Modelado de sistemas informaticos y de telecomunicación

Eitan Altman,
INRIA, BP93, 2004 Route des Lucioles 06902 Sophia-Antipolis Cedex, France
y C.E.S.I.M.O., Universidad de Los Andes, Faculdad de Ingeneria, Merida, Venezuela.

November 1, 2002

Contenido

Capítulo 1

Introducción

Estas notas estan destinadas para estudiantes con un poquito de conocimientos de teoria de probabilidad que quieren aprender las bases de modelización aleatoria de sistemas informáticos. El libro corresponde a un curso de 28 horas y cubre nociones básicas de la teoria de colas, la modelización de redes locales y de la internet. Para profundizar en la teoria de colas, pueden leer [13, 6]. Más sobre la modelización de redes locales se encuentra en [8, 14]. Sobre la modelisación de la internet, pueden ver en [1, 3, 4].

En este curso, presentamos herramientas para primero representar una sistema informático por un modelo que tiene las características de base del sistema original pero es más simple, y después, usando este modelo, mostramos como analizar su rendimiento. Las medidas de rendimiento que son importantes son las esperanzas y distribuciones de tiempo de espera, de retardos, de probabilidad de perdidas de paquetes en sistemas (causado por la congestión), etc.

Porque necesitamos modelos? Con el sistema original podemos tratar de medir el rendimiento. Pero, a menudo

- tales experimentaciones necesitan mucho tiempo,
- es difícil o imposible cambiar los parámetros en el sistema real para experimentar su influencia
- A veces podemos experimentar solamente el comportamiento global, y no lo que pasa dentro del sistema.

Un ejemplo es la internet. Podemos experimentar las características del tráfico en la fuente, pero es difícil de medir el tráfico en los enrutadores. Es dificil o imposible cambiar los parámetros de los enrutadores que pueden estar en otros paises.

Cuando tenemos un modelo podemos estudiarlo con simulaciones o con herramientas matemáticas. La teoria de colas es una herramienta matemática que usa métodos probabilisticos. La teoria de colas ha tenido mucho exito en el análisis de sistemas informáticos. A veces, una sola ecuación puede darnos la distribución o la esperanza de retardos, tiempo de espera o tasa de pérdidas. Un ejemplo de esto son las redes de Jackson, que pueden contenir decenas de colas (que representan enrutadores o conmutadores), donde la esperanza de retardos o tiempo de espera puede representarse por una ecuación sencilla.

Otro exito de los modelos matemáticos de colas ha sido en las redes telefónicas, que son modeladas como "redes de perdidas". Estos modelos permiten frecuentemente estudiar varias políticas de enrutamiento de conexiones telefónicas para minimizar la probabilidad de encontrar la linea ocupada.

La mayoria de modelos que usamos son probabilisticos. La razón para el uso de probabilidades en nuestros modelos es que los objetos que influyen el rendimiento de sistemas informáticos y de redes son frecuentemente aleatorios: el tamaño de programas o de archivos, el tiempo de llegadas de tareas al sistema, el tráfico sobre la Internet, la duración de llamadas telefónicas, etc. Por eso, necesitamos hacer un repaso sobre herramientas de la teoria de probabilidad, como las funciones generadoras de probabilidad, las transformadas de Laplace-Stieltjes y las cadenas de Markov.

Escribi estas notas cuando era Profesor invitado en el C.E.S.I.M.O. en la facultad de Ingenieria de la Universidad de Los Andes, Mérida, Venezuela, en el ao 2001. Quisiera expresar mi gratidud a la Prof. T. Jiménez que me ayudo mucho preparando estas notas, a la Prof. M. Ablan, Prof. J. Dávila y el Prof. G. Tonella que me invitaron al C.E.S.I.M.O. y me ayudaron en mi trabajo, así que a la Prof. N. Morenos Salas y a F. Zerba.

Quisiera expresar mi gratitud también a Prof. U. Yechiali que era mi Profesora del curso de teoria de colas de espera. Durantes mi estudios en el Technion en Haifa, viaje de Haifa a Tel-Aviv dos veses por semana mas que 100Km para escuchar sus cursos interesantes. Parte de la materia aqui y de los ejercicios vienen de su curso, de cuillo tengo todavia mi cuaderno.

Capítulo 2

Probabilidad: repaso

Vamos a recordar algunas bases de la teoria de probabilidad. Vamos a ver algunas distribuciones y el cálculo de probabilidades condicionales. Pueden leer más en http://csic1.csic.edu.uy/EMC/licest/probabilidad1/ sobre los bases de probabilidad.

2.1 Probabilidad y espacio de probabilidad

La probabilidad es una función que tiene por dominio una familia de eventos cuya occurencia es posiblemente incierta. A cada evento, la probabilidad da un número entre 0 y 1, tanto más grande cuanto mayor sea la confianza que ocurra.

Para definir una probabilidad sobre un espacio Ω (que represente todos los eventos elementales o los estados posibles del sistema), tenemos entonces que definir primero el dominio de la probabilidad que es una familia $\mathcal{A}$ de partes de Ω.

$\mathcal{A}$ tiene que ser una *álgebra*, que significa que

1. A es no vacía,
2. Si $A \in \mathcal{A}$ entonces $A^c \in \mathcal{A}$ (A^c es el complemento de A, es la parte de Ω que tiene todos los elementos que no estan en A).
3. Si $A_i \in \mathcal{A}$, $i = 1, ..., n$, entonces $\cup_{i=1}^{n} A_i \in \mathcal{A}$.

Ejemplo 2.1.1 *Arrojamos un dado cuyas caras están numeradas de 1 a 6. Sea Ω todos los resultados posibles. Las familias siguientes son álgebras:*

- $\mathcal{A}_1 =$ *todos los subconjuntos de* Ω
- $\mathcal{A}_2 = \{\Omega, \emptyset\}$,
- $\mathcal{A}_3 = \{\Omega, \{1, 2, 3\}, \{4, 5, 6\}, \emptyset\}$.

Para Ω que tiene un número infinito de elementos definimos una σ-álgebra como una álgebra $\mathcal{A}$ que tiene la propiedad que si $A_i \in \mathcal{A}$, $i = 1, 2, ...,$, entonces $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Si $\mathcal{A}$ es un σ-álgebra, entonces $(\Omega, \mathcal{A})$ se llama un espacio probabilizable.

