 1)\}, \\ n = 3 \quad & \{(2, N+2), (2, N+1), ..., (2, 2), ..., (N+2), (N+1, 2)\}, \\ \cdots, etc. & \end{aligned}$$

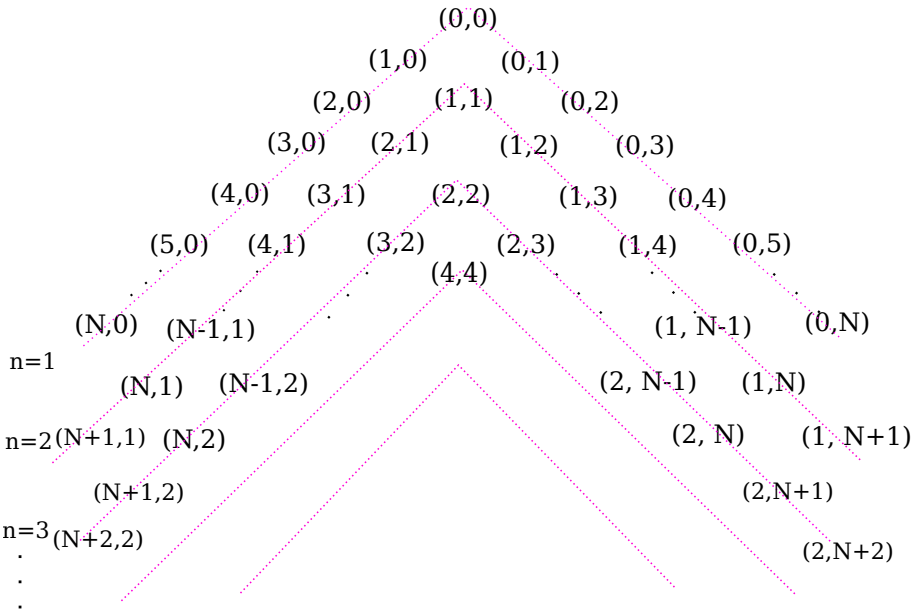


Figura 4.4: Diagrama para umbral N arbitrario, con los niveles en forma de gorro.

Así X_0, X_1, Y_0, Y_1, Y_2 son submatrices definidas de la siguiente forma, con a, b, y c definidos como anteriormente y $J = 2N + 1$:

$$X_0 = \begin{pmatrix} b & \mu_2 & & & & & & \\ & b & \mu_2 & & & & & \\ & & \ddots & \ddots & & & & \\ & & & b & \mu_2 & & & \\ & & & \lambda/2 & -\lambda & \lambda/2 & & \\ & & & & \mu_1 & a & & \\ & & & & & \ddots & \ddots & \\ & & & & & & \mu_1 & a \end{pmatrix}_{J\times J}, \quad X_1 = \begin{pmatrix} c & \mu_1+\mu_2 & & & & & & \\ & c & \mu_2 & & & & & \\ & & \ddots & \ddots & & & & \\ & & & c & \mu_2 & & & \\ & & & \lambda/2 & c & \lambda/2 & & \\ & & & & \mu_1 & c & & \\ & & & & & \ddots & \ddots & \\ & & & & & & \mu_1+\mu_2 & c \end{pmatrix}_{J\times J},$$

$$Y_1 = \begin{pmatrix} 0 & \lambda & & & & & & \\ & 0 & \lambda & & & & & \\ & & \ddots & & & & & \\ & & & \lambda & & & & \\ & & & 0 & & & & \\ & & & \lambda & \ddots & & & \\ & & & & & 0 & & \\ & & & & & \lambda & & 0 \\ & & & & & & \lambda & 0 \end{pmatrix}_{J\times J}, \qquad Y_2 = \begin{pmatrix} 0 & & & & & & & \\ \mu_1 & 0 & & & & & & \\ & \mu_1 & 0 & & & & & \\ & & & \ddots & & & & \\ & & & \mu_1 & 0 & \mu_2 & & \\ & & & & & \ddots & & \\ & & & & & & 0 & \mu_2 \\ & & & & & & & 0 \end{pmatrix}_{J\times J}.$$

La segunda forma de la *matriz generadora infinitesimal* $\widetilde{Q}$ tiene la ventaja de que todas las filas excepto la primera tiene la misma forma, es decir la estructura tridiagonal en bloques que deseamos. Al usar el tipo de matrices $\widetilde{Q}$ como anteriormente se describieron, ahora somos capaces calcular las probabilidades estacionarias.

La distribución estacionaria

Ahora estamos listos para calcular la distribución estacionaria para el modelo de dos colas en paralelo donde ocurre la estrategia *"salto"* cuando la diferencia entre las dos colas alcanza el valor N, para algún N arbitrario dado, y el sistema tiene tasas de servicio distintas, a saber μ_1 y μ_2.

Sea π el vector de probabilidad estacionaria, donde $\pi=(\pi_0, \pi_1, \pi_1, ...)$, con

$$\pi_0 = (\pi(N,0), \pi(N-1,0), ..., \pi(1,0), \pi(0,0), \pi(0,1), ...\pi(0,N-1), \pi(0,N)),$$
$$\pi_1 = (\pi(N+1,1), \pi(N,1), ..., \pi(2,1), \pi(1,1), \pi(1,2), ...\pi(1,N), \pi(1,N+1)),$$

$$\pi_n = (\pi(N+n,n), \pi(N+n-1,n), ..., \pi(n,n), ...\pi(n,N+n-1), \pi(n,N+n)), \qquad n \geq 2.$$

Donde $\pi(i,j)$ nos denota la probabilidad estacionaria de que el proceso está en el estado (i,j). Los vectores de probabilidad pueden ser expresados como

$$\pi_{n+1} = \pi_n R, \qquad n \geq 1, \tag{4.9}$$

donde la matriz R es llamada la *matriz de tasas* R que es la solución mínima no negativa de la ecuación matriz cuadrática

$$Y_1 + RX_1 + R^2 Y_2 = \mathbf{0}\,. \tag{4.10}$$

Método de la matriz geométrica

Volviendo a describir la *matriz generadora infinitesimal* $\widetilde{Q}$

$$\widetilde{Q} = \begin{pmatrix} X_0 & Y_0 & & & & & \\ Y_2 & X_1 & Y_1 & & & & \\ & Y_2 & X_1 & Y_1 & & & \\ & & Y_2 & X_1 & Y_1 & & \\ & & & Y_2 & X_1 & Y_1 & \\ & & & & \ddots & \ddots & \ddots \end{pmatrix},$$

donde Y_1, X_1 y Y_2 son matrices cuadradas de orden $2J$. Las matrices Y_0, Y_1 y Y_2 son no negativas y las matrices X_0 y X_1 tienen elementos no negativos fuera de la diagonal y estrictamente negativos en la diagonal. La suma de los renglones de $\widetilde{Q}$ es igual a cero.

El PNM de Markov $\widetilde{Q}$ es irreducible y aperiódico trivialmente. Además, ya que los estados en los niveles mayores a N pueden comunicarse unos con otros vía sus rutas, no pasando de niveles $\geq N$, el generador $A_0 + A_1 + A_2$ es también irreducible. Así por el teorema 1.7.1 del libro de Neuts [24] se puede aplicar fácilmente. En concreto, el proceso de Markov Q es ergódico si y solo si satisface la condición

$$\pi Y_0 \mathbf{e} < \pi Y_2 \mathbf{e} \tag{4.11}$$

donde $\pi = (\pi_0, \pi_1, ...)$ es la distribución de equilibrio de el generador $Y_1 + X_1 + Y_2$ y $\mathbf{e}$ es el vector unitario. Cuando la condición (4.11) se cumple, la distribución estacionaria del PNM existe.

Así, la condición de estabilidad para que exista $\widetilde{Q}$ según (4.11) es

$$2\lambda < \mu_1 + \mu_2$$

Designando por $\pi(i,j)$ la probabilidad estacionaria de que el proceso se encuentre en el estado (i,j), y el uso de la notación vectorial $\pi_n = (\pi(N+n,n), \pi(N+n-1,n), ..., \pi(n,n), ...\pi(n,N+n-1)$, las ecuaciones de balance del PNM están dadas por

$$\pi_{n-1}Y_1 + \pi_n X_1 + \pi_{n+1}Y_2 = \mathbf{0}, \qquad n \geq 2, \tag{4.12}$$

y

$$\pi_0 X_0 + \pi_1 Y_2 = 0, \tag{4.13}$$
$$\pi_0 Y_0 + \pi_1 X_1 + \pi_2 Y_2 = 0. \tag{4.14}$$

Introduciendo la *matriz de tasas* R como la solución mínima no negativa de la ecuación matricial no líneal

$$Y_1 + RX_1 + R^2 Y_2 = \mathbf{0}\ , \tag{4.15}$$

se puede demostrar que las probabilidades de equilibrio satisfacen (véase [24], pp. 80-83).

$$\pi_{n+1} = \pi_n R, \qquad n \geq 2. \tag{4.16}$$

Los vectores π_1 y π_0 se derivan de las condiciones límite (4.13) y (4.14)

$$\pi_0 X_0 + \pi_1 Y_2 = \mathbf{0}\ , \tag{4.17}$$

$$\pi_0 Y_0 + \pi_1 (X_1 + RY_2) = \mathbf{0}\ ,$$

y la condición de normalización

$$\sum_{n=0}^{\infty} \pi_n = \pi_0 \mathbf{e} + \pi_1 (I - R)^{-1} \mathbf{e} = 1\ , \tag{4.18}$$

donde I representa la matriz identidad.

Cálculo de la *matriz de tasas* R

Con el fin de obtener la distribución estacionaria, debemos obtener la *matriz de tasas* R. Existen varios algoritmos iterativos para resolver (4.15) (ver [26] y [1]). Uno de ellos (el más común) es el que desarrollaremos aquí. Antes, probamos un lema.

Lema 4.2.1 *La inversa de X_1 existe y su determinante puede ser calculado por*

$$det(X_1) = c^N$$

Demostración: Probamos el lema mediante el cálculo de la inversa X_1^{-1}. La idea del método es evaluar el determinante de X_1, esto es, reducir la matriz a la forma triangular superior por las operaciones elementales sobre las filas, obteniendo así:

$$det(B_k) = \begin{vmatrix} c & x_0 & 0 & 0 & & & & & 0 \\ 0 & c & \mu_2 & 0 & 0 & & & & 0 \\ \vdots & 0 & \ddots & \ddots & 0 & 0 & & & \vdots \\ & \vdots & 0 & c & \mu_2 & 0 & 0 & & \\ & & \vdots & 0 & 1 & x_1 & 0 & & \\ & & & \vdots & 0 & 1 & x_2 & & \\ & & & & & \ddots & \ddots & \ddots & \\ & & & & & & 0 & 1 & 0 \\ 0 & 0 & 0 & & & & & 0 & c \end{vmatrix}$$

donde $x_0 = \mu_1 + \mu_2$, $x_1 = \frac{\mu_2}{c}$, $x_2 = \frac{2\lambda c}{-2\lambda\mu_1 + 2c^2}$, hay que mencionar que el primer renglón que tiene 1 en la diagonal, corresponde al N-ésimo renglón.
Ahora simplemente por las propiedades del determinante, el valor es la multiplicación de la diagonal

$$det(X_1) = c^{N-1}c = c^N.$$

Claramente, el $det(X_1) \neq 0$, pues recordemos que $c = -(\lambda + \mu_1 + \mu_2)$, entonces existe la inversa de X_1. □

Ahora bien, de la ecuación (4.15) y del lema 4.2.1

$$\begin{aligned} RX_1 &= -Y_1 - R^2Y_2, \\ RX_1X_1^{-1} &= -[Y_1 + R^2Y_2]X_1^{-1}, \\ RI &= -[Y_1 + R^2Y_2]X_1^{-1}, \\ R &= -[Y_1 + R^2Y_2]X_1^{-1}, \end{aligned}$$

$$R_{k+1} = -[Y_1 + R_k^2\, Y_2]\, X_1^{-1} \qquad k = 0, 1, ...\cdot \tag{4.19}$$

y tomando el valor inicial de R igual a la matriz cero, se puede resolver iterativamente para R y se puede verificar la exactitud de esta aproximación usando la igualdad $RA_2\mathbf{e} = A_0\mathbf{e}$. El valor de R convergerá ya que el valor de $-A_1^{-1}$ y $[A_0 + R^2A_2]$ son positivos. Por lo que después de cada iteración, los elementos de R aumentarán de manera monótona.

Resumen para el cálculo de las probabilidades estacionarias

PASO 1: Obtener la *matriz de tasas* R con el método iterativo (4.19).

PASO 2: Calcular los valores π_0 y π_1 resolviendo el sistema de ecuaciones (4.17) y la condición de normalización (4.18), es decir resolver el sistema

$$[\ \pi_0 \quad \pi_1\ ] \begin{bmatrix} \mathbf{e} & X_0* & Y_0 \\ (I-R)^{-1}\mathbf{e} & Y_2* & X_1+RY_2 \end{bmatrix} = [\ 1 \quad \mathbf{0}\ ]\ .$$

Donde M^* es M con la primer columna eliminada.

PASO 3: Cálculo de las demás probabilidades estacionarias π_n, para $n \geq 2$, con la condición (4.16).

4.3. Dos colas en paralelo con "salto" y capacidad restringida a L

4.3.1. Descripción del modelo

- En este modelo, el proceso de llegadas de los clientes es un proceso de Poisson con tasa de arribos λ. El sistema de colas consiste de dos servidores independientes en paralelo que ofrecen servicio (tiempos distribuidos exponencialmente) con diferentes tasas de servicio μ_1 y μ_2, respectivamente.
- Si las longitudes de las dos colas no son iguales, entonces un cliente al llegar se une a la fila más corta.
- Si las longitudes de las dos colas tienen misma longitud, entonces ellos eligen una primer cola con probabilidad α o la segunda con probabilidad β, donde $\alpha + \beta = 1$.
- La capacidad de cada cola está restringida a L, $L \geq 1$ incluyendo el cliente que está siendo servido.
- En el momento en que el servidor se vuelve inactivo y hay un cliente esperando en la otra cola, el cliente inmediatamente después del cliente que está recibiendo el servicio en ese servidor se transfiere a la cola del servidor inactivo.
- Sea $N_1(t)$ y $N_2(t)$, respectivamente el número de clientes en cada cola al tiempo t.