Una función P sobre un espacio probabilizable $\{\Omega, \mathcal{A}\}$ se llama una probabilidad si

1. $P(\Omega)$=1,

2. Si $A_i \in \mathcal{A}$, $i = 1, 2, ...$, es una sucesión de eventos disjuntos, entonces

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

Esta definición implica que

- $P(\emptyset) = 0$,
- $P(A^c) = 1 - P(A)$,
- Si $A \subset B$ entonces $P(A) \leq P(B)$.
- La segunda propiedad se cumple también con un numero finito de elementos A_i: Si $A_i \in \mathcal{A}$, $i = 1, 2, ..., N$, es una sucesión de eventos disjuntos, entonces

$$P(\cup_{i=1}^{N} A_i) = \sum_{i=1}^{N} P(A_i).$$

$(\Omega, \mathcal{A}, P)$ se llama un *espacio de probabilidad.*

2.1.1 Probabilidad Condicional

En nuestro ejemplo del dado, podemos definir una probabidad simétrica a partir de $P(\{i\}) = 1/6$. Tenemos entonces $P(\{1, 2, 3\}) = P(\{4, 5, 6\}) = 1/2$ gracias a la última propiedad.

En este ejemplo, todos los resultados tienen la misma factibilidad.

Ahora, suponemos que sabemos que el resultado es más grande que 3. La factibilidad de obtener $i = 4, 5, 6$ deviene más grande. Llamamos este nueva probabilidad de un evento A una *probabilidad condicional* de A. Aqui, condicionamos esta probabilidad sobre el evento $B = \{4, 5, 6\}$, y la escribimos $P(A|B)$.

Tenemos la relación:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

En nuestro ejemplo logramos para $i = 4, 5, 6$:

$$P(\{i\}|B) = \frac{1/6}{1/2} = \frac{1}{3},$$

y para $i = 1, 2, 3$, $P(\{i\}|B) = 0$.

Suponemos que sabemos que solo un evento de $B_1, B_2, ..., B_k$ puede ocurrer. Entonces $\Omega = \cup_k B_k$, y $\emptyset = B_i \cap B_j$, $i \neq j$. Entonces,

$$P(A) = P(A \cap \Omega) = P(A \cap (\cup_k B_k)) = P(\cup_k (A \cap B_k)) = \sum_k P(A \cap B_k).$$

Entonces:

$$P(A) = \sum_k P(A|B_k) P(B_k).$$

Ahora,

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A|B_j)P(B_j}{P(A)} = \frac{P(A|B_j)P(B_j}{\sum_k P(A|B_k)P(B_k)}.$$

Aqui $P(B_j|A)$ esta definido usando $\{P(A|B_k)\}$. Esta ecuación se llama la Ley de inverción de Bayes.

2.2 Variables aleatorias

Llamamos σ-álgebra de Borel de $\mathbb{R}$ a la mínima σ-álgebra que contiene los intervalos de la forma $[a, b]$. Lo escribimos $\mathcal{B}$.

Definición 2.2.1 *Dado un espacio probabilizable $(\Omega, \mathcal{A})$, $X : \Omega \to \mathbb{R}$ se llama una variable aleatoria real si la preimagen de todo conjunto B de $\mathcal{B}$ es un evento en $\mathcal{A}$.*

Definición 2.2.2 *Dada una variable aleatoria (VA) real X, la función $P_X : \mathbb{R} \to \mathbb{R}$ definida por medio de $P_X(B) = P(X^{-1}(B)) = P(\{\omega : X(\omega) \in B)\}$ se llama la distribución de probabilidad de la variable aleatoria X.*

Teorema 2.2.1 *La terna $(\mathbb{R}, \mathcal{B}, P_X)$ es un nuevo espacio de probabilidad.*

2.3 Distribuciones de probabilidad

Definición 2.3.1 $F_X : \mathbb{R} \to \mathbb{R}^+$ *definida por*

$$F_X(x) = P_X(\{\omega : X(\omega) \leq x\})$$

se llama la función de distribución de probabilidad de la variable aleatoria real X.

Una variable aleatoria se dice absolutamente continua cuando su función de distribución tiene derivada seccionalmente continua, y, la función de distribución puede escribirse como integral de su derivada. A este derivada se le llama *función de densidad* de la distribución de probabilidad. Si la densidad de la distribución de probabilidad es f_X, entonces

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \qquad \text{y} \qquad P(a < X \leq b) = \int_a^b f_X(t)dt.$$

Ejemplo 2.3.1 *X es una VA exponencial con parámetro μ si*

$$F_X(a) = 1 - \exp(-\mu a).$$

Para obtener la densidad de su distribución de probabilidad derivamos:

$$f_X(t) = \mu \exp(-\mu t).$$

Ejemplo 2.3.2 *Sea X una VA exponencial con parámetro μ. Calcular $P(X > a + b|X > b)$, donde $a, b \geq 0$.*

Solución:

$$P(X > a+b, X > b) = P(X > a+b) = e^{-\mu(a+b)}.$$

Entonces,

$$P(X > a+b|X > b) = \frac{e^{-\mu(a+b)}}{e^{-\mu b}} = e^{-\mu a}.$$

Vemos que $P(X > a+b|X > b) = P(X > a)$. Una VA con esta propiedad se llama VA sin memoria, o con rata de falla constante (constant failure rate). Una VA donde $P(X > a+b|X > b) \leq P(X > a)$ tiene una rata de falla creciente (increasing failure rate).

2.4 Distribuciones conjuntas

Dadas dos variables aleatorias X, Y, definimos la distribución conjunta $P_{X,Y}$:

$$P_{X,Y}((a,b],(c,d]) = P(\{a < X \leq b\} \cap \{c < Y \leq d\}) = P(a < X \leq b, c < Y \leq d)$$

Esta permite definir la distribución conjunta $P_{X,Y}$ de todos los subconjuntos de $\mathcal{B}^{(2)}$, que es la σ-álgebra mínima que contiene a los rectángulos.