Otra vez, como en las secciones 4.2.1 y 4.2.2, ya que las llegadas son procesos de Poisson y los tiempos de servicio son exponenciales, el proceso estocástico $\{(N_1(t), N_2(t)), t \geq 0\}$ describe un PNM de Markov con espacios de estados (bidimensioanal) en $\{0, 1, 2, ...L\} \times \{0, 1, 2, ...L\}$. Sea (i, j) el estado del sistema que denota que hay i clientes en la colas 1 y j clientes en la cola 2 (incluyendo los clientes que están siendo servidos). Sea $\pi_{i,j} = P(N_1 = i, N_2 = j)$ la probabilidad de estado de equilibrio del sistema y como $\mu_1 \neq \mu_2$, entonces tenemos que las probabilidades $\pi_{i,j} \neq \pi_{j,i}$ para todo i, j.

A partir de los supuestos anteriores, la probabilidad $\pi_{i,j}$ satisface el siguiente sistema de ecuaciones de equilibrio y el diagrama que forma este modelo se muestra en la Figura 4.5.

$$\lambda\pi_{0,0} = \mu_1\pi_{1,0} + \mu_2\pi_{0,1}, \tag{4.20}$$

$$(\lambda + \mu_2)\pi_{0,1} = \lambda\beta\pi_{0,0} + \mu_1\pi_{1,1}, \tag{4.21}$$

$$(\lambda + \mu_1)\pi_{1,0} = \alpha\beta\pi_{0,0} + \mu_2\pi_{1,1}, \tag{4.22}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{1,1} = \lambda(\pi_{0,1} + \pi_{1,0}) + (\mu_1 + \mu_2)(\pi_{1,2} + \pi_{2,1}), \tag{4.23}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{1,2} = \lambda\beta\pi_{1,1} + \mu_1\pi_{2,2} + (\mu_1 + \mu_2)\pi_{1,3}, \tag{4.24}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{2,1} = \lambda\alpha\pi_{1,1} + \mu_2\pi_{2,2} + (\mu_1 + \mu_2)P_{3,1}, \tag{4.25}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{2,2} = \lambda(\pi_{1,2} + \mu_1\pi_{2,1}) + \mu_1\pi_{3,2} + \mu_2\pi_{2,3}, \tag{4.26}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{1,n} = (\mu_1 + \mu_2)\pi_{1,n+1} + \mu_1\pi_{2,n}, \qquad 3 \le n \le L-1, \tag{4.27}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{n,1} = (\mu_1 + \mu_2)\pi_{n+1,1} + \mu_2\pi_{n,2}, \qquad 3 \le n \le L-1, \tag{4.28}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{k,n} = \lambda\pi_{k-1,n} + \mu_1\pi_{k+1,n} + \mu_2\pi_{k,n+1}, \qquad 2 \le k \le n-2, \tag{4.29}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{n,k} = \lambda p_{n,k-1} + \mu_1 p_{n,k+1} + \mu_2\pi_{n+1,k}, \qquad 2 \le k \le n-2, \tag{4.30}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{n-1,n} = \lambda p_{n-2,n} + \beta\lambda\pi_{n-1,n-1} + \mu_1\pi_{n,n} + \mu_2\pi_{n-1,n+1}, \tag{4.31}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{n,n-1} = \lambda\pi_{n,n-2} + \beta\lambda\pi_{n-1,n-1} + \mu_1\pi_{n+1,n-1} + \mu_2\pi_{n,n}, \tag{4.32}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{n,n} = \lambda(\pi_{n,n-1} + \pi_{n-1,n}) + \mu_1\pi_{n+1,n} + \mu_2\pi_{n,n+1}, \tag{4.33}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{1,L} = \mu_1\pi_{2,L}, \tag{4.34}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{L,1} = \mu_2\pi_{L,2}, \tag{4.35}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{k,L} = \lambda\pi_{k-1,L} + \mu_1\pi_{k+1,L}, \qquad 2 \le k \le L-2, \tag{4.36}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{L,k} = \lambda\pi_{L,k-1} + \mu_2\pi_{L,k+1}, \qquad 2 \le k \le L-2, \tag{4.37}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{L-1,L} = \lambda\pi_{L-2,L} + \beta\lambda\pi_{L-1,L-1} + \mu_1\pi_{L,L}, \tag{4.38}$$

$$(\lambda+\mu_1+\mu_2)\pi_{L,L-1}=\lambda\pi_{L,L-2}+\alpha\lambda\pi_{L-1,L-1}+\mu_2\pi_{L,L}, \tag{4.39}$$

$$(\mu_1+\mu_2)\pi_{L,L}=\lambda(\pi_{L-1,L}+\pi_{L,L-1}). \tag{4.40}$$

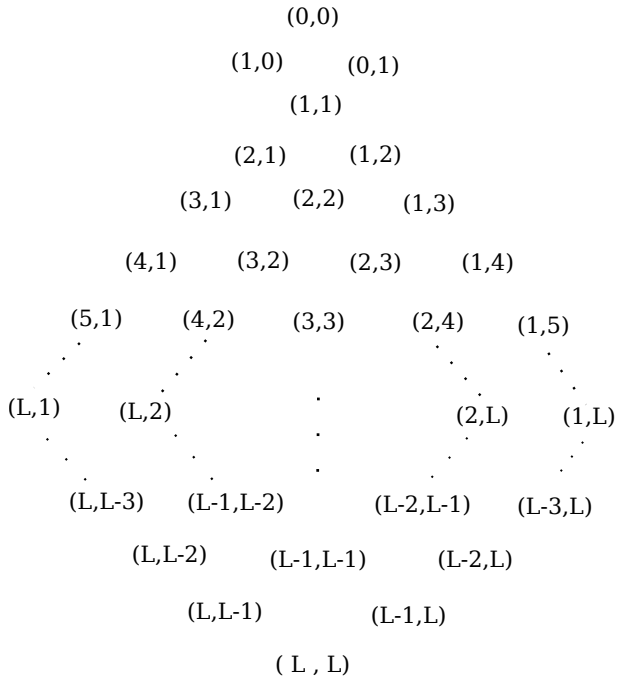


Figura 4.5: Diagrama para umbral N

Definimos el vector columna $\pi_k=(\pi_{1,k},\pi_{k,1},\pi_{2,k},\pi_{k,2},\ldots,\pi_{k-1,k},\pi_{k,k-1},\pi_{k,k})^t$, $k=1,2,3,...,L,$ y $\pi_k^+=(\pi_{0,1},\pi_{1,0},\pi_k)^t$, $k=1,2,3,...,L$. También, sea $\rho=\frac{\alpha}{\mu_1+\mu_2}$ y $\gamma_1=\frac{\mu_1}{\mu_1+\mu_2}$ y $\gamma_2=\frac{\mu_2}{\mu_1+\mu_2}$. Ignorando las ecuaciones (4.20)-(4.23) el sistema anterior del (4.24)-(4.26) puede ser re-escrito en la siguiente forma de matriz de bloques:

$$A_1\pi_1+B_2\pi_2+C_3\pi_3=0,$$

donde

$$A_1=\begin{pmatrix}\beta\rho\\ \alpha\rho\\ 0\end{pmatrix}, \qquad \mathrm{B}_2=\begin{pmatrix}-(1+\rho) & 0 & \gamma_1\\ 0 & -(1+\rho) & \gamma_2\\ \rho & \rho & -(1+\rho)\end{pmatrix}, \qquad \text{y}$$

$$C_3 \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \gamma_2 & \gamma_1 & 0 \end{pmatrix}.$$

Así por el *método de solución geométrica* [24] ya bien conocido en las subsecciones previas, el sistema anterior del (4.24)-(4.39) se pueden rescribir como

$$A_{k-1}\pi_{k-1} + B_k\pi_k + C_{k+1}\pi_{k+1} = 0, \qquad k = 2, 3, 3, ..., L-1, \tag{4.41}$$

$$A_{L-1}\pi_{L-1} + B_L\pi_L = 0, \qquad k = L, \tag{4.42}$$

y la ecuación de normalización

$$\sum_{i=0}^{L}\sum_{j=0}^{L}\pi_{i,j} = 1,$$

donde

$A_k = (a_{i,j})_{(2k+1)\times(2k-1)}, \quad k = 1, 2, ..., L-1,$

$a_{2k-1,2k-1} = \beta\rho, \qquad a_{2k,2k-1} = \alpha\rho \quad$ y $\quad a_{i,j} = 0$ en otro caso.

$B_k = (b_{i,j})_{(2k-1)\times(2k-1)}, \quad k = 2, 3, ..., L-1,$

con

$b_{2k-1,2k-2} + b_{2k-1,2k-3} = \rho, \qquad b_{2k-3,2k-1} = \gamma_1, \qquad b_{2k-2,2k-1} = \gamma_2, \qquad k = 2, 3, ..., L\ ,$

$b_{i,i} = -(1+\rho), \quad i = 1, 2, 3, ..., 2k-2,$

$b_{i+2,i} = \rho, \qquad i = 1, 2, 3, ..., 2k-4,$

$b_{2i-1,2i+1} = \gamma_1, \qquad b_{2i,2i+2} = \gamma_2, \qquad i = 1, 2, 3, ..., k-2,$

$b_{2k-1,2k-1} = -(1+\rho), \qquad k \neq L,$

$b_{2L-1,2L-1} = -1 \quad$ y

$b_{i,j} = 0 \qquad$ en otro caso. Finalmente

$C_k = (c_{i,j})_{(2k-3)\times(2k-1)}\ , \qquad k = 3, 4, ..., L,$

con
$c_{1,1} = c_{2,2} = 1,$

$c_{2k-3,2k-2} = \gamma_1, \qquad c_{2k-3,2k-3} = \gamma_2, \qquad k = 3, 4, ..., L,$

$c_{2i,2i} = \gamma_1, \qquad c_{2i-1,2i-1} = \gamma_2, \qquad i = 2, 3, ..., k-2, \quad$ y

$c_{i,j} = 0, \qquad$ en otro caso.

4.3.2. Resultados teóricos

En esta sección, formularemos la solución para la distribución estacionaria de el sistema (4.20)-(4.40). Claramente, usando las ecuación (4.41) y (4.42) la distribución estacionaria de el proceso de Markov tiene una matriz de solución y puede ser obtenida recursivamente resolviendo la ecuación (4.41). Más específicamente, tenemos primero que probar el siguiente lema que nos ayudará en la obtención de la distribución estacionaria del sistema anterior.

Lema 4.3.1 *Las inversas de* $B_k, \quad k = 2, 3, ..., L$ *existen y sus determinantes pueden ser calculados por*

$$det(B_k) = \left(-s + \frac{\gamma_1\rho}{d_{k-1}}\frac{\gamma_2\rho}{e_{k-1}}\right)\Pi_{i=1}^{k-1} d_i e_i, \qquad k = 2, 3, ..., L,$$

donde

$d_1 = -(1+\rho), \quad d_i = (1+\rho) - \frac{\gamma_1\rho}{d_{i-1}}, \quad i = 2, 3, ..., k-1,$

$e_1 = -(1+\rho), \quad e_i = (1+\rho) - \frac{\gamma_2\rho}{e_{i-1}}, \quad i = 2, 3, ..., k-1,$

$s = (1+\rho)$ *para* $k \neq L$ *y* $s = 1$ *para* $k = L$.

Demostración: Probamos el lema mediante el cálculo de la inversa B_k^{-1}. La idea del método es evaluar el determinante de B_k, esto es, reducir la matriz dada a la forma triangular superior por las operaciones elementales de fila, obtenemos

$$det(B_k) = \begin{vmatrix} -d_1 & 0 & \gamma_1 & 0 & & & & & 0 \\ 0 & -e_1 & 0 & \gamma_2 & 0 & & & & 0 \\ \vdots & 0 & -d_2 & 0 & \gamma_1 & 0 & & & \vdots \\ & \vdots & 0 & -e_2 & 0 & \gamma_2 & & & \\ & & \vdots & 0 & -d_3 & 0 & & & \\ & & & & \ddots & \ddots & \ddots & & \\ & & & & & & e_{k-1} & 0 & \gamma_2 \\ 0 & 0 & 0 & & & & & & -s + \frac{\gamma_1\rho}{d_{k-1}} + \frac{\gamma_2\rho}{e_{k-1}} \end{vmatrix}$$

donde

$d_1 = -(1+\rho), \qquad d_i = (1+\rho) - \frac{\gamma_1\rho}{d_{i-1}}, \quad i = 2,3,...,k-1,$

$e_1 = -(1+\rho), \qquad e_i = (1+\rho) - \frac{\gamma_2\rho}{e_{i-1}}, \quad i = 2,3,...,k-1,$

$s = (1+\rho)$ para $k \neq L$ y $s = 1$ para $k = L$.

Ahora simplemente tenemos que

$$det(B_k) = \left(-s + \frac{\gamma_1\rho}{d_{k-1}}\frac{\gamma_2\rho}{e_{k-1}}\right)\Pi_{i=1}^{k-1} d_i e_i,$$

Claramente, $det(B_k) \neq 0$ para cualquier valor de ρ, entonces B_k es invertible para $k = 2,3,...,L$. Por lo que se completa la demostración. □

Teorema 4.3.1 *Sea π_k^+ la única solución positiva del sistema (4.20)-(4.40), entonces*

$$\pi_k = \quad (-1)^{k+1} R_k A_{k-1} R_{k-1} A_{k-2} \cdots R_2 A_1 \pi_{1,1}, \quad k = 1,2,3,...,L, \tag{4.43}$$

$$\pi_{1,0} = \frac{\rho_1[\beta(\mu_1+\mu_2)+\lambda]}{(\mu_1+\mu_2)+2\lambda}\pi_{0,0}\ , \tag{4.44}$$

$$\pi_{0,1} = \frac{\rho_2[\alpha(\mu_1+\mu_2)+\lambda]}{(\mu_1+\mu_2)+2\lambda}\pi_{0,0}\ , \tag{4.45}$$

$$\pi_{1,1} = \frac{\rho_1\rho_2[\beta\mu_1+\lambda\mu_2)+\lambda]}{(\mu_1+\mu_2)+2\lambda}\pi_{0,0}\ , \tag{4.46}$$

con

$$\pi_{0,0} = \frac{(1+2\rho)(1-\rho)}{(1+2\rho)(1-\rho)+(1-\rho)[\rho_1(\beta+\rho)+\rho_2(\alpha+\rho)]+\rho_1\rho_2[\beta\gamma_1+\alpha\gamma_2+\rho](1-\rho^{2L-1})}, \qquad \rho \neq 1$$

$$\pi_{0,0} = \frac{3}{3+\rho_1\beta+\rho_2\alpha+\rho_1\rho_2+\rho_1\rho_2(1+\beta\gamma_1+\alpha\gamma_2)(2L-1)}, \qquad \rho = 1 \tag{4.47}$$

$\rho_1 = \lambda/\mu_1$, $\rho_2 = \lambda/\mu_2$, $R_1 = A_0 = 1$, $R_L = B_k^{-1}$, $\gamma_1 = \frac{\mu_1}{\mu_1+\mu_2}$, $\gamma_2 = \frac{\mu_2}{\mu_1+\mu_2}$ *y* $R_k = [B_k - C_{k+1}R_{k+1}A_k]^{-1}$, $k = L-1, L-2, ..., 3, 2.$,

Demostración: Por el lema 4.3.1, probamos que la matriz B_k es invertible. Ahora usando la ecuación (4.42) obtenemos:

$$\begin{aligned} B_L\pi_L &= -A_{L-1}\pi_{L-1}, \\ B_L^{-1}B_L\pi_L &= -B_L^{-1}A_{L-1}\pi_{L-1}, \qquad (B_L^{-1} existe) \\ \pi_L &= -B_L^{-1}A_{L-1}\pi_{L-1}, \end{aligned}$$

$$\pi_L = -\pi_L^{-1}A_{L-1}\pi_{L-1}, \qquad k = L \tag{4.48}$$

donde $R_L = B_L^{-1}$. Por lo tanto $R_{L-1}R_{L-2}, ..., R_2$ se pueden resolver recursivamente por la siguiente formula:

$$R_k = [B_k - C_{k+1}R_{k+1}A_k]^{-1}, \qquad k = L-1, L-2, ..., 3, 2.$$

Además. de la ecuación (4.41) obtenemos

$$\pi_k = -R_kA_{k-1}\pi_{k-1}, \qquad k = 2, 3, ..., L-1. \tag{4.49}$$

Por lo tanto $\pi_{L-1}, \pi_{L-2}, \pi_{L-3}, ..., \pi_1$ puede obtenerse recursivamente usando (4.48) y (4.49) en términos de $\pi_{1,1}$

$$\begin{aligned} \pi_2 &= -R_2A_1\pi_1 = \ -R_2A_1\pi_{1,1}\ , \\ \pi_3 &= -R_3A_2\pi_2 = \ R_3A_2R_2A_1\pi_{1,1}\ , \\ \pi_4 &= -R_4A_3\pi_3 = \ -R_4A_3R_3A_2R_2A_1\pi_{1,1}\ , \\ &\vdots \end{aligned}$$

por lo tanto,

$$\pi_k = (-1)^{k+1}R_kA_{k-1}R_{k-1}A_{k-2}\cdots R_2A_1\pi_{1,1}. \qquad k = 1, 2, 3, ..., L. \tag{4.50}$$

Ahora por simples manipulaciones algebraicas, las ecuaciones (4.44)-(4.46) pueden ser obtenidas por el sistema (4.20)-(4.22) en términos de $\pi_{0,0}$.

$$\begin{cases} \mu_1\pi_{1,0} - \mu_2\pi_{0,1,0} - \lambda\pi_{0,0} = 0 \\ \lambda\beta\pi_{0,0} - \mu_1\pi_{1,1} - (\lambda+\mu_2)\pi_{0,1} = 0 \\ \lambda\alpha\pi_{0,0} - \mu_2\pi_{1,1} - (\lambda+\mu_1)\pi_{1,0} = 0 \end{cases}$$

$$\begin{pmatrix} \mu_1 & 0 & -(\lambda+\mu_2) & -\lambda\beta \\ \mu_2 & -(\lambda+\mu_2) & 0 & \lambda\beta \\ 0 & \mu_1 & \mu_2 & -\lambda \end{pmatrix} \begin{pmatrix} \pi_{1,1} \\ \pi_{1,0} \\ \pi_{0,1} \\ \pi_{0,0} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Para resolver este sistema de ecuaciones, procederemos a utilizar el método de eliminación de Gauss, donde la matriz asociada al sistema de ecuaciones homogéneo está dada por:

$$\begin{pmatrix} 1 & 0 & -\frac{\lambda+\mu_2}{\mu_1} & \frac{\lambda\beta}{\mu_1} \\ 0 & -(\lambda+\mu_1) & \frac{\mu_2(\lambda+\mu_2)}{\mu_1} & +\lambda\beta - \frac{\lambda\mu_2\beta}{\mu_1} \\ 0 & 0 & \frac{2\lambda\mu_2+\mu_2^2+\mu_2\mu_1}{\lambda+\mu_1} & \frac{-\lambda(\mu_2\beta-\alpha\mu_1+(\lambda+\mu_1))}{\lambda+\mu_1} \end{pmatrix}.$$

Finalmente para completar la demostración, supongamos que $N = N_1 + N_2$, el número de clientes en el sistema y sea $g_k = Pr(N = k)$. Ahora, por inducción sobre k necesitamos mostrar que $g_k = \rho^{k-2} g_2$, $k = 2, 3, ..., 2L$.

De los siguientes casos surge: para el caso simple, usando (5.4), obtenemos

$$(\lambda+\mu_1+\mu_2)\pi_{1,1} = \lambda(\pi_{0,1}+\pi_{1,0}) + (\mu_1+\mu_2)(\pi_{1,2}+\pi_{2,1}),$$

$$(\lambda+\mu_1+\mu_2)g_1 = \quad \lambda g_1 \quad + \quad (\mu_1+\mu_2)g_3,$$

$$g_3 = \left(1 + \frac{\lambda}{\mu_1+\mu_2}\right) g_2 - \frac{\lambda}{\mu_1+\mu_2} g_1 .$$

Por lo que

$$g_3 = \rho g_2.$$

Similarmente , trabajando sobre las ecuaciones (4.24) y (4.25), obtenemos

$$(\lambda+\mu_1+\mu_2)(\pi_{2,1}+\pi_{1,2}) = \lambda\pi_{1,1} + (\mu_1+\mu_2)(\pi_{2,2}+\pi_{3,1}+\pi_{1,3}) ,$$

$$g_4 = (1+\rho)g_3 - \rho g_2 ,$$

$$g_4 = (1+\rho)\rho g_2 - \rho g_2 ,$$

$$g_4 = \rho^2 g_2.$$

De (4.25)

$$(\lambda+\mu_1+\mu_2)\pi_{2,2} = \lambda(\pi_{1,2}+\pi_{2,1}) + (\mu_1+\mu_2)(\pi_{3,2}+p_{2,3}) \ ,$$

$$(\lambda+\mu_1+\mu_2)g_2 = \lambda g_3 + (\mu_1+\mu_2)g_5 \ ,$$

$$g_5 = \left(1+\frac{\lambda}{\mu_1+\mu_2}\right)g_4 - \frac{\lambda}{\mu_1+\mu_2}g_3 \ ,$$

$$g_5 = (1+\rho)g_4 - \rho g_3 \ ,$$

$$g_5 = (1+\rho)\rho^2 g_2 - \rho^2 g_2 \ ,$$

$$g_5 = \rho^3 g_2.$$

En general para $k=2r$, remplazando n por $(2r-1),(2r-2),(2r-3),...,r$ en las ecuaciones (4.27)-(4.33) respectivamente e insertando, obtenemos:

$$(\lambda+\mu_1+\mu_2)\left[\sum_{i=1}^{r-1}\pi_{i,2r-i}+\sum_{i=1}^{r-1}\pi_{2r-i,i}+\pi_{r,r}\right] = \lambda\left[\sum_{i=1}^{r-1}\pi_{i,2r-i-1}+\sum_{i=1}^{r-1}\pi_{2r-i+1,i}\right]$$
$$+(\mu_1+\mu_2)\sum_{i=2}^{r}\pi_{i,2r-i+1}+(\mu_1+\mu_2)\sum_{i=2}^{r}\pi_{2r-i+1,i}+(\pi_{1,2r}+\pi_{2r,1}) \ ,$$

$$(\lambda+\mu_1+\mu_2)g_{2r} = \lambda g_{2r-1} \ , +(\mu_1+\mu_2)g_{2r+1}+(\mu_1+\mu_2)g_{2r+1}+g_{2r+1} \ ,$$

$$(\lambda+\mu_1+\mu_2)g_{2r} = \lambda g_{2r-1}+(\mu_1+\mu_2)g_{2r+1} \ ,$$

$$g_{2r+1} = (1+\rho)g_{2r} - \rho g_{2r-1} \ , \qquad r=3,4,...,L-1. \tag{4.51}$$

Trabajando del mismo modo para el caso $k=2r+1$, remplazando n por $2r,(2r-1),(2r-2),...,r+1$ se puede fácilmente establecer la siguiente relación:

$$g_{2r+2} = (1+\rho)g_{2r+1} - \rho g_{2r} \ , \qquad r=3,4,...,L-1. \tag{4.52}$$

De manera general, podemos escribir

$$g_k = (1+\rho)g_{k-1} - \rho g_{k-2} \ , \qquad k=3,4,...,2L. \tag{4.53}$$

Claramente, el caso para $k = 2L$ puede ser fácilmente obtenida usando las ecuaciones (4.38) y (4.39). Ahora podemos decir que la relación $g_k = \rho^{k-2} g_2$, es verdadera para algún $m \leq k$. Al hacer uso de la ecuación (4.49), obtenemos

$$\begin{aligned} g_{m+1} &= (1+\rho) g_m - \rho g_{m-1} \ , \\ &= (1+\rho)\rho^{m-2} g_2 - \rho\rho^{m-3} g_2 \ , \\ &= \left[(1+\rho)\rho^{m-2} - \rho\rho^{m-3}\right] g_2 \ , \\ &= \left[\rho^{m-2} + \rho^{m-1} - \rho^{m-2}\right] g_2 \ , \\ &= \rho^{m-1} g_2 . \end{aligned}$$

Por lo tanto la relación $\mathrm{g}_k = \rho^{k-2} g_2$, es verdadera para toda k, $k = 2, 3, ..., 2L$. Usando la ecuación de normalizadora $\sum_{k=0}^{2L} \mathrm{g}_k = 1$, obtenemos

$$\begin{aligned} g_0 + g_1 + g_2 + g_3 + g_4 + ... + g_{2L} &= 1 \ , \\ g_0 + g_1 + g_2 \sum_{k=2}^{2L} \rho^{k-2} &= 1 \ . \end{aligned}$$

Usando el hecho de que $\mathrm{g}_0 = \pi_{0,0}$, $\mathrm{g}_1 = \pi_{1,0} + \pi_{0,1}$, $\mathrm{g}_2 = \pi_{1,1}$[1] y usando las ecuaciones (4.44)-(4.46) en la ecuación pasada, el teorema queda establecido por completo. □

4.3.3. Casos particulares

1. Si ponemos $\alpha = \beta$ en el Teorema 4.3.1, obtenemos $\pi_{1,0} = \frac{1}{2}\rho_1\pi_{0,0}$, $\pi_{0,1} = \frac{1}{2}\rho_2\pi_{0,0}$ y $\pi_{1,1} = \rho_1\rho_2\pi_{0,0}$, donde

$$\pi_{0,0} = \begin{cases} \frac{2\mu_1\mu_2(1-\rho)}{2\mu_1\mu_2(1-\rho)+\rho(1-\rho)(\mu_1+\mu_2)^2+\lambda^2(1-\rho^{2L-1})}, & \rho \neq 1, \\ \frac{1}{1+L\rho_1\rho_2}, & \rho_1 = 1. \end{cases}$$

2. Además, otro resultado interesante puede ser obtenido por el sistema de colas espacio de tiempo de espera finito poniendo $\alpha = \beta$, $\mu_1 = \mu_2 = \mu$ y $\rho = \frac{\lambda}{2\mu}$, obtenemos $\pi_{1,0} = \rho\pi_{0,0}$ y $\pi_{1,1} = 2\rho^2\pi_{0,0}$, donde

$$g_n = \begin{cases} \frac{1-\rho}{1+\rho-2\rho^{2L+1}} & n = 0, \quad \rho \neq 1, \\ \frac{1}{1+4L}, & n = 0, \quad \rho = 1, \\ 2\rho^n\pi_{0,0}, & 1 \geq n \geq 2L, \end{cases}$$

[1] Recordemos que por el quinto supuesto en la descripción del modelo, el estado (2,0) y (0,2) nunca ocurren. Véase también el diagrama en la figura 4.5.

el cual da las probabilidades de estados estable para el modelo $M/M/2/2L$. En adición, tomando $L \to \infty$ uno puede fácilmente obtener la bien conocida distribución de estado estacionario para el espacio infinito de espera del sistema de colas $M/M/2$.

4.4. Conclusiones

Hemos sido capaces de calcular las probabilidades estacionarias de un sistema en paralelo el cual permite la estrategia *"salto"* cuando la diferencia de la longitud de las dos colas es suficientemente grande. Hemos resuelto explícitamente las probabilidades estacionarias para los casos específicos cuando $N = 1$ con tasas de servicio $\mu_1 = \mu_2$ y $3 \leq N < \infty$ con tasas de servicio $\mu_1 \neq \mu_2$ con la suposición de que si la longitud de las dos colas son iguales, una llegada de un cliente se une a cualquiera de las dos con probabilidad 0.5. Y el caso cuando hay una capacidad restringida $L \geq 1$, $1 < N \leq L$ con tasas de servicio $\mu_1 \neq \mu_2$ con la propiedad de que si la longitud de las dos colas son iguales, una llegada de un cliente se une a cualquiera de las dos filas con probabilidad α o a la segunda con probabilidad β, donde $\alpha+\beta = 1$.