Definición 2.4.1 *$F_{X,Y} : \mathbb{R}^2 \to \mathbb{R}^+$ definida por*

$$F_{X,Y}(x,y) = P_{X,Y}(X \leq x, Y \leq Y)$$

se llama la funcitón de distribución de probabilidad de la pareja de variables aleatorias X, Y.

Tenemos las propiedades:

- $F_{X,Y}$ es no decreciente en cada uno de su argumentos,
- $F_{X,Y}(-\infty, y) = F_{X,Y}(x, -\infty) = 0$ para todo x, y.
- $F_{X,Y}(\infty, y) = F_Y(y)$, y $F_{X,Y}(x, \infty) = F_X(x)$. Llamamos F_Y y F_X las distribuciones marginales de (X, Y).

Definición 2.4.2 *La pareja X, Y es absolutamente continua cuando su función de distribución $F_{X,Y}$ puede representarse como la integral*

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} dx' \int_{-\infty}^{y} f_{X,Y}(x',y')dy'.$$

$f_{X,Y}$ se llama la función de densidad de la distribución conjunta.

Tenemos

$$F_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \qquad F_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx.$$

2.4.1 Variables aleatorias independientes

Los eventos $\{ A_i,\ i \in I\}$ son independientes si $P(\cap_{i\in N} A_i) = \prod_{i\in N} P(A_i)$, para todo $N \subset I$, donde N es finito (I puede ser infinito).

La familia de variables aleatorias $\mathcal{X} = \{X_i : i \in I\}$ es independiente cuando para cualesquiera conjuntos $B_i \in \mathcal{B}$, la familia de eventos $\{\{X_i \in B_i\}, i \in I\}$ es independiente.

Ejemplo 2.4.1 *Sea X una VA exponencial con parámetro μ, y Y una VA independienta de X con distribución exponencial con parámetro λ. Sea $\theta = \min(X, Y)$. Cual es la distribución de θ?*

Solución:

$$P(\theta > a) = P(X > a, Y > a) = \exp(-\mu a)\exp(-\lambda a) = \exp(-(\mu + \lambda)a).$$

Entonces θ tiene una distribución exponencial con parámetro $\mu + \lambda$.

2.5 Esperanza y esperanza condicional

2.5.1 Esperanza de variables aleatorias

La esperanza de una VA X con recorrido discreto $\{x_1, x_2, ...\}$ esta definida como

$$E[X] = \sum_j x_j P(X = x_j).$$

Como una función g de una VA es también una VA, podemos ver que

$$E[g(X)] = \sum_j g(x_j) P(X = x_j).$$

Para una pareja X, Y de VAs, una función $g(X, Y)$ es de nuevo una VA, podemos ver que

$$E[g(X, Y)] = \sum_{i,j} g(x_i, y_j) P(X = x_j, Y = y_j).$$

Si X tiene una distribución absolutamente continua con densidad f_X, entonces

$$E[X] = \int_{-\infty}^{\infty} t f_X(t) dt$$

y también

$$E[g(X)] = \int_{-\infty}^{\infty} g(t) f_X(t) dt$$

(si las integrales son absolutamente convergentes).

Para una pareja X, Y de VAs que tienen una dunción de densidad,

$$E[g(X, Y)] = \int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} g(t, y) f_X(t, s) ds.$$

La esperanza es un operador lineal:

$$E[aX + bY] = aE[X] + bE[Y],$$

donde X y Y son VAs, y a y b son constantes.

2.5.2 Esperanza condicional

Ya vimos la probabilidad condicional. Podemos definir ahora la distribución condicional como

$$P_{X|Y}(A|B) = \frac{P_{X,Y}(X \in A, Y \in B)}{P_Y(Y \in B)}$$

si el denominador noe es zero. La función de distribución condicional está definida por:

$$F_{X|Y}(a|b) = \frac{F_{X,Y}(a,b)}{F_Y(b)}.$$

Para (X,Y) con recorridos discretos definimos

$$E[g(X)|Y=y] = \sum_{i,j} g(x_i) P_{X|Y}(X = x_j|Y = y).$$

Condicionar sobre una VA Cuando escribimos $P(X \in A|Y)$, la condición misma es una VA. Entonces, $P(X \in A|Y)$ es una VA. Si Y tiene el recorrido $y_1, ...y_k$ entonces

$$P(X \in A|Y) = P(X \in A|Y = y_k) \text{ con prob. } P_Y(y_k).$$

Podemos entonces calcular la esperanza de $P[A|Y]$:

$$E[P(A|Y)] = \sum_k P(X \in A|Y = y_k) P_Y(y_k) = \sum_k P(X \in A, y_k) = P(X \in A).$$

Cuando escribimos $E[X|Y]$, la condición misma es una VA. Entonces, $E[X|Y]$ es una VA. Si Y tiene el recorrido $y_1, ...y_k$ entonces

$$E[X|Y] = E[X|Y_k] \text{ con prob. } P_Y(y_k).$$

Podemos entonces calcular la esperanza de $E[X|Y]$:

$$\begin{aligned} E(E[X|Y]) &= \sum_k E[X|y_k] P_Y(y_k) = \sum_{k,j} x_j P_{X|Y}(x_j|y_k) P_Y(y_k) = \sum_{k,j} x_j P_{X,Y}(x_j, y_k) \\ &= \sum_j x_j P_{X,Y}(x_j) = E[X]. \end{aligned}$$

Este expreción sirve también cuamdo el recorrido de la VA que aparaesse en la condición no es discreto. Si (X,Y) tienen una densidad, entonces existe una densidad condicional

$$f_{X|Y} = \frac{f_{X,Y}}{f_Y}.$$

Entonces

$$\begin{aligned} E[P(X \in A|Y)] &= E[\int_A f_{X|Y}(x)dx] = \int_{-\infty}^{\infty} f_Y(y)dy \int_A f_{X|Y}(x|y)dx = \int_{-\infty}^{\infty} dy \int_A f_{X,Y}(x,y)dx \\ &= \int_A \int_{-\infty}^{\infty} dy f_{X,Y}(x,y)dx = \int_A f_X(x)dx = P(X \in A) \end{aligned}$$

Ejemplo 2.5.1 La distribución hyper-exponencial:
Sea X una VA con distribución exponencial con un parámetro Y, donde Y es una VA también; Y tiene valores en $\{\mu_1, ..., \mu_k\}$, y definimos $p_i = P(Y = \mu_i)$. ? Cual es la distribución y la esperanza de X ?

Solución:

$$P(X > a) = E[P(X > a|Y)] = \sum_{i=1}^{k} p_i e^{-\mu_i a},$$

y además, como $E(X|Y = \mu_i) = \mu_i^{-1}$,

$$E[X] = E(E(X|Y)) = \sum_{i=1}^{k} \mu_i^{-1}.$$

2.5.3 Esperanzas y variables aleatorias independientes

Si X y Y son independientes, y tienen esperanzas finitas, entonces

$$E[XY] = E[X]E[Y].$$

Ademas, para cada función f y g tal que $E[f(X)]$ y $E[g(Y)]$ son finitos,

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)].$$

Ejemplo 2.5.2 *Sea X una VA exponencial con parámetro μ, y Y una VA exponencial con parámetro λ. Sea $\theta = \min(X, Y)$. Ya vimos que θ tiene una distribución exponencial con parámetro $\mu + \lambda$.*

1. *Calcular $P(X > Y)$.*
2. *Calcular $P(\theta > a, X > Y)$ y calcular $P(\theta > a|X > Y)$.*

Solución:
1.

$$P(X > Y) = E[P(X > Y|Y)] = E[\exp(-\mu Y)] = \int_0^\infty e^{-\mu y} \lambda e^{-\lambda y} \mathrm{d}y = \frac{\lambda}{\lambda + \mu}.$$

2.

$$\begin{aligned} P(\theta > a, X > Y) &= P(X > Y, Y > a) = \int_a^\infty f_Y(y) \int_y^\infty f_X(x) \mathrm{d}x \mathrm{d}y \\ &= \int_a^\infty \lambda e^{-\lambda y} e^{-\mu y} \mathrm{d}y = \frac{\lambda}{\lambda + \mu} e^{-(\lambda+\mu)a} = P(X > Y) P(\theta > a). \end{aligned}$$

Entonces,

$$P(\theta > a|X > Y) = e^{-(\lambda+\mu)a} = P(\theta > a)$$

y $\theta = \min(X, Y)$ y $1\{X > Y\}$ son independientes!

2.5.4 Momentos, varianza y covarianza

$E[X^n]$ se llama el momento de orden n de la variable X. La varianza de X esta definida por

$$\mathrm{var}[X] := E([X - E(X)]^2).$$

Claro que

$$\mathrm{var}[X] = E([X]^2) - (E[X])^2.$$

Tenemos siempre que $\text{var}[X] \geq 0$, y cuando X es una constante, $\text{var}[X] = 0$. $\text{var}[X] = 0$ se usa como una medida de variabilidad de una VA.

Más generalmente tenemos para toda función f convexa de una VA X que

$$E[f(X)] \geq f(E[X]).$$

Esta relación se llama la desigualdad de Jensen.

Sean $Y = \sum_{i=1}^{n} X_n$ donde X_n son independientes con la misma distribución. Entonces

$$E[Y] = nE[X],$$

y también

$$\text{var}[Y] = E(\sum_{i=1}^{n} X_n)^2 - (E[\sum_{i=1}^{n} X_n])^2 = nE[X^2] + n(n-1)(E[X])^2 - n^2(E[X])^2 = n\text{var}[X].$$

La covarianza de dos VAs X y Y esta definida como $\text{Cov}[XY] = E[XY] - E[X]E[Y]$. Si X y Y son independientes entonces $\text{Cov}[XY] = 0$. Llamamos a $E[XY]$ la correlación entre X y Y.

2.5.5 Ejemplos de variables aleatorias

VA Bernoulli
Es la VA X que vale 1 si un evento A ocurre, y 0 si non. Si $P(A) = p$ y $P(A^c) = 1 - p = q$, entonces $E[X^k] = p$ para todos $k = 1, 2, ...$ y $\text{var}[X] = p - p^2 = pq$.

Distribución binomial Bin(n, p)
Consideramos n VAs de Bernoulli independientes. Estas representan n experimentos independientes sucesivos. Llamamos X_i el resultado del experimento i, y sea $B_n = \sum_{i=1}^{n} X_i$. Entonces

$$P(B_n = k) = \binom{n}{k} p^k q^{n-k}.$$

Ahora, $E[B_n] = np$, $E[B_n^2] = npq + n^2p^2$ y $\text{var}[B_n] = npq$.

Distribución geométrica Geo(p)
Repetimos de nuevo N experimentos independientes como antes, donde en cada experimento, A ocurre con probabilidad p. Sea N el número (aleatorio) de ensayos que deben realizarse antes de obtener por primera vez el resultado A. Entonces

$$P(N = k) = q^{k-1}p, \qquad k = 1, 2, 3, ...$$

La distribución de N se llama geométrica con parámetro p. Obtenemos:

$$\begin{aligned} E[N] &= \sum_{i=k}^{\infty} kq^{k-1}p = p\sum_{i=k}^{\infty} \frac{\mathrm{d}q^k}{\mathrm{d}q} = p\frac{\mathrm{d}\sum_{i=k}^{\infty} q^k}{\mathrm{d}q} \\ &= p\frac{\mathrm{d}(1-q)^{-1}}{\mathrm{d}q} = p\frac{1}{(1-q)^2} = \frac{1}{p}. \end{aligned}$$

Distribución multinomial
Esta distribución generaliza la distribución binomial. Ahora tenemos n experimentos independientes. En cada experimento hay k resultados (eventos) posibles, $A_1, ..., A_k$, con probabilidades $p_1, p_2, ..., p_k$. Sea Y_i el número de veces que obtenemos A_i durante todos los n experimentos. Entonces, por $n_1 + ... + n_k = n$, tenemos

$$P(Y_1 = n_1, ..., Y_k = n_k) = \frac{n!}{n_1! n_2! ... n_k!} p_1^{n_1} p_2^{n_2} ... p_k^{n_k}.$$