En el primer caso ($N = 1$, $\mu_1 = \mu_2 = \mu$, $\alpha = \beta = 0.5$) solo fue necesario resolver un sistema de ecuaciones (no despreciando su grado de dificultad) con manipulación algebraica. Para el segundo caso ($3 \leq N < \infty$, $\mu_1 \neq \mu_2$, $\alpha = \beta = 0.5$) ya no nos fue posible resolver con solo manipulación algebraica, pues las ecuaciones de equilibrio formaban un sistema entre ellas con mucho más incógnitas que en el anterior. Así, por la *matriz generadora infinitesimal* que generó la cadena de Markov de PNM descrita, y después modificando sin alterar el proceso, pudimos utilizar el *método de la solución geométrica* estudiado en el capítulo de preliminares. Finalmente, para el último caso ($L \geq 1$, $1 \leq N \leq L$ y $\alpha + \beta = 1$), otra vez, por el *método de la solución geométrica* pudimos calcular las probabilidades estacionarias.

A todo esto, es importante mencionar la relevancia del *método de solución geométrica*, ya que de manera general permite obtener una solución explícita de las probabilidades estacionarias del modelo de dos colas en paralelo y sus variaciones (en N, μ_1, μ_2, α y β), para que pueda ser lo más parecido a la vida real.

CAPÍTULO 5

Valor de la información en un sistema de dos colas con "salto" umbral.

5.1. Introducción

Para los dos modelos descritos en el capítulo anterior, *salto instantáneo* y *no salto*, observemos que son los dos casos extremos, $N = 2$ y $N = \infty$, respectivamente. Para valores finitos de N estos modelos son más tratables que en el caso $N = \infty$ ya que la distribución estacionaria sigue un patrón de matriz geométrica donde el tamaño de la matriz es a lo más de $N \times N$. Una vez que la distribución estacionaria se conoce es posible calcular el número esperado de clientes en el sistema y el tiempo esperado de espera [18], [16].

La cuestión que se aborda, en particular en este capítulo, es un asunto diferente de las distribuciones estacionarias y tiempos de espera tradicionales. Esto es concerniente con la cuestión de el valor de la información. Supongamos que a su llegada un cliente no conoce cual cola (o fila) es más corta y por lo tanto él se une a cada cola con probabilidad 0.5. También supongamos que él tiene la opción de adquirir la información sobre cuál cola es más corta. Si él ejerce la opción, él se forma en la cola más corta. Esta opción no viene gratis: supongamos que el costo tiene algún valor el cuál, por conveniencia, es medido en unidades de tiempo. Así una llegada que busca su auto optimización compara el costo asociado con adquirir la información con la ganancia esperada de unirse a la fila más corta en lugar de hacerlo con una probabilidad 0.5. Por supuesto en el caso de *"salto" instantáneo* ($N = 2$) la información no tiene valor. Pero en otros casos si es importante. Cabe señalar que en nuestro modelo, los clientes informados y desinformados se comportan de la misma manera con respecto al *"salto"*.

Una observación fundamental para este modelo es que el *valor de la información* para un cliente (y por lo tanto la decisión de adquirir o no esta información) dependerá de lo que decidan los otros clientes. Así, el concepto de solución de referencia es de una *estrategia de equilibrio de Nash*. Al limitarnos a estrategias simétricas, una *estrategia de equilibrio de Nash* se caracteriza por un parámetro p, $0 \leq p \leq 1$, donde p es la probabilidad de adquirir la información. Por supuesto, cuando $p = 0$ quiere decir que nadie adquiere la información y $p = 1$ quiere decir que todo mundo adquiere la información. Si $p = 0$ o $p = 1$ la estrategia es llamada *estrategia pura*, en otro caso es llamada *estrategia mixta*.

Con el fin de encontrar una *estrategia de equilibrio de Nash*, primero tenemos que deducir el *valor de la información* para una llegada individual, es decir, la reducción en el tiempo de espera estimado debido a la adquisición de información, donde el resto de los clientes están usando la estrategia p. Llamaremos a este valor $g(p)$. El valor de la información tiene que ser comparado con el *costo de la información* (medido en unidades de tiempo) el cual es denotado aquí por C. Una probabilidad p, $0 < p < 1$ específica una *estrategia de equilibrio de Nash* si y sólo si $g(p) = C$. También si $g(p) \leq C$ entonces $p = 0$ es una *estrategia de equilibrio de Nash* y si $g(1) \geq C$ entonces $p = 1$ es una *estrategia de equilibrio de Nash*. Puede suceder que existan más de una *estrategia de equilibrio de Nash*.

Vamos a decir que el valor de la actual información para un cliente depende solo del estado del sistema a su llegada al tiempo t. Si los clientes que llegan más tarde son o no informados, no es de importancia para él. Por lo tanto, todo lo que importa para determinar el *valor de la información* para un cliente individual son las probabilidades que él ve en los estados donde la información es útil y la ganancia correspondiente a esos estados (el cual no es función de p). Por ejemplo, en el caso cuando $N = 3$ estos son los estados en los cuales la diferencia entre las dos longitudes de las colas es exactamente uno.

Otro aspecto que aparece en este problema, es la cuestión de las *externalidades* asociadas con la compra de información (véase definición B.0.2). Lo cual nuestra intuición nos dice que cuanto mayor sea la porción de llegadas que adquieren la información menos es el tiempo total de espera. Esto parecería que es verdad, ya que la probabilidad de que un servidor esté inactivo mientras que un cliente está esperando por un servicio tendría que disminuir. Sin embargo, cuando $C > 0$, no podemos decir, sin un análisis más cuidadoso que los clientes deben ser alentados a comprar más información. La razón es que la reducción en el tiempo de espera de los que compran la información puede ser debido a que otro tendrá que esperar más. Por lo tanto una cuestión cualitativa interesante es si el efecto (externalidad) de la acción individual debido a la adquisición de la información en otros es positiva o negativa. Si ésta es negativa, entonces para algunos posibles valores de C se optimiza de manera individual (es decir, se comportan como se prescribe por la estrategia de equilibrio), por lo que van a comprar más información de lo que es socialmente deseable. Para $N = 3$, obtenemos una expresión cerrada para las partes positivas y negativas de las *externalidades*. Tratar el problema de *externalidades* para un valor arbitrario de N parece ser una tarea mucho más difícil.

5.2. La distribución estacionaria

En esta sección mostraremos como calcular la distribución estacionaria para el modelo de dos colas en paralelo $M/M/2/FCFS$ cuando ocurre un *"salto"* entre ellas tan pronto como la diferencia entre las dos colas alcanza el valor de N para algún N dado, con $3 \leq N < \infty$.

Utilizaremos las técnicas de matriz geométrica descritas por Neuts [24] en la sección 2.4. La técnica requiere una descomposición en el espacio estado. Esta descomposición no es la única y la que hemos seleccionado es la más conveniente para nuestro propósito.

5.2.1. Descripción del modelo

- Supongamos que a su llegada un cliente no conoce cuál cola es más corta (cola no observable) y por lo tanto tiene dos opciones:

  1. unirse a cada cola con probabilidad 0.5 ó bien,
  2. adquirir la información sobre cuál de las colas es más corta para después formarse en ella.

  Así, los clientes que llegan toman una sola cola de espera basados en el orden de su llegada con tasas λ_1 y λ_2 respectivamente, donde $\lambda_1 = \lambda(1+p)/2$ y $\lambda_2 = \lambda(1-p)/2$.

- Suponemos que los dos servidores están distribuidos exponencialmente con tasas de servicio μ donde se supone una escala de λ y μ de tal manera que $\lambda + 2\mu = 1$

Sea $N_1(t)$ y $N_2(t)$ las variables aleatorias que describen el número de clientes al tiempo t en la cola 1 y 2 respectivamente. Entonces $\{(N_1(t), N_2(t)), t \geq 0\}$ es un proceso de Markov bidimensional con espacio de estados en un conjunto Ω. Sea N_1 y N_2 las variables aleatorias estacionarias, por lo que podemos denotar como la distribución estacionaria por

$$\pi_{i,j} = P\{N_1 = i, N_2 = j\} = \lim_{t \longrightarrow \infty} P\{N_1(t) = i, N_2(t) = j\}, \quad (i,j) \in \Omega.$$

Donde $\pi_{i,j}$ denota que i clientes están en una cola 1 y j clientes están en la cola 2 en cualquier momento [1]. Estos números incluyen a los clientes en servicio. Sin perdida de generalidad, supongamos que $i \leq j$. Para $i \geq 0$, sea $L(i)$[2] el conjunto de N estados (i,j) tal que $i \leq j \leq i+N-1$. Ordenando estos N estados por (i,i), $(i,i+1)$, ..., $(i, i+N-1)$ obtenemos el espacio de estados descrito de la siguiente forma:

$$\begin{aligned} L(0) &= \{\,(0,j)|\ 0 \leq j \leq N-1\} \\ L(1) &= \{\,(1,j)|\ 1 \leq j \leq N\} \\ &\vdots \\ L(i) &= \{\,(i,j)|\ i \leq j \leq i+N-1\} \\ &\vdots \end{aligned}$$

por lo que

$$\Omega = \{L(0) \cup L(1) \cup L(2) \cup \cdots \cup L(i) \cup \ldots\}.$$

[1] En general, no se es necesario etiquetar cada cola como 1 y 2, pues $\pi_{i,j}$ puede denotar la distribución estacionaria de que i clientes estén en una cola (no importa cual) y j clientes estén en la otra cola.

[2] Cada conjunto $L(i)$ tiene cardinalidad N, estos son los n niveles (ver sección 2.2)

Es decir

$$\Omega = \{(0,0),(0,1),(0,2),...,(0,N-1)\} \cup \{(1,1),(1,2),(1,3),...,(1,N)\} \cup$$

$$\{(2,2),(2,3),(2,4),...,(2,N+1)\} \cup \cdots \cup \{(i,i),(i,i+1),(i,i+2),...,(1,i+N-1)\} \cup \cdots$$

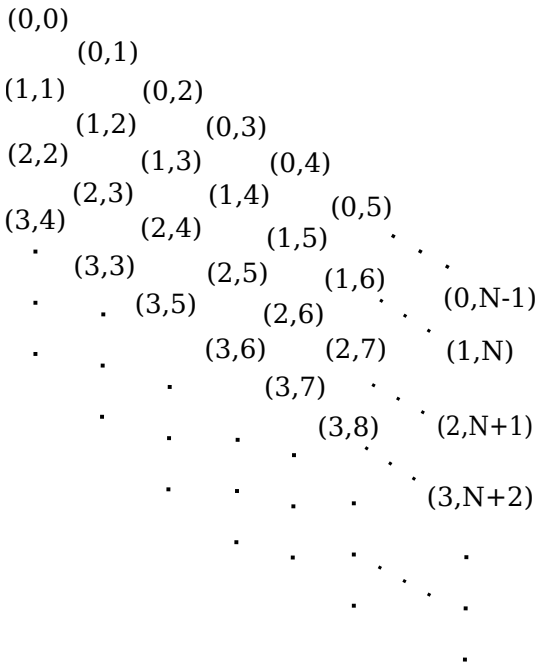


Figura 5.1: Diagrama para umbral $3 \leq N < \infty$

Sea π_i el vector renglón de la probabilidad estacionaria de los estados en $L(i)$ ordenados como antes. Entonces, estos vectores están particionados de la siguiente manera:

$$\pi_0 = (\pi_{0,0}, \pi_{0,1}, \pi_{0,2}, ..., \pi_{0,N-1}),$$

$$\pi_1 = (\pi_{1,1}, \pi_{1,2}, \pi_{1,3}..., \pi_{1,N}) \quad y$$

$$\pi_i = (\pi_{i,i}, \pi_{i,i+1}, \pi_{i,i+2}, ..., \pi_{i,i+N-1}), \quad i \geq 2.$$

Este PNM origina la *matriz generadora infinitesimal* Q que tiene la siguiente estructura tridiagonal en bloques:

$$Q = \begin{pmatrix} B_0 & C_0 & & & & & \\ B_1 & A_1 & A_0 & & & & \\ & A_2 & A_1 & A_0 & & & \\ & & A_2 & A_1 & A_0 & & \\ & & & A_2 & A_1 & A_0 & \\ & & & & \ddots & \ddots & \ddots \end{pmatrix}$$

Donde las submatrices A_0, A_1 y A_2 son matrices cuadradas de orden $2N$ donde $\lambda_1 = \lambda(1+p)/2$ y $\lambda_2 = \lambda(1-p)/2$, definidas de la siguiente manera:

$$A_0 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \lambda_1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & \lambda & 0 \end{pmatrix}$$

$$A_0(i,j) = \begin{cases} \lambda_1 & i = 2, ..., N-1, j = i-1 \\ \lambda & i = 1, j = N-1 \\ 0 & otro \quad caso \end{cases}$$

$$A_1 = \begin{pmatrix} -1 & \lambda & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \mu & -1 & \lambda_2 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & \mu & -1 & \lambda_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & \mu & -1 & \lambda_2 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & -1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & -1 & \lambda_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & \mu & -1 & \lambda_2 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 2\mu & -1 \end{pmatrix}$$

$$A_1(i,j) = \begin{cases} 1 & j = i, i = 1, ..., N \\ \lambda & i = 1, j = 2 \\ \mu & i = 2, ..., N-1, j = i-1 \\ \lambda_2 & i = 2, ..., N-1, j = i+1 \\ 2\mu & i = N, j = N-1 \\ 0 & otro \quad caso \end{cases}$$

$$A_2 = \begin{pmatrix} 0 & 2\mu & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & \mu & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \mu \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \end{pmatrix}$$

$$A_2(i,j) = \begin{cases} 2\mu & i = 1, j = 2 \\ \mu & i = 2, ..., N-1, j = i+1 \\ 0 & otro \quad caso \end{cases}$$

Y las matrices de la parte inicial

$$B_0 = \begin{pmatrix} -\lambda^* & \lambda^* & 0 & 0 & \dots & 0 & 0 \\ 2\mu & -1 & \lambda_2 & 0 & \dots & 0 & 0 \\ 0 & 2\mu & -1 & \lambda_2 & \dots & 0 & 0 \\ 0 & 0 & 2\mu & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & -1 & \dots & -1 & \lambda_2 \\ 0 & 0 & 0 & 0 & \dots & 2\mu & -1 \end{pmatrix}$$

$$B_1 = \begin{pmatrix} 0 & 2\mu & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \mu & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \mu & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & \mu \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

El PNM de Markov Q es irreducible y aperiódico trivialmente. Además, ya que los estados en los niveles mayores a N pueden comunicarse unos con otros vía sus rutas, no pasando de niveles $\geq N$, el generador $A_0 + A_1 + A_2$ es también irreducible. Así por el teorema 1.7.1 en el libro de Neuts [24] junto con el capítulo 2, sección 2.3 se puede obtener la condición de estabilidad.