Distribución de Poisson
Una VA X tiene la distribución de Poisson can parámetro λ si

$$P(X = n) = \frac{(\lambda)^n}{n!} e^{-\lambda}, \qquad n = 0, 1, 2, ...$$

Entonces,

$$\begin{aligned} E[X] &= \sum_{n=0}^{\infty} n \frac{(\lambda)^n}{n!} e^{-\lambda} = \sum_{n=1}^{\infty} n \frac{(\lambda)^n}{n!} e^{-\lambda} = \lambda \sum_{n=1}^{\infty} \frac{(\lambda)^{n-1}}{(n-1)!} e^{-\lambda} \\ &= \lambda \sum_{k=0}^{\infty} \frac{(\lambda)^k}{k!} e^{-\lambda} = \lambda. \end{aligned}$$

2.6 Proceso de Poisson

En un proceso de Poisson con parámetro λ, el número de llegadas en un intervalo de longitud t es una VA de Poisson con parámetro λt. La probabilidad de que lleguen exactamente n clientes durante un intervalo de longitud t esta dada por la ley de Poisson:

$$q_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

donde λ es la velocidad media de llegadas.

2.6.1 La relación con la distribución exponencial

Montramos que para este proceso de llegadas, el tiempo entre llegadas tiene la distribución exponencial. La probabilidad $a(t)\delta t$ de que un tiempo entre llegadas se encuentre entre t y $t + \delta t$ es la probabilidad de que no haya llegadas durante un tiempo t, multiplicada por la probabilidad de que exista una sola llegada durante el intervalo de longitud δt. Entonces

$$a(t)\Delta(t) = q_0(t) q_1(\Delta t) = e^{-\lambda t} \Delta t e^{-\lambda \Delta t}$$

Ahora, si $\Delta t \to 0$, obtenemos

$$a(t) = \lambda e^{-\lambda t}$$

que es la densidad de la distribución exponencial.

2.6.2 Particionamente y suma de Procesos de Poisson

Suponemos que tenemos k colas y un proceso de llegadas de Poisson. Cuando el cliente i llega, va a la cola $C_i \in \{1, 2, ..., k\}$. C_i son independientes y $P(C_i = j) = p_j, j = 1, ..., k$. ¿Cual es el proceso de llegada a cada cola?

Sea $N(t)$ el número de llegadas durante un intervalo de longitud t, y sea $N_j(t)$ el número de llegadas a la cola j durante este intervalo. Entonces

$$P(N_1(t) = n_1, ..., N_k(t) = n_k | N(t) = n) = \frac{n!}{n_1! n_2! ... n_k!} p_1^{n_1} p_2^{n_2} ... p_k^{n_k}$$

donde $n_1 + ... + n_k = n$. Entonces,

$$\begin{aligned} P(N_1(t) = n_1, ..., N_k(t) = n_k) &= P(N_1(t) = n_1, ..., N_k(t) = n_k | N(t) = n) P(N(t) = n) \\ &= \frac{n!}{n_1! n_2! ... n_k!} p_1^{n_1} p_2^{n_2} ... p_k^{n_k} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \\ &= \prod_{i=1}^{n} e^{-\lambda p_i t} \frac{(\lambda p_i t)^n}{n!} = \prod_{i=1}^{n} P_{Poisson(\lambda p_i)}(N_i = n_i). \end{aligned}$$

Entonces, obtenemos un particionamente a n procesos de Poisson independientes!

El contrario es también cierto: Si $N_i(t)$ son procesos de Poisson independientes con parámetros λ_i, $i = 1, ..., k$, entonces $N(t) = \sum_{i=1}^{k} N_i(t)$ es también un proceso de Poisson con parámetro $\lambda = \sum_{i=1}^{k} \lambda_i$. Demonsramos este propiedad más tarde con detalles. Pero, una explicación de eso, es que como $N_i(t)$, $i = 1, ..., k$, son procesos de Poisson, entonces para cada tiempo s, el tiempo $X_i(s)$ hasta la proxima llegada en el processo i tiene la distribución exponencial con parámetro λ_i, como lo vimos en la subsección 2.6.1. La primera llegada en el proceso N hasta el tiempo s es ocurre despues el tiempo $\theta = \min_{i=1,...,k} X_i(s)$. θ tiene la distribución exponencial con parámetro λ, como lo vimos ya en Ejemplos 2.4.1 y 2.5.2. Usando de nuevo la relación entre proceso de Poisson y la distribucion exponencial entre llegadas de la subsección 2.6.1, vemos que $N(t)$ es un proceso de Poisson con parámetro λ.

Finalmente, otra razón para trabajar con Procesos de Poisson es que cuando una cola tiene un procesa de Poisson como llegada, el proceso de salida es también Poisson si los tiempos de servicios tienen la distribución exponencial; este resultado se generaliza también cuando hay más que un servidor (a sistemas que se llaman $M/M/C/\infty$). Esta propriedad nos permite analizar redes de colas que tienen una topologia de arboles. Pero, se pueden analizar también redes con topologia general cuando los procesos de llegadas son Poisson (redes de Jackson [10, 11]).

2.7 Ejercicios

1. Calcular las segundos momentos y las varianzas de VAs Uni(0,1), Bin(n,p), Geo(p), Exp(λ), Poisson(λ).

2. Sea Y una VA con distribución $Geo(p)$. ¿Cual es la probabilidad $P(Y > i)$? ¿Cual es la probabilidad que Y sea un número par?

3. Sea X una VA con distibución exponencial con parámetro λ. Sea $Y = j$ si $j \leq X < j + 1$, donde $j = 0, 1, 2, ...$ ¿Cual es la distribución de Y?

4. Sean X_i VAs exponenciales independientes con parámetros λ_i, $i = 1, 2, ..., k$. Mostra que

$$P(X_j \leq X_i, \forall i = 1, ..., k) = \frac{\lambda_j}{\sum_{i=1}^{k} \lambda_i}.$$

Capítulo 3

Funciones Generadoras, Transformada de Laplace-Stieljes

Veremos aqui las Funciones Generadoras de Probabilidad (FGP) y la Transformada de Laplace Stieltjes (TLS) que nos facilitan el cálculo de distribuciones y de momentos de las variables aleatorias. La primera se usa para variables aleatorias con valores discretos no negativos, y la segunda para valores reales no negativos.

3.1 Funciones Generadoras de Probabilidad de Variables Aleatorias

Definición 3.1.1 *Si* $\{a_0, a_1, ...\}$ *es una secuencia de números reales, entonces* $A(z) := \sum_{i=0}^{\infty} a_i z^i$, *se llama la* **función generadora** *de esta secuencia.*

$A(z)$ tiene las propiedades siguientes:

1. **Convergencia.** Hay un número $R \geq 0$, llamado el **radio de convergencia**, tal que la suma converge absolutamente si $|z| < R$, y diverge si $|z| > R$. La suma converge uniformemente sobre conjuntos con la forma $\{z : |z| \leq R'\}$ donde $R' < R$.

2. **Derivación e Integración.** De la convergencia uniforme tenemos que para derivar (o integrar) G, podemos cambiar entre la suma y la derivada (o la integración).

3. **El Teorema de Abel.** Si $a_i \geq 0$ para todo i, entonces $\lim_{z \uparrow 1} A(z) = \sum_{i=0}^{\infty} a_i$, aún si la suma no converge.

4. **Obtener** a_i**.** Si $R > 0$ entonces
$$a_i = \frac{A^{(i)}(0)}{i!}$$
donde $A^{(i)}$ es la i-ésima derivada de A.

5. **Convolución.** Sea $\{b_0, b_1, ...\}$ otra secuencia de números reales y definimos
$$c_n = \sum_{i=0}^{n} a_i b_{n-i}, \qquad n = 0, 1, 2, ...$$

La secuencia c se llama la convolución de $\{a_i\}$ y $\{b_i\}$. Si a_i y b_i son no-negativos, entonces podemos cambiar el orden de sumatorias y obtener:

$$\sum_{n=0}^{\infty} c_n z^n = \sum_{n=0}^{\infty}\sum_{i=0}^{n} a_i b_{n-i} z^n = \sum_{i=0}^{\infty}\sum_{n=i}^{\infty} a_i b_{n-i} z^n = \sum_{i=0}^{\infty} a_i z^i \sum_{k=0}^{\infty} b_k z^k.$$

Vemos que la FGP de la convolución de $\{a_i\}$ y $\{b_i\}$ nos da el producto de las FGPs de las dos secuencias.

Sea X una variable aleatoria discreta con valores en $0, 1, 2, ...$, y $a_i = P(X = i), i = 0, 1, 2, ...$ Entonces $G_X(z) := \sum_{i=0}^{\infty} a_i z^i$ se llama la **función generadora de probabilidad de** X. Podemos escribirla como

$$G_X(z) = E\left(z^X\right) = \sum_{i=0}^{\infty} P(X = i) z^i.$$

Tenemos en convergencia para todo z t.q. $|z| \leq 1$, entonces $R \geq 1$.

3.1.1 Propiedades y ejemplos de FGPs

Sea $W := X + Y$, donde $X, Y \in \{0, 1, 2, 3, ...\}$ son independientes.

$$P(W = n) = \sum_{i=0}^{n} P(X = i) P(Y = n - i), \qquad n = 0, 1, 2, ...$$

Entonces $G_W(z) = G_X(z) G_Y(z)$. Podemos verlo también de

$$G_W(z) = E\left[z^{X+Y}\right] = E\left[z^X z^Y\right] = E\left[z^X\right] E\left[z^Y\right] = G_X(z) G_Y(z). \tag{3.1}$$

Obtener las probabilidades
Usamos $x^0 = 1$ para todo x (aun $0^0 = 1$), y $0^x = 0$ para todo $x \neq 0$.

$$P(X = 0) = 0^0 P(X = 0) = \sum_{i=0}^{\infty} 0^i P(X = i) = G_X(0),$$

$$P(X = k) = \frac{G_X^{(k)}(0)}{k!}, \quad k \geq 1.$$

Cálculo de momentos: $G(1) = 1$, $E[X] = G'(1)$, porque

$$G'(1) = \left.\frac{\mathrm{d}E\left[z^X\right]}{\mathrm{d}z}\right|_{z=1} = E\left[\frac{\mathrm{d}z^X}{\mathrm{d}z}\right]\Big|_{z=1} = E\left[X z^{X-1}\right]\Big|_{z=1} = E[X].$$

Otros momentos:

$$E[X(X-1)...(X-k+1)] = G^{(k)}(1), \quad k \geq 1.$$

Cálculo de la varianza:

$$Var[X] = G''(1) + G'(1) - (G'(1))^2.$$

Usamos la convención que si el radio de convergencia de G es 1, entonces $G^{(k)}(1) := \lim_{z\uparrow 1} G^{(k)}(z)$.

Ejemplo 3.1.1 *Si $X = c$ casi seguro, entonces $G_X(z) = a^c$.*

Ejemplo 3.1.2 *Si X es Bernoulli: $X = 1$ con probabilidad p y $X = 0$ con probabilidad $1-p$, entonces $G_X(z) = 1 - p + pz$.*
Ahora, vemos que

$$P(X = 0) = G_X(0) = 1 - p, \qquad P(X = 1) = G'_X(0) = p,$$

y además,

$$E[X] = G'(1) = p, \qquad G^{(n)}(z) = 0, \ \forall n > 1, \forall z,$$

entonces $E[X(X-1)] = 0$ y $E[X^2] = E[X] = p$.