En efecto, el proceso de Markov Q es ergódico si y solo si

$$\pi A_0 \mathbf{e} < \pi A_2 \mathbf{e} \tag{5.1}$$

donde $\mathbf{e}$ es el vector columna de unos y π es la solución de

$$\pi(A_0 + A_1 + A_2) = 0, \qquad \pi\mathbf{e} = 1.$$

NOTA: En [18], p. 425 se afirma que la distribución estacionaria existe para

$$\lambda < 2\mu$$

que se sigue de (5.1).

Ahora bien, vamos a resolver las ecuaciones de balance del PNM correspondiente

para la parte repetitiva:

Para $i \geq 1$, las transiciones de los estados en $L(i)$ puede llevarse acabo solo en los estados de $L(i-1)$, $L(i)$, o $L(i+1)$. Así para las matrices A_0, A_1 y A_2 en $\mathbb{R}^{N \times N}$, se cumple que

$$\pi_i A_0 + \pi_{i+1} A_1 + \pi_{i+2} A_2 = \mathbf{0}, \qquad i \geq 1 \tag{5.2}$$

Por lo tanto, π_j es una función solo de las tasas de transiciones entre estados con $j-1$ clientes en la cola y estados con j clientes en la cola. Luego, de la condición (2.12) de la sección 2.3

$$\pi_i = \pi_{i-1}\mathbf{R}\ , \qquad i \geq 2,$$

o bien

$$\pi_i = \pi_1 \mathbf{R}^{i-1}, \qquad i \geq 2, \tag{5.3}$$

Luego, sustituyendo (5.3) en (5.2), tenemos:

$$\pi_1 R^{i-1} A_0 + \pi_1 R^i A_1 + \pi_1 R^{i+1} A_2 = \mathbf{0}$$

De donde obtenemos la ecuación cuadrática:

$$R^2 A_2 + R A_1 + A_0 = \mathbf{0}. \tag{5.4}$$

donde R es llamada la *matriz de tasas* y ésta es la solución mínima no negativa de la ecuación cuadrática.

Para la parte inicial:

$$\pi_0 B_0 + \pi_1 B_1 \quad = \quad \mathbf{0}, \tag{5.5}$$

$$\pi_0 C_0 + \pi_1 A_1 + \pi_2 A_2 \quad = \quad \mathbf{0}, \tag{5.6}$$

y finalmente utilizando el factor de que las suma de las probabilidades es uno

$$\pi_0 \mathbf{e} + \pi_1 (I - R)^{-1} \mathbf{e} = 1 \tag{5.7}$$

donde I es la matriz identidad y $\mathbf{e}$ es el vector columna de puros unos.

5.2.2. Cálculo de la matriz de tasas R

Una alternativa para el cálculo de la *matriz de tasas* R, es utilizar su representación espectral (véase el apéndice C). Específicamente sean $w_1, w_2, ..., w_N$ los valores propios de R. Alguno de ellos pueden coincidir pero supondremos que la matriz R no es mal condicionada, es decir, existe una base ortonormal de $\mathbb{R}^{N\times N}$ para R formada por los vectores propios existentes de R. Por lo tanto, existen N matrices (de proyección) de rango 1 $E_1, E_2, ..., E_N$ con la propiedad de que $E_i E_j = 0$ si $i \neq j$ y $E_i E_i = E_i$, $RE_i = E_i R = w_i E_i$ para $1 \leq i, j \leq N$. Por lo que se puede representar a R como:

$$R = \sum_{i=1}^{N} w_i E_i \tag{5.8}$$

que es la *representación espectral de* R. Por lo tanto,

$$R^k = \sum_{i=1}^{N} w_i^k E_i, \quad k \geq 1. \tag{5.9}$$

Así estamos listos para calcular cualquier potencia de R.

5.2.3. Vectores de probabilidades límite

Si desglosamos la ecuación (5.3)

$$\begin{aligned} i = 0 \qquad & \pi_1 = \pi_0 R \\ i = 1 \qquad & \pi_1 = \pi_1 \\ i \geq 2 \qquad & \pi_i = \pi_1 R^{i-1} \end{aligned}$$

podemos observar que nos indica como calcular π_i, para $i \geq 2$ una vez que π_0 es conocido. Por lo tanto, dirijamos nuestra atención ahora al cálculo de π_0.

Con el fin de obtener los vectores de probabilidad limite π_0 y π_1, necesitamos resolver las siguientes ecuaciones:

$$\begin{aligned} \pi_0 B_0 + \pi_1 B_1 &= \mathbf{0}, \\ \pi_0 C_0 + \pi_1 (A_1 + R A_2) &= \mathbf{0}, \end{aligned} \tag{5.10}$$

o bien

$$[\ \pi_0 \quad \pi_1\ ] \begin{bmatrix} B_0 & C_0 \\ B_1 & A_1 + R A_2 \end{bmatrix} = \mathbf{0}.$$

Pero también necesitamos la condición de normalización

$$[\ \pi_0 \quad \pi_1\ ] \begin{bmatrix} \mathbf{e} \\ (I - R)^{-1}\mathbf{e} \end{bmatrix} = 1.$$

Por lo que entonces tenemos que resolver el sistema

$$[\ \pi_0 \quad \pi_1\ ] \begin{bmatrix} \mathbf{e} & B_0 * & C_0 \\ (I - R)^{-1}\mathbf{e} & B_1 * & A_1 + R A_2 \end{bmatrix} = [\ 1 \quad \mathbf{0}\ ].$$

donde M^* es M con la primer columna eliminada.

5.3. El tiempo de espera previsto

La siguiente observación, se cumple para cualquier valor de N para el *"salto umbral"*, esto es fundamental para nuestro análisis.

"En cualquier caso, el tiempo de permanencia (futuro) de un cliente depende del número total de clientes en el sistema y la cantidad de ellos que están delante de él en su propia fila. Sin embargo, dado estos dos números, el tiempo de permanencia no depende de la división de los clientes entre las dos filas".

Definición 5.3.1 *De acuerdo con la observación anterior, para $k \geq 0$, sea $M_{k,i}$ el tiempo de permanencia (futuro) estimado de un cliente que ve k clientes en frente de él (en esa fila) y un total de i clientes detrás de él en su fila.*

Note que bajo la suposición del modelo de si un cliente ve k clientes enfrente de él en su fila para algún k con $k > N-2$, entonces todas las posiciones en la otra fila desde la posición 1 y hasta la posición $k-N+2$ están ocupadas. Claramente, $M_{k,i} = (k+1)/\mu$ para $0 \leq k \leq N-2$, $i \geq 0$.

Antes de resolver para $M_{k,i}$ señalamos que nuestro interés es el cálculo del valor esperado de la información como una función de p, véase [16]:

$$g(p) = \sum_{k=0}^{\infty} \sum_{i=0}^{N-2} \pi_{k,k+i} \left[\frac{M_{k+i,N-i-2} + M_{k,N-2+i}}{2} - M_{k,N-2+i} \right]$$

$$= \sum_{k=0}^{\infty} \sum_{i=0}^{N-2} \frac{\pi_{k,k+i}}{2} \left(M_{k+i,N-i-2} - M_{k,N-2+i} \right). \tag{5.11}$$

Aquí nosotros mostramos como calcular los valores de la esperanza condicional y sus funciones generadoras. La misma función $g(p)$ se dará en las siguientes 2 secciones.

Si suponemos que $M_{-1,i} = 0$ entonces la siguiente ecuación en diferencias parcial está satisfecha por la $M_{k,i}$ (suponiendo $\lambda + 2\mu = 1$),

$$M_{k,i} = 1 + \lambda M_{k,i+1} + \mu M_{k,i-1} + \mu M_{k-1,i+1}, \quad k \geq 0, \quad i \geq 1. \tag{5.12}$$

También, si suponemos que para $k \leq N-2, M_{k-N+1,2N-3} = (k+1)/\mu$, entonces las condiciones de frontera son

$$M_{k,0} = 1 + \lambda M_{k,1} + \mu M_{k-1,1} + \mu M_{k-N+1,2N-3}, \quad k \geq 0 \tag{5.13}$$

$$M_{k,i} = \frac{k+1}{\mu}, \quad 0 \leq k \leq N-2, \quad i \geq 0. \tag{5.14}$$

y finalmente

$$\lim_{i \longrightarrow \infty} M_{k,i} \leq \frac{k+1}{\mu}, \quad k \geq 0. \tag{5.15}$$

Note que las ecuaciones (5.12)-(5.15) y por lo tanto su solución es independiente de p. Sin embargo, las probabilidades estacionarias y por lo tanto la esperanza incondicional del tiempo de espera y el valor de la información, son funciones de p. También notemos que la ecuación en diferencia (5.12) pero con diferentes condiciones de frontera aparece en [23] (ecuación (23)).

Ahora estamos listos para resolver (5.12)-(5.15).

Teorema 5.3.1 *La solución de las ecuaciones (5.12)-(5.15) están dadas por*

$$G_i(t) = \frac{1}{\mu(1-t)^2} + C_-(t)(f_-(t))^i + C_+(f_+(t))^i \tag{5.16}$$

para algunas funciones $C_-(t)$ y $C_+(t)$.

Demostración: Utilizando el método de funciones generadoras. Primero para $i \geq 0$, y para un número complejo t con $|t| < 1$, sea $G_i(t) = \sum_{k=0}^{\infty} M_{k,i} t^k$. Segundo, fijar un valor de i con $i \geq 1$. Entonces, multiplicando la k-ésima ecuación en (5.12) con t^k se suman éstas, y recordando que $M_{-1,i} = 0$, obtenemos

$$\sum_{k=0}^{\infty} M_{k,i}\, t^k = \sum_{k=0}^{\infty} t^k + \lambda \sum_{k=0}^{\infty} M_{k,i+1}\, t^k + \mu \sum_{k=0}^{\infty} M_{k,i-1}\, t^k + \mu \sum_{k=0}^{\infty} M_{k-1,i+1}\, t^k$$

$$G_i(t) = (1 + t + t^2 + \cdots) + \lambda G_{i+1}(t) + \mu G_{i-1}(t) + \mu(M_{-1,i+1} + M_{0,i+1}t + M_{1,i+1}t^2 + \cdots)$$

$$G_i(t) = \frac{1}{1-t} + \lambda\, G_{i+1}(t) + \mu G_{i-1}(t) + \mu(M_{0,i+1}t + M_{1,i+1}t^2 + \cdots)$$

$$G_i(t) = \frac{1}{1-t} + \lambda\, G_{i+1}(t) + \mu G_{i-1}(t) + \mu\, t G_{i+1}(t).$$

Por lo que

$$(\lambda + t\mu)G_{i+1}(t) - G_i(t) + \mu G_{i-1}(t) = \frac{-1}{1-t}, \quad i \geq 1. \tag{5.17}$$

Notemos que para un valor dado de t, la ecuación en diferencia (anterior) es ordinaria y de segundo orden. Una solución particular es:

$$G_i(t) \equiv \frac{1}{\mu(1-t)^2}, \quad i \geq 0.$$

Mirando a la parte homogénea vemos que ambas secuencias $(f_-(t))^i$ y $(f_+(t))^i$, $0 \leq i < \infty$, con

$$f_-(t) = \frac{1 - \sqrt{1 - 4\mu(\lambda + \mu t)}}{2(\lambda + \mu t)}$$

y

$$f_+(t) = \frac{1 + \sqrt{1 + 4\mu(\lambda + \mu t)}}{2(\lambda + \mu t)}$$

como raíces. Por lo tanto

$$G_i(t) = \frac{1}{\mu(1-t)^2} + C_-(t)(f_-(t))^i + C_+(t)(f_+(t))^i \tag{5.18}$$

para algunas funciones $C_-(t)$ y $C_+(t)$. □

Ya que supusimos $\lambda + 2\mu = 1$ entonces $\mu \leq 0.5$ y por lo tanto $|f_+(t)| \geq 1$ para cualquier t con $|t| \leq 1$ (con igualdad si y solo si $t = 1$). Esto unido con (5.11) implica que $C_+(t) = 0$. Igualmente $|f_-(t)| \leq 1$ para cualquier t con $|t| \leq 1$ (con igualdad si y solo si $t = 1$).