Ejemplo 3.1.3 *Si X es Binomial $B(n,p)$: $X_i = 1$ con probabilidad p y $X_i = 0$ con probabilidad $1-p$, $X = \sum_{i=1}^n X_i$, X_i son i.i.d. Entonces $G_X(z) = (1 - p + pz)^n$ (usando ecuación (3.1)).*

Ejemplo 3.1.4 *Si X tiene la distribución de Poisson con parámetro λ:*

$$P(X = i) = \frac{\lambda^i}{i!}\exp(-\lambda),$$

entonces

$$G_X(z) = \sum_{i=0}^{\infty} P(X = i)z^i = \exp(-\lambda)\sum_{i=0}^{\infty}\frac{(\lambda z)^i}{i!} = \exp(-\lambda)\exp(\lambda z) = \exp(-\lambda(1-z)).$$

Ejemplo 3.1.5 *Si X tiene la distribución geométrica con parámetro p: $P(X = i) = (1-p)^{i-1}p$, $i = 1, 2, 3, ...$, entonces*

$$G_X(z) = \sum_{i=1}^{\infty}(1-p)^{i-1}pz^i = pz\frac{1}{1-(1-p)z}.$$

3.1.2 Suma aleatoria de VAs

Sea $\{X_i\}$ una secuencia de VA i.i.d. y

$$S = \begin{cases} 0, & \text{si } N = 0 \\ \sum_{i=1}^N X_i & \text{si } N \geq 1. \end{cases}$$

Entonces, $G_S(z) = G_N(G_X(z))$, $|z| \leq 1$:

$$\begin{aligned} G_S(z) &= E\left[z^S\right] = E\left(E\left[z^S \middle| N\right]\right) = E\left(E\left[z^{X_1+\ldots+X_N} \middle| N\right]\right) \\ &= E\left((E[z^X])^N\right) = E\left(\left(G_X(z)\right)^N\right) = G_N(G_X(z)) \end{aligned}$$

3.1.3 Suma de Procesos de Poisson

Sean $N_i(t)$ procesos de Poisson independientes con parámetros λ_i, $i = 1, ..., k$, entonces $N(t) = \sum_{i=1}^{k} N_i(t)$ es tambien un procesos de Poisson con parámetro $\lambda = \sum_{i=1}^{k} \lambda_i$. Para mostrarlo, recordamos que la función generadora de probabilidad de una variable aleatoria de Poisson con parámetro $\lambda_i t$ es dada por

$$G_{N_i(t)}(z) = e^{-\lambda_i t(1-z)}.$$

Entonces,

$$\begin{aligned} G_{N(t)}(z) &= E\left[z^{N(t)}\right] = E\left[z^{\sum_{i=1}^{k} N_i(t)}\right] = \prod_{i=1}^{k} E\left[z^{N_i(t)}\right] \\ &= \prod_{i=1}^{k} e^{-\lambda_i t(1-z)} = e^{-\lambda t(1-z)}. \end{aligned}$$

3.2 Transformada de Laplace Stieltjes de Variables Aleatorias

Sea X una variable aleatoria no-negativa. La **Transformada de Laplace Stieltjes (TLS)** de X (o de su distribución F_X) esta definida como

$$X^*(s) = E\left(e^{-sX}\right) = \int_0^\infty e^{-sx} dF_X(x)$$

donde $F_X(x)$ es la distribución de probabilidad de X.

Si X es un variable contínua con densidad de probabilidad f_X, entonces

$$X^*(s) = E\left(e^{-sX}\right) = \int_0^\infty e^{-sx} f_X(x) dx$$

y $X^*(s)$ se llama la transformada de Laplace de f_X.

Estas transformadas son definidas para todo $s \geq 0$.

3.2.1 Propiedades de la TLS

1. **Momentos de X:** $X^*(0) = 1$, $E[X] = -(X^*)'(0)$, porque

   $$\left.\frac{\mathrm{d}X^*(s)}{\mathrm{d}s}\right|_{s=0} = \left.\frac{\mathrm{d}E[\exp(-sX)]}{\mathrm{d}s}\right|_{s=0} = E\Big[\frac{\mathrm{d}\exp(-sX)}{\mathrm{d}s}\Big]\Big|_{s=0} = E\Big[-X\exp(-sX)\Big|_{s=0} = -E[X].$$

   Otros momentos:

   $$E[X^k] = (-1)^k (X^*)^{(k)}(0), \quad k > 1.$$

2. **El converso:** Si $[X^k] < \infty$ entonces

   $$X^*(s) = \sum_{j=0}^{k} \frac{E[X^j]}{j!}(-s)^j + o(s^k).$$

3. Si $Z = X + Y$, y X y Y son independientes, entonces

   $$Z^*(s) = X^*(s)Y^*(s),$$

   porque

   $$Z^*(s) = E\left(-e^{sZ}\right) = E\left(-e^{s(X+Y)}\right) = E\left(-e^{sX}\right)E\left(-e^{sY}\right) = X^*(s)Y^*(s).$$

3.2.2 Ejemplos

Ejemplo 3.2.1 *X tiene la distribución uniforme sobre (a,b), $a \geq 0$, si su función de densidad de distribución es*

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in= [a,b], \\ 0, & x \notin= [a,b]. \end{cases}$$

Entonces,

$$X^*(s) = \begin{cases} \dfrac{e^{-as} - e^{-bs}}{s(b-a)}, & s \neq 0, \\ 1, & s = 0. \end{cases}$$

Ejemplo 3.2.2 *Si X tiene la distribución exponencial con parámetro λ: $P(X > r) = \exp(-\lambda r)$, entonces*

$$X^*(s) = \int_0^\infty e^{-sx}\lambda e^{-\lambda x}dx = \lambda \int_0^\infty e^{-(\lambda+s)x}dx = \frac{\lambda}{\lambda+s}.$$

Cálculo de momentos:

$$\frac{\mathrm{d}X^*(s)}{\mathrm{d}s} = \frac{-\lambda}{(\lambda+s)^2} \qquad \textit{entonces } E[X] = -X^*(0)' = \frac{1}{\lambda};$$

$$\frac{\mathrm{d}^2X^*(s)}{\mathrm{d}s^2} = \frac{2\lambda}{(\lambda+s)^3} \qquad \textit{entonces } E[X^2] = X^*(0)^{(2)} = \frac{2}{\lambda^2},$$

y $Var(X) = 1/\lambda^2$.

3.2.3 La distribución de Erlang y de Gamma

Una distribución se llama de Gamma con parámetros λ y n para n positivo y real (no necesariamente entero) si la densidad de probabilidad es

$$f(t) = e^{-\lambda t}\frac{\lambda^n t^{n-1}}{\Gamma(n)}. \tag{3.2}$$

La definición de $\Gamma(\alpha)$ es

$$\Gamma(\alpha) = \int_0^\infty e^{-\lambda t}\lambda^\alpha t^{\alpha-1}dt.$$

Si n es entero entonces la distribución se llama también Erlang(n), y tenemos $\Gamma(n) = (n-1)!$. Para verlo, integramos en partes:

$$dU = e^{-\alpha t}dt, \qquad V = \lambda^\alpha t^{\alpha-1}dt.$$

Entonces,

$$U = -\alpha^{-1}e^{-\alpha t}, \qquad dV = (\alpha-1)\lambda^\alpha t^{\alpha-2}$$

y logramos

$$\Gamma(\alpha) = \int UdV = UV\Big|_0^\infty - \int VdU = (\alpha-1)\Gamma(\alpha-1).$$

Además,

$$\Gamma(1) = \int_0^\infty e^{-\lambda t}\lambda dt = 1.$$

Sean $X_i, i = 1, ..., n$ V.A.s exponenciales con parámetro λ, independientes. Sea $Y_n = \sum_{i=1}^{n} X_i$. Vamos a mostrar que Y_n tiene la distribución E_n. Tenemos

$$Y^*(s) = \left(\frac{\lambda}{\lambda + s}\right)^n, \tag{3.3}$$

y $E[Y_n] = n/\lambda$.

Vemos si obtenemos lo mismo de (3.2):

$$\begin{aligned}
\int_0^\infty f(t)e^{-st}dt &= \int_0^\infty e^{-st}e^{-\lambda t}\frac{\lambda^n t^{n-1}}{(n-1)!}dt \\
&= \lambda^n \int_0^\infty e^{-(\lambda+s)t}\frac{t^{n-1}}{(n-1)!}dt \\
&= \left(\frac{\lambda}{\lambda+s}\right)^n \int_0^\infty e^{-(\lambda+s)t}\frac{(\lambda+s)^n t^{n-1}}{(n-1)!}dt \\
&= \left(\frac{\lambda}{\lambda+s}\right)^n.
\end{aligned}$$

La última integral es igual a 1 porque es la integral de una densidad de probabilidad (de la distribución E_n con parámetro $\lambda + s$).

Manera directa de obtener la densidad:
Sea $N(t)$ el número de llegadas durante $[0, t]$. Tenemos $N(t) \geq n$ si y solamente si $Y_n \leq t$. Como las llegadas siguen un proceso de Poisson,

$$P(N(t) \geq n) = \sum_{j=n}^{\infty} e^{-\lambda t}\frac{(\lambda t)^j}{j!}.$$

Entonces,

$$\begin{aligned}
f_Y(t) &= \frac{dP(Y_n \leq t)}{dt} = \sum_{j=n}^{\infty}\left[\lambda e^{-\lambda t}\frac{\lambda^{j-1}t^{j-1}}{(j-1)!} - \lambda e^{-\lambda t}\frac{(\lambda t)^j}{j!}\right] \\
&= \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}
\end{aligned}$$

Congergencia a una constante
Consideramos una secuencia X_i de VAs todos con la misma esperanza, y donde X_i es Erlang$(i\lambda, i)$ (como es la suma de i VAs exponenciales con párametro $i\lambda$, su esperanza es λ y no depende de i). Vamos a mostrar que X_i converge a una constante (en distribución). Por eso vamos a mostrar que su transformada de Laplace converge a $\exp(-s/\lambda)$:

$$E[\exp(-sX_i)] = \left(\frac{i\lambda}{i\lambda + s}\right)^i = \left(1 - \frac{s}{i\lambda + s}\right)^i,$$

Entonces

$$\lim_{i\to\infty} E[\exp(-sX_i)] = \lim_{i\to\infty}\left(1 - \frac{s}{i\lambda + s}\right)^i = \lim_{i\to\infty}\left(1 - \frac{s}{i\lambda}\right)^i = \exp(-s/\lambda).$$

3.2.4 Suma aleatoria de VAs

Sea $\{X_i\}$ una secuencia de VA i.i.d. y

$$\mathcal{B}_N = \begin{cases} 0, & \text{if } N = 0 \\ \sum_{i=1}^{N} X_i & \text{if } N \geq 1. \end{cases}$$

Entonces, $\mathcal{B}_N^*(s) = G_N(X^*(s))$:

$$\begin{aligned} \mathcal{B}_N^*(s) &= E\left[e^{-s\mathcal{B}_N}\right] = E\left(E\left[e^{-s\mathcal{B}_N}\,\middle|\, N\right]\right) = E\left(E\left[e^{-s(X_1+\ldots+X_N)}\,\middle|\, N\right]\right) \\ &= E\left((E[e^{-sX}])^N\right) = E\left(\left(X^*(s)\right)^N\right) = G_N(X^*(s)) \end{aligned}$$

Ejemplo 3.2.3 *Sea X_i VAs con distribución exponencial con parámetro μ, y sea N una VA con distribución geométrica con parámetro p. Recordamos que*

$$G_N(z) = \frac{pz}{1-(1-p)z}.$$

Entonces:

$$\mathcal{B}_N^*(s) = \frac{pX^*(s)}{1-(1-p)X^*(s)} = \frac{\frac{\lambda p}{\lambda+s}}{1-(1-p)\frac{\lambda}{\lambda+s}} = \frac{\lambda p}{\lambda+s-(1-p)\lambda} = \frac{\lambda p}{s+\lambda p}$$

Vemos que $\mathcal{B}_N$ tiene una distribución exponencial con parámetro λp!

Ejemplo 3.2.4 *(Particionamiento de Procesos de Poisson y otros procesos)*
Suponemos que tenemos un sistema de k colas. Los tiempos entre llegadas al sistema X_i, $i = 1, 2, \ldots$ X_i son independientes y tienen transformada de Laplace-Stieltjes (TLS) $X^*(s)$. Cuando el cliente i llega, va a la cola $C_i \in \{1, 2, \ldots, k\}$. C_i son independientes y $P(C_i = j) = p_j, j = 1, \ldots, k$.

1. *¿Cual es la TLS $S_j^*(s)$ del tiempo entre llegadas a la cola j?*
2. *Si el proceso de llegadas al sistema es Poisson con parámetro λ, cual es el proceso de llegadas a cada cola?*

Solución:
(1) El tiempo entre llegadas a la cola j esta dada por:

$$S_j = \sum_{i=1}^{N(j)} X_i,$$

donde $N(j)$ tiene una distribución geométrica con parámetro p_j. Usando el ejemplo precedente