Por lo tanto todo lo que queda es encontrar el valor de $C_-(t)$. Esto se hará para utilizar (5.13). En realidad, multiplicando cada una de las ecuaciones en (5.13) por t^k y resumiendo que desde $k = 0$ a infinito, conduce a una identidad (con respecto a la variable t) la cual implica $G_0(t)$, $G_1(t)$ y $G_{2N-3}(t)$. Usando la forma $\frac{1}{\mu(1-t)^2} + C_-(t)(f_-(t))^i$ para $G_i(t)$ cuando i toma los valores 0, 1 y $2N-3$ en la identidad, conduce a una ecuación para una sola incógnita $C_-(t)$ cuyo valor a su vez se encuentra que es

$$C_-(t) = \frac{(N-1)t^{N-1}}{(1-t)[\mu t^{N-1}(f_-(t))^{2N-3} + (\lambda + \mu t)f_-(t) - 1]}. \tag{5.19}$$

5.4. El valor de la información y estrategias de equilibrio de Nash

Con el fin de calcular el valor de la información (véase la definición de $g(p)$ en la ecuación (5.11) es necesario de calcular primero el valor de alguna de las transformaciones $G_i(t)$ donde t corre sobre los valores propios no nulos de R. Específicamente, el valor de la información es la mitad de

$$\sum_{k=0}^{\infty}\pi_{k,k+1}(M_{k+1,N-3}-M_{k,N-1})+\sum_{k=0}^{\infty}\pi_{k,k+2}(M_{k+2,N-4}-M_{k,N}) \\ +\cdots+\sum_{k=0}^{\infty}\pi_{k,k+N-2}(M_{k+N-2,0}-M_{k,2N-4}).$$

Sea I el conjunto de valores propios no nulos de R y sea $(x)_i$ la i-ésima entrada del vector x. Entonces el valor de la suma anterior es igual a

$$\pi_{0,1}(M_{1,N-3}-M_{0,N-1})+\sum_{i\in I}(\pi_0E_i)_2\sum_{k=1}^{\infty}w_i^k(M_{k+1,N-3}-M_{k,N-1})+ \\ \pi_{0,2}(M_{2,N-4}-M_{0,N})+\sum_{i\in I}(\pi_0E_i)_3\sum_{k=1}^{\infty}w_i^k(M_{k+2,N-4}-M_{k,N})+\cdots+ \\ \pi_{0,N-2}(M_{N-2,0}-M_{0,2N-4})+\sum_{i\in I}(\pi_0E_i)_{N-1}\sum_{k=1}^{\infty}w_i^k(M_{k+N-2,0}-M_{k,2N-4})$$

que es igual a

$$\frac{\pi_{0,1}}{\mu}+\sum_{i\in I}(\pi_0E_i)_2\left\{\frac{1}{w_i}\left[G_{N-3}(w_i)-\frac{1}{\mu}-\frac{2w_i}{\mu}\right]-\left[G_{N-1}(w_i)-\frac{1}{\mu}\right]\right\}+$$

$$2\frac{\pi_{0,2}}{\mu}+\sum_{i\in I}(\pi_0E_i)_3\left\{\frac{1}{w_i^2}\left[G_{N-4}(w_i)-\frac{1}{\mu}-\frac{2w_i}{\mu}+\frac{3w_i^2}{\mu}\right]-\left[G_N(w_i)-\frac{1}{\mu}\right]\right\}+\cdots+$$

$$(N-2)\frac{\pi_{0,N-2}}{\mu}+\sum_{i\in I}(\pi_0E_i)_{N-1}\left\{\frac{1}{w_i^{N-2}}\left[G_0(w_i)-\frac{1}{\mu}-\frac{2w_i}{\mu}-\cdots-\frac{(N-1)w_i^{N-2}}{\mu}\right]-\left[G_{2N-4}(w_i)-\frac{1}{\mu}\right]\right\}.$$

Así, una tarea posterior (no se realiza en este trabajo) sería evaluar la función $g(p)$ para valores seleccionados de N y $\rho=\lambda/2\mu$. Poniendo todos los valores en la misma escala de tiempo y con $\mu\leq 0.5$ (recordemos $\lambda+2\mu=1$) para analizar el comportamiento del valor $g(p)$ respecto a p. Es de un particular interés, si g es una función monótona creciente o decreciente respecto a la proporción de personas que adquieren la información.

5.5. Las externalidades de comprar información para $N=3$

La acción de adquirir información para un individuo en la cual la fila es corta, tiene un efecto sobre el tiempo de espera en otros. Este efecto es la *externalidad* asociado con la acción. En esta subsección se

analizará las *externalidades* por la compra de información para $N = 3$. Tenga en cuenta que no es del todo claro si las *externalidades* son *positivas o negativas*. Por un lado la información puede ayudar a un individuo en rebasar a otros y afectarlos con una demora adicional (i.e., *externalidades negativas*), pero por otro lado puede conducir a una mejor utilización de los dos servidores y reducir el tiempo de espera de los otros clientes por lo tanto los clientes que de otro modo podrían haber esperado para este individuo, no tendrían que hacerlo (i.e., *externalidades positivas*). Observemos que definimos las *externalidades* de una acción sobre el resto de la sociedad [3] como la diferencia esperada entre el costo esperado del resto de los clientes cuando esta acción no se toma y cuando se toma. El efecto sobre el individuo que lleva a cabo la acción no se considera como parte de la *externalidad*.

Comenzando con una observación (que se cumple para $N = 3$):

A menos que el estado es (0,1), la acción de adquirir la información (a una llegada) afectará a nadie o solo a la próxima llegada.

A continuación, argumentamos que éste es de hecho el caso. Primero, es fácil ver que aquellos que están ya en el sistema no son afectados por la posibilidad de que una nueva llegada adquiera la información sobre el tamaño de la cola. Segundo, supongamos que a la llegada el estado es $(i, i+1)$ para algún $i \geq 1$. Note que éste es el único caso cuando la información tiene cualquier valor. Si el próximo evento es una salida, entonces el próximo estado y las suposiciones relativas a los clientes son las mismas, independientemente si la llegada actual se une a la cola más corta o la más larga. Si el próximo evento es una llegada, entonces no importa si el nuevo arribo adquiere la información o no, el estado $(i+1, i+2)$ se alcanzará. La única posible diferencia puede ser que la nueva llegada y la actual llegada intercambiarán lugares dependiendo si la actual llegada se une a la cola más corta o más larga. Los que llegan después no se ven afectados.

Considere de nuevo un estado $(i, i+1)$ para $i \geq 1$, si la actual llegada se une a la fila más corta y si el próximo evento es una llegada (este último con probabilidad λ), entonces el nuevo que viene tendrá un tiempo de espera estimado de $M_{i+1,1}$, mientras que si la actual llegada se une a la cola más larga, el valor correspondiente es $M_{i,3}$. Por lo tanto la externalidad negativa de que una llegada informada para el estado $(i, i+1)$ afecta sobre la próxima llegada (si el próximo evento es efectivamente una llegada) es $(0.5M_{i,3} + 0.5M_{i+1,1}) - M_{i+1,1} = 0.5(M_{i,3} - M_{i+1,1})$. Por lo tanto, la porción de la externalidad debido a la información adquirida y consecuentemente la posibilidad de intercambiar posiciones entre un cliente informado y una llegada a su lado es

$$EX_1 \equiv \sum_{i=0}^{\infty} 0.5\lambda(M_{i,3} - M_{i+1,1})(\pi_i)_2.$$

Por lo tanto,

[3] Entendemos por sociedad a todos los clientes que se encuentran en el sistema.

$$EX_1 = 0.5\lambda\ \pi_{0,1}(M_{0,3} - M_{1,1}) + 0.5\lambda \sum_{k=1}^{\infty}(M_{k,3} - M_{k+1,1}) \sum_{i\in I} w_i^k(\pi_0 E_i)_2$$

$$= -0.5\lambda\frac{\pi_{0,1}}{\mu} + 0.5\lambda \sum_{i\in I}(\pi_0 E_i)_2 \left(G_3(w_i) - \frac{1}{\mu}\right) - 0.5 \sum_{i\in I}(\pi_0 E_i)_2 \frac{1}{w_i}\left(G_1(w_i) - \frac{1}{\mu} - \frac{2w_i}{\mu}\right).$$

Una mejora en la utilización de los dos servidores debido al uso de información es una fuente de *externalidad positiva*. Como se señaló anteriormente, estos efectos pueden verse sólo cuando se enfrenta una llegada al estado (0,1). En otros estados, adquirir la información puede afectar la posición relativa de los clientes pero no afecta a la utilización de los servidores. Por lo tanto dirigimos nuestra atención al estado (0,1). Aquí muchas llegadas futuras pueden ser afectadas por esperar menos ya que el servidor se utiliza mejor cuando la llegada es informada. Ahora investigaremos la diferencia de tiempo (futuro) de espera total estimado cuando la llegada se une a la fila más larga para alcanzar el estado (0,2), o la fila más corta para alcanzar el estado (1,1). Esta diferencia viene dada que es posible que el mismo proceso aleatorio de los acontecimientos en el sistema bajo el último caso causará una transición del estado (1,1) en el estado (0,1) mientras que en virtud del estado en el primer caso (0,2) permanecerá como está. (En particular el servidor inactivo (0,2) será considerado como el que sirve a un cliente ficticio). Esto sucederá con probabilidad μ. En este caso, habrá un cliente más en el sistema que inicia con (0,2) en comparación con el sistema que inicia con (0,1) y éste será el caso siempre que los sistemas difieran en dos.

Por lo tanto, tenemos interés en calcular el tiempo estimado hasta que los estados en los dos sistemas coincidan (o se acoplen) ya que es el beneficio social de la adquisición de información. Más aún, para este valor tenemos que sustraer $M_{1,0} = 2/\mu$ la cual es la ganancia esperada del cliente en cuestión. Este cliente, si él selecciona la fila más corta, se coloca en segundo lugar en la fila en el momento inicial. (y entonces el estado (0,2) se alcanza), y no queremos incorporar su ganancia, mientras el cálculo de las externalidades. Para ver porque $2/\mu$ es justamente el valor a restar de la ganancia total, considere una llegada el estado (0,1). Por supuesto, el siguiente estado es (1,1) o (0,2) dependiendo de si él se une a la cola más corta o a la más larga. Tengamos en cuenta que en el primer caso, el servicio comienza en uno de los servidores a la llegada. Suponga que el siguiente evento (para ambos casos) es completar el servicio en éste servicio (una probabilidad μ evento). Hasta este evento, la última llegada ha pasado al mismo tiempo en ambos sistemas. Sin embargo, si se hubiera unido a la cola más larga, tendría él que permanecer en el sistema durante un tiempo adicional cuyo valor esperado es $2/\mu$.

Con el fin de encontrar el tiempo esperado hasta el acoplamiento, tenemos que incorporar el problema en uno más general: dados dos sistemas que se diferencia por un cliente, ¿cuál es el tiempo de espera hasta acoplarse dado que los dos sistemas están sujetos a los mismos sucesos (o eventos) aleatorios (llegadas, salidas y adquisición de información)?. Para $j \geq 3$, sea f_j el tiempo esperado hasta que se acoplan, cuando el número total inicial de un sistema es j mientras que en el segundo es $j + 1$. No es difícil ver que para $j \geq 3$ el tiempo esperado para acoplarse es invariante con respecto a como estos clientes se dividen entre las dos colas paralelas en los dos sistemas (siempre que las disciplinas de la cola se mantengan). Por lo tanto, para $j \geq 4$,

$$f_j = 1 + \lambda f_{j+1} + 2\mu f_{j-1}. \tag{5.20}$$

Para tres clientes o menos en total en el sistema menos congestionado, es necesario dar más detalles. Por lo tanto, denotamos su estado por un superíndice, mientras que el subíndice se refiere al sistema más congestionado. Por ejemplo $f_{2,1}^{2,0}$ es el tiempo esperado hasta acoplarse cuando el estado inicial de un sistema es (2,0), mientras que el estado inicial del otro es (2,1). Notemos que nuestro interés final es en $f_{2,0}^{1,0}$.
A continuación escribimos seis ecuaciones las cuales son satisfechas por estos valores.

$$f_3 = 1 + \lambda f_4 + \mu f_{2,1}^{2,0} + \mu f_{2,1}^{1,1}$$

$$f_{2,1}^{2,0} = 1 + \lambda f_3 + \mu f_{1,1}^{1,0}$$

$$f_{2,1}^{1,1} = 1 + \lambda f_3 + \mu f_{1,1}^{1,0} + \mu f_{2,0}^{1,0}$$

$$f_{1,1}^{1,0} = 1 + \lambda_1 f_{2,1}^{1,1} + \lambda_2 f_{2,1}^{2,0} + \mu f_1$$

$$(1-\mu) f_{2,0}^{1,0} = 1 + \lambda_1 f_{2,1}^{1,1} + \lambda_2 f_{2,1}^{2,0} + \mu f_1$$

$$(1-\mu) f_{0,1}^{0,0} = 1 + \lambda_1 f_{1,1}^{1,0} + \lambda_2 f_{2,0}^{1,0} + \mu f_1.$$

A continuación describimos como resolver las ecuaciones acopladas anteriores con la ecuación en diferencia (5.20). Luego la ecuación en diferencia definida en la ecuación (5.20) es resuelta por

$$\frac{j}{2\mu - \lambda} + k_1 + k_2 \rho^{-j}, \quad j \geq 3, \tag{5.21}$$

para alguna constante k_1 y k_2. Claramente, $k_2 = 0$, de lo contrario el valor esperado será (asintóticamente) un factor mayor que 1. Este por supuesto no es el caso, ya que en las condiciones de frontera, las f_j son como el tiempo hasta el primer éxito de cero hasta iniciar en j en un paso aleatorio con un desplazamiento diferente de cero a cero. Por lo tanto, para $j \geq 3$,

$$f_j = \frac{j}{2\mu - \lambda} \tag{5.22}$$

para alguna constante k. Insertando esta expresión para f_3 y para f_4 en las seis ecuaciones definidas anteriormente y obtener un sistema de seis ecuaciones lineales con las seis variables k, $f_{1,0}^{0,1}$, $f_{1,1}^{0,1}$, $f_{2,0}^{1,0}$, $f_{2,1}^{2,0}$ y $f_{2,1}^{1,1}$. Este sistema puede ser resuelto. En particular, el valor de $f_{2,0}^{1,0}$ es el que buscamos encontrar.

Podemos concluir aquí que la contribución al valor de las externalidades, debido a la adquisición de la información cuando se está en el estado (0,1) es

$$EX_2 \equiv \pi_{0,1} 0.5\mu \left[f_{2,0}^{1,0} - \frac{2}{\mu} \right].$$

Como se ha dicho anteriormente EX_2 es positivo. También notemos que $0.5\mu f_{2,0}^{1,0}\pi_{0,1}$ es la ganancia esperada de toda la sociedad, cuando una llegada adquiere la información.

5.6. Conclusiones

En la vida real, es posible que uno esté dispuesto a pagar más con el propósito de saber cuál cola es más corta. Así se introdujo el concepto de *valor de información* en un sistema de dos colas y obtenemos una expresión para determinar dicho valor de la información para una llegada individual en la cual una parte de los clientes compran la información. Mostramos también que con la ayuda del método de solución de *la matriz geométrica* el valor de la información puede calcularse para un sistema de dos colas paralelas sin memoria con un esquema de *"salto" umbral* para cualquier valor *umbral* N. Una vez que la distribución estacionaria se conoce será posible calcular el número de clientes en el sistema y el tiempo de espera.

También calculamos las *externalidades* asociadas con la información adquirida para un valor umbral $N = 3$ (son los estados en los cuales la diferencia entre las dos longitudes de las colas es uno) y en el cual fuimos capaces de descomponer las *externalidades* en dos fuentes: una para la posibilidad de que un cliente informante adelanta a otro cliente; el otro debido a la mejora del servidor cuando más clientes utilizan la información. Tratar el problema de externalidades para un valor arbitrario de N es una tarea mucho más difícil.

Finalmente, se deja como un problema a futuro, el cálculo con valores específicos (que cumplan con la condición de existencia de probabilidad estacionaria) para el valor de la información y el valor de las *externalidades* con la proporción de personas que adquieren la información, $0 \leq q \leq 1$.

CAPÍTULO 6

Conclusiones y Perspectivas

En el presente trabajo hemos estudiado dos tipos de modelos de colas: una cola con un servidor donde la tasa de servicio es creciente y dos colas en paralelo, cada una con su servidor, donde es permitido la estrategia *"salto"* entre ellas.

Modelo de una sola cola con tasa creciente

El primer modelo, fue analizado como un juego *no cooperativo*, donde los clientes representan los jugadores y cada uno de ellos busca su propio beneficio. Se estudió el comportamiento de los clientes al unirse a un sistema donde varía la tasa del servicio de manera creciente, en función de las personas que se van acumulando en la cola. Por lo anterior, nos dimos cuenta que nuestro modelo tenía una estructura dinámica, por lo que fuimos capaces de presentar las variables aleatorias que generan tal comportamiento respecto a reglas de decisión que los clientes pudieran tomar y respecto a las características que el sistema de espera mantiene (véase apéndice A.1). Después de realizar un acoplamiento entre el tiempo de servicio y técnicas de inducción pudimos así caracterizar las condiciones bajo las cuales se presenta un *equilibrio de Nash*, es decir cuando éste existe.

Modelo de dos colas paralelas con estrategia salto

En el segundo modelo nos enfocamos de manera principal, al análisis del cálculo de las probabilidades estacionarias, esto es, las dificultades que contiene este modelo (hay pocos ejemplos en la literatura) respecto a las variaciones en los valores: umbral (N), tasas de servicio (μ) y probabilidades de elegir una de las colas para unirse (α y β). Todo esto con el propósito de llevar este problema lo más posible a las distintas variaciones que existen en la vida real. No olvidando la importante herramienta que permitió este análisis y posteriormente el cálculo de las probabilidades estacionarias: el *método de la matriz geométrica* descrito por Neuts [24]. También, entre el análisis del cálculo de las probabilidades estacionarias, introducimos un concepto importante: el *valor de información* en un sistema de dos colas paralelas, donde la obtención de

las probabilidades estacionarias es diferente a los métodos tradicionales estacionarias y tiempos de espera.

Trabajo a futuro

Para el primer modelo, el trabajo futuro que se deja en este modelo, es encontrar a través de experimentos de simulación, valores tales como las medias empíricas y las probabilidades simuladas de entrada, y así mostrar si existe estrecha correspondencia en los tiempos de permanencia esperados y en las probabilidades de ingreso bajo las *políticas de equilibrio de Nash simétricas SNEP* en el juego estacionario asociado. Propiedades de convergencia y estabilidad en el caso de múltiples *equilibrios de Nash* no se conocen bien en este modelo, sin embargo, con experimentos de simulación podríamos saber si se sugiere que los *SNEPs* son viables de atracción y proporcionar una guía aproximada de los puntos de funcionamiento del sistema.

Para el segundo modelo el trabajo futuro que requiere, igualmente que en el primer modelo, es encontrar a través de experimentos de simulación estimar, el cálculo de las probabilidades estacionarias y analizar el comportamiento del *valor de la información* respecto a la proporción p (proporción de población que la adquiere).

APÉNDICE A

Teoría de colas

A.1. Descripción de un sistema de espera

En el lenguaje de la teoría de colas, un sistema de colas es un lugar donde los clientes llegan de acuerdo a un "proceso de llegada" para obtener los servicios de un centro de servicio. El centro de servicio pueden contener más de un servidor y se supone que un servidor puede servir un cliente a la vez. Si un cliente que llega encuentra todos los servidores ocupados, se une a una cola de espera. Este cliente recibirá su servicio más tarde, ya sea cuando llega a la cabeza de la cola de espera o de acuerdo a una cierta disciplina de servicio. Él sale del sistema al término de su servicio.
En términos generales un sistema de espera consiste de una unidad de servicio, un proceso de arribo de clientes que deben ser atendidos por la unidad y un proceso de servicio. El diagrama esquemático de un sistema de colas se representa en la Figura 2.1.

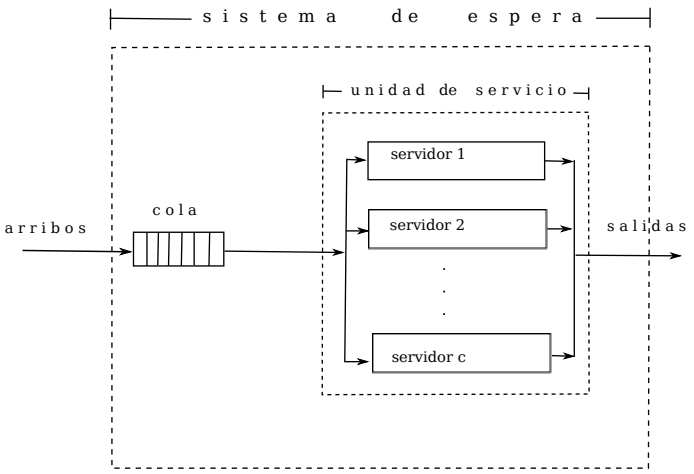


Figura A.1: Componentes de un sistema de espera

Entonces, dada una imagen dinámica de un sistema de colas, ¿cómo podemos describir analíticamente? ¿Cómo formular un modelo matemático que refleja esta dinámica? ¿Cuáles son los parámetros que caracterizan un sistema de colas completamente? Antes de continuar vamos a examinar la estructura de un

sistema de colas. Básicamente, un sistema de cola consta de tres componentes principales (véase [10]):

- Proceso de entrada
- La estructura del sistema
- El proceso de salida
- Diseño de la instalación

Características del proceso de entrada

Cuando hablamos sobre el proceso de entrada, de hecho, estamos preocupándonos por los siguientes tres aspectos del proceso de llegada:

(1) El tamaño de la población de llegada:
El tamaño de la población cliente que llega puede ser infinito en el sentido de que el número de clientes potenciales procedentes de fuentes externas es muy grande en comparación con los del sistema, de modo que la tasa de llegada no se ve afectada por el tamaño. El tamaño de la población que llega tiene un impacto en los resultados de colas. Una población infinita tiende a hacer que el análisis de colas más manejable y, a menudo capaz de ofrecer simples soluciones de forma cerrada, por lo tanto, este modelo se supone para nuestros sistemas de colas posteriores, salvo se indique lo contrario. Por otro lado, el análisis de un sistema de colas con el tamaño finito población cliente es más complicada debido a que el proceso de llegada se ve afectada por el número de clientes ya en el sistema.

(2) Patrones de llegada
Los clientes pueden llegar a un sistema de colas, ya sea en un patrón regular o de una forma totalmente aleatoria. Cuando los clientes llegan con regularidad en un intervalo fijo, el patrón de llegada puede ser fácilmente descrito por un único número - la tasa de llegada. Sin embargo, si los clientes llegan de acuerdo con algún modo aleatorio, entonces tenemos que ajustar una distribución estadística con el patrón de llegar a fin de hacer el análisis de colas matemáticamente factible.
El parámetro que se utilizan comúnmente para describir el proceso de llegada es el tiempo entre llegadas entre dos clientes. Por lo general, adaptarse a una probabilidad de distribución a fin de que podamos pedir a los vastos conocimientos de la teoría de la probabilidad. El patrón más comúnmente de llegada es el proceso de Poisson cuyo tiempos entre llegadas están distribuidos exponencial mente. La popularidad del proceso de Poisson reside en el hecho de que describe muy bien un patrón de llegada completamente al azar, y también conduce a resultados muy sencillo y elegante.
A continuación enumeramos algunas distribuciones de probabilidad que se utilizan comúnmente para describir el tiempo entre la llegada de un proceso de llegada. Estas distribuciones son generalmente denotado por una sola letra, como se muestra:

M: Markov (o sin memoria), implica el proceso de Poisson;
D: los tiempos entre llegadas constantes, los deterministas;
Ek: distribución de Erlang de orden K de los tiempos entre llegadas;
G: distribución de probabilidad general de tiempos entre llegadas;
GI: distribución general e independiente de tiempo entre llegadas.

(3) Comportamiento de los clientes que llegan:
Se dice que los clientes son impacientes si al llegar al sistema ven que este está lleno

Los clientes que llegan a un sistema de colas puede comportarse de manera diferente cuando el sistema está lleno debido a una cola de espera muy larga o cuando todos los servidores están ocupados. Si un cliente que llega encuentra que el sistema está lleno y deja siempre al sistema sin entrar, el sistema de formación de colas se conoce como bloqueo o si el tiempo que creen que les queda por esperar es suficientemente corto se conoce como renegase. Un tercer tipo de impaciente es el que va cambiando de cola entre colas paralelas, es decir

1. No querer unirse a ninguna cola (Balk)
2. Salirse de la cola por mucho tiempo de espera. (renegade)
3. Saltar entre colas para reducir tiempo (Jockey)

Es importante mencionar que la literatura únicamente considera los primeros dos tipos de clientes impacientes; los que no se unen a a cola o los que abandonan antes de tiempo. En particular, en este trabajo se analizara este tercer tipo de comportamiento utilizando nuevos métodos para su estudio. Cabe mencionar que al igual para los modelos de colas simples, debe cumplir con ciertas condiciones para que el sistema pueda ser estudiado (o tratable), es decir que cumpla con condiciones de estabilidad.

Características de la Estructura del Sistema

(1) Número de servidores (c):
En los sistemas de espera más simples se tiene sólo un servidor, $c = 1$, pero en muchas situaciones prácticas se tiene un número $c > 1$ de servidores. También se puede considerar sistemas con un número infinito de servidores. En tales sistemas no se forman colas porque siempre se dispone de servidores libres para atender a los nuevos clientes que llegan. Algunas veces, se puede usar un sistema con $c = \infty$ para aproximar un sistema con un número finito pero muy grande de servidores.

(2) La capacidad máxima del sistema (K):
La capacidad del sistema se refiere al número máximo de clientes que puede contener el sistema y en el caso de que haya c servidores se tiene que

$$K = C_q + c,$$

en donde C_q es la capacidad de la cola, es decir el número máximo de clientes que se permiten estar en la cola esperando servicio.
Hay sistemas de espera llamados "sistemas con perdida", en los $C_q = 0$, y por lo tanto $K = c$. El nombre se los sistemas de pérdida se debe al hecho de que $C_q = 0$ significa que si un cliente llega cuando todos los servidores están ocupados, no se les permite esperar y es rechazado, es decir, el cliente se "pierde".

Características del proceso de salida

(1) Disciplina de cola o disciplina de servicio:
Disciplina de colas, a veces conocido como disciplina de servicio, se refiere a la forma en que los clientes en la cola de espera se seleccionan para el servicio. En general, tenemos:

- En primero en llegar, es el primero en ser servido (FCFS)

- El último que llega es al primero que se atiende (LCFS)
- Prioridad
- Procesador de intercambio
- Aleatorio.

La disciplina de cola FCFS no asigna las prioridades y atiende a clientes en el orden de sus llegadas. Al parecer esta es la disciplina más frecuente en una cola ordenada. La disciplina LCFS es justo lo contrario de la FCFS. Los clientes que llegan al último se sirven en primer lugar. Este tipo de disciplina de encolamiento se encuentra comúnmente en las operaciones de la pila, donde los puntos (en nuestra terminología a los clientes) se apilan y las operaciones se producen sólo en la parte superior de la pila. En la disciplina de colas de prioridad, los clientes se dividen en varias clases de prioridad de acuerdo a sus prioridades asignadas. Aquellos que tienen una mayor prioridad que otros se sirven antes que otros que reciben su servicio. Hay dos subclasificaciones: de suscripción preferente y no preferente.
En el procesador de intercambio, la capacidad se dividen por igual entre todos los clientes en la cola, que es cuando hay k clientes, el servidor dedica $1/k$ de su capacidad de cada uno. Del mismo modo, cada cliente obtiene el servicio en $1/k$ de velocidad y sale del sistema al término de su servicio.

A.2. Tiempo de servicio residual

Ahora dirigimos nuestra atención a otro enfoque de análisis - *Tiempo de servicio residual*. En este enfoque, nos fijamos en las épocas de llegada en lugar de las épocas de salida y de obtener el tiempo de espera de un cliente que llega [10].
Considere el instante en que un nuevo cliente (por ejemplo el i-ésimo cliente) llega al sistema, el tiempo de espera en la cola de este cliente debe ser igual a la suma de los tiempos de servicio de los clientes por delante de él en la cola y el tiempo de servicio residual del cliente actualmente en servicio. *El tiempo de servicio residual* es el tiempo restante hasta la finalización de servicios del cliente en el servicio y se denota como r_i para poner de relieve el hecho de que este es el momento que el cliente i tiene que esperar al cliente en el servicio para completar su servicio, como se muestra en la Figura A.2.

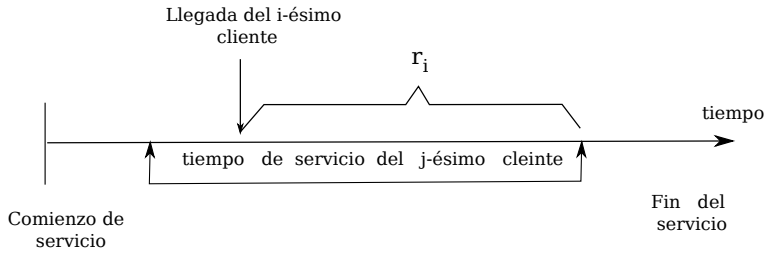


Figura A.2: Tiempo de servicio residual

El tiempo restante de servicio r_i es igual a cero si no hay clientes en el servicio cuando el i-ésimo cliente llega.

Supongamos que hay n clientes delante de él en la cola de espera, entonces el tiempo de espera (w_i) se da como

$$w_i = u(k)r_i + \sum_{j=i-n}^{i-1} x_j \tag{A.1}$$

donde $u(k)$ se define como sigue para tener en cuenta ya sea un sistema que se encuentra vacío o un sistema con k clientes:

$$u(k) = \begin{cases} 1 & k \geq 1 \\ 0 & otro\ caso \end{cases} \tag{A.2}$$

A.3. Notación Kendall

Vemos que hay muchos procesos estocásticos y una multiplicidad de parámetros (variables aleatorias) que participan en un sistema de colas, por lo que dada una situación tan compleja, ¿cómo clasificarlos y describirlos en una forma abreviada matemática? David G Kendall, un estadista británico, ideó una notación abreviada, se muestra a continuación, para describir un sistema de colas que contiene una sola cola de espera. Esta notación se conoce como *notación de Kendall*:
donde

A / B / X / Y / Z

A: Distribución de tiempos de llegada;
B: distribución de tiempo de servicio;
X: número de servidores;
Y: la capacidad del sistema;
Z: disciplina de la cola.

Por ejemplo, $M/M/1/\infty/FCFS$ representa un sistema de colas, donde los clientes llegan de acuerdo a un proceso de Poisson y los tiempos de servicio del servidor tienen una distribución exponencial. El sistema tiene sólo un servidor, una cola infinita de espera y los clientes son atendidos de forma FCFS. En muchas situaciones, sólo utilizamos los tres primeros parámetros, por ejemplo, $M/D/1$. Los valores predeterminados para los dos últimos parámetros son $Y = \infty$ y $Z = FCFS$.

APÉNDICE B

Teoría de juegos

La teoría de juegos es una herramienta que ayuda a analizar problemas de optimización interactiva. La mayoría de las situaciones estudiadas por la teoría de juegos implican conflictos de intereses, estrategias y trampas.

La teoría de juegos se divide en dos ramas, la *cooperativa* y la *no cooperativa*. La distinción puede ser muy difícil a veces, pero esencialmente podemos decir que en la teoría de los *juegos no cooperativos* el individuo que participa en el juego tratando de obtener lo "máximo" posible para él, sin cooperar con el resto de los jugadores, sujeto a las reglas y posibilidades claramente definidas. En otro caso, si los individuos muestran una conducta cooperativa esto nos llevaría, obviamente, a la situación de *juegos cooperativos* (véase [13]).

Hay dos formas o tipos básicos de modelos formales empleados en la teoría de juegos no cooperativos. La primera y más simple es una forma estratégica o un juego de forma normal. Esta clase de modelo tiene tres elementos:

a) una lista de *jugadores*;
b) para cada jugador, una lista de *estrategias*; y
c) para cada conjunto de estrategias (una para cada jugador), una lista de *pagos* que reciben los jugadores.

Cuando un jugador tiene en cuenta las reacciones de otros jugadores para realizar su elección, se dice que el jugador tiene una *estrategia*. Una estrategia es un plan completo de acciones que se llevan a cabo cuando se desarrolla el juego. Se prescribe, antes de que comience el juego, cada decisión que los jugadores deben tomar durante el transcurso de éste, dada la información disponible para el jugador. La estrategia puede incluir movimientos aleatorios.

El segundo tipo de modelo que se utiliza en la teoría de juegos no cooperativos es el *juego de forma extensa*. En un juego de forma extensa se presta atención al tiempo de las acciones que pueden realizar los jugadores, y a la información que tendrán cuando deban realizar tales acciones.

Por tanto, tenemos dos clases diferentes de modelos en un juego: el modelo de forma *estratégica* y el modelo de forma *extensa*. ¿Cuáles conexiones existen entre las dos? Para cada juego de forma extensa hay un juego de forma estratégica correspondiente, donde imaginamos a los jugadores escogiendo simultáneamente estrategias que pondrán en práctica. Pero un juego de forma estratégica puede corresponder a varios juegos de forma extensa diferentes (véase [21]).

Si intentamos predecir los resultados de los juegos, podríamos hacernos las siguientes preguntas: ¿qué acciones pueden descartar los jugadores?, y ¿existen en estos juegos un modo obvio de jugarlos? Lo que cada jugador se supone que va a hacer, debe ser la mejor respuesta a lo que se supone que harán los otros jugadores.

Definición B.0.1 *Un* ***juego simétrico*** *es un juego en el que las recompensas por jugar una estrategia en particular depende sólo de las estrategias que empleen los otros jugadores y no de quién los juega. es decir, las identidades de los jugadores pueden cambiarse sin que cambie las recompensas de las estrategias.*

Definición B.0.2 *Se puede definir una* ***externalidad*** *como la situación en la cual los costos o beneficios de producción y/o consumo de algún bien o servicio no son reflejados en el precio de mercado de los mismos. En otras palabras, son externalidades aquellas actividades que afectan a otros para mejorar o para empeorar, sin que éstos paguen por ellas o sean compensados.*

Las ***externalidades negativas*** *son las que llevan a los mercados a producir cantidades superiores a las socilamente deseables. Las* ***externalidades positivas*** *llevan a los mercados a producir cantidades inferiores a las socialmente deseables.*

APÉNDICE C

Descomposición espectral de una matriz real

Las matrices simétricas reales constituyen uno de los tipos más importantes de matrices para las cuales puede garantizarse la diagonalización. Además, dicha diagonalización se puede obtener matrices de paso ortogonales.

Diagonalización

Teorema C.0.1 *Sea A uma matriz real simétrica. Entonces:*
(a) Todos los autovalores de A son reales.
(b) Si v_1 y v_2 son autovectores (reales) de A asociados a autovalores distintos λ_1 y λ_2, entonces v_1 y v_2 son ortogonales.

Teorema C.0.2 ***(espectral para matrices simétricas)*** *Sea A una matriz cuadrada real $n \times n$. Son equivalentes:*
(a) A es simétrica.
(b) A es diagonalizable mediante una matriz de paso ortogonal, es decir, existe una matriz ortogonal Q tal que $Q^{-1}AQ = Q^TAQ = D$ es diagonal.

En este caso, las columnas de la matriz $\{q_1, ..., q_2\}$ de Q son un conjunto de **autovectores** de A que forman una **Base Ortonormal** de $\mathbb{R}^n$ y, ademas, tenemos que

$$A = QDQ^T = \left[q_1 \middle| \dots \middle| q_2 \right] \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix} \begin{bmatrix} q_1^T \\ q_2^T \\ \vdots \\ q_n^T \end{bmatrix} \tag{C.1}$$

$$= \lambda_1 q_1 q_1^T + \lambda_2 q_2 q_2^T + \cdots + \lambda_n q_n q_n^T$$

Cada matriz $q_k q_k^T$ es la matriz de proyección ortogonal sobre el subespacio generado por el correspondiente vector q_k (es una matriz de rango 1). Así obtenemos la expresión

$$A = \sum_{k=1}^{n} \lambda_k q_k q_k^T \tag{C.2}$$

que se llama **descomposición espectral** de A. Esta expresión nos da la matriz simétrica real A como una combinación lineal de matrices de proyección de rango 1. El conjunto de valores propios se llama el espectro de A. Si dos o más valores propios de A son idénticos, el espectro de la matriz se llama **degenerada**. La nomenclatura "espectro" es una analogía exacta con la idea del espectro de la luz como se muestra en un arco iris. El brillo de cada color del espectro nos dicen "cuánta" luz de longitud de onda que existía en la luz blanca no disperso. Por esta razón, el procedimiento se refiere a menudo como una descomposición espectral.

A la hora de obtener una diagonalización ortogonal de una matriz simétrica real A pueden aparecer dos situaciones distintas:

- *Todos los valores de A son simples.* En este caso, los autovectores correspondientes tienen que ser ortogonales dos a dos y formarán una base ortogonal de $\mathbb{R}^n$. Normalizando dichos autovectores (dividiendo cada uno con su norma) seguiremos teniendo autovectores ortogonales que además serán unitarios. Una matriz Q que tenga (como columnas)0 a dichos autovectores ortonormales será una matriz de paso que diagonaliza A ortogonalmente.
- *La matriz A tiene algún autovalor de multiplicidad m.* En este caso, cuando calculamos los autovectores asociados a uno de los autovalores λ, obtenemos una base del espacio propio asociado $Nul(A - \lambda I)$. En general esta base no es una base ortogonal de dicho subespacio. Ortogonalizando primero y normalizando a continuación, tendremos una base ortonormal de autovectores asociados a dicho autovalor. Haciendo esto con cada uno de los autovalores múltiples y normalizando los autovectores asociados a autovalores simples tendremos una base ortonormal de $\mathbb{R}^n$ formada por autovectores de A. Basta considerar una matriz Q cuyas columnas sean los vectores de dicha base para obtener una diagonalización ortogonal de A.

Este es un resultado útil, ya que ayuda a facilitar el cálculo de las matrices A^{-1}, $A^{1/2}$ y $A^{-1/2}$.

$$A^{-1} = QD^{-1}Q^T = \sum_{k=1}^{n} \frac{1}{\lambda_k}\, q_k q_k^T$$

$$A^{1/2} = QD^{1/2}Q^T = \sum_{k=1}^{n} \sqrt{\lambda_k}\, q_k q_k^T$$

$$A^{-1/2} = QD^{-1/2}Q^T = \sum_{k=1}^{n} \frac{1}{\sqrt{\lambda_k}}\, q_k q_k^T$$

Bibliografía

[1] Adan, I.J.B.F., Wessels, J. and Zijm, W.H.M. Matrix-geometric analysis of the shortest queue problem with thereshold jockeying. Operations Research Letters, No. 13, pp. 107-112. 1993.

[2] Altman, E. Aplications of dynamic games in queues. Advances in dynamic games. No. 7, Part IV, pp. 309-342. 2005.

[3] Altman, E. and Shimkin, N. Individual equilibrium and learning in processor sharing system. Operations Research, No. 6, pp. 776-784. 1998.

[4] Ancker, C. J. Jr. and Gafarian, A. V. Some Queueing Problems with Balking and Renging. Operations Research, No. 11. Part II. 1963.

[5] Barred, D. Y. Queueing with impatient customers and ordered service. Oprerations Research, No. 5, pp. 650-656. 1999.

[6] Brooms, A.C. On the Nash equilibria for the FCFS queueing system with load-increasing service rate. Advances in Applied Probability No. 37, pp. 461-481. 2004.

[7] Cohen, J. W. Analysis of the assymetrical shorter two-server queueing model. Appl. Math. Stoch. Anal. No. 11 (2), pp. 115-162. 1998.

[8] Cohen, J. W. and O. J. Boxman. *Boundary Value Problems in Queueing System Analysis.* North-Holand Publish Company Amsterdam. 1983.

[9] Cohen J. W. On the analysis of the symmetrical shortest queue. Report BS-R9420, Amsterdam. 1994.

[10] Chee-Hock, Ng and Boon-Hee, Soong. *Queueing Modelling Fundamentals With Applications in Communication Networks.* Segunda edición. Edt. John Wiley Sons Ltd. 2008.

[11] Fayolle, G. and Iasnogorodski, R. Two couple processes: the reduction to a Riemann-Hilbert problem. Gebiete No. 47, pp. 325-351. 1979.

[12] Flatoo, L. and McKean, H. P. Two queues in parallel. Comm. Pure. Appl. Math. No.30, pp. 255-263. 1977.

[13] Fundenberg, D. and Tirole, J. *Game Theory.* Cambridge: MIT Press. 1991.

[14] Gertsbakh, I. Shorter queue problem: a numerical study using tha matrix-geometric solution. European Journal of Operations Research. No. 15, pp. 374-381. 1984.

[15] Haight, F.A. Two queues in parallel. Biometrika No. 45, pp. 401-410. 1958.

[16] Hassin, R. and Haviv, M. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems.* Kluwer International Series. Kluwer, Dordrecht. 2003.

[17] Hassin, R. and Haviv, M. Nash equilibrium and subgame perfection in observable queues. Annals of operations research No. 113, pp. 15-26. 2002.

[18] Hassin, R. and Haviv, M. Equilibrium strategies and the value of information in a two line queueing system with thereshold jockeying, Statist-Stochastic Model. No. 10, pp. 415-435. 1994.

[19] Hiller, Frederick S. and Lieberman, Gerald J. *Investigación de operaciones.* Editorial McGraw Hill. 2002.

[20] Kingman, J.F.C. Two similar queues in parallel. Ann. Math. Stat. No. 32, pp. 1314-1323. 1961.

[21] Kreps, D. *Curso de Microeconómica.* McGraw-Hill. 1995.

[22] Lippman, S. Applying a new device in the optimization of exponential queuing systems. Operations Research, No. 23, pp. 687-709. 1975.

[23] Maekawa, M. Queueing models for computer system conected by communication line. The Journal of thr Association for Computing Machinery. No. 24, pp. 566-582. 1997.

[24] Neuts, M. F. *Matrix-Geometric Solution in Stochastic Models: An algotithmic Approach.* The Johns Hopkins University Press, Baltimore. 1991.

[25] Pérez, N. J, Jimeneo, P. J y Cerdá, T. E. *Teoría de Colas.* Editorial Pearson Prentice Hall. 2005.

[26] Ramswami, V and Latouche, G. A General Class of Markov Processes with Explicit Matrix-Geometric Solutions. OR Spektrum, No. 8, pp. 209-218. 1986.

[27] Sheldon, R. M. *Introduction to probability Models.* 9th Edición. 2007.

[28] Tarabia, A.M.K. Analysis of two queues in parallel with jockeying and restricted capacities. Applied Mathematical Modelling. No. 32, pp. 802-810. 2008.

[29] Van den Berg, L. Stochastic comparison of Markov queueing networks using coupling. Master thesis. May 17, 2010.

[30] Van Leeuwaarden, J. S. H and Winands, E. M. M. Quasi-birth-death processes with an explicit rate matrix. Departament of Mathematics and Computer Science and Department of Technology Management. May 2004.

[31] Willing, A. A short introduction to queueing theory. July 21, 1999.

[32] Yiqiang, Q. Zhao and Winfried, K. Grassman. Queueing analysis of a jockeying model. Operations Research. Vol. 43, No. 3, pp. 520-529. 1996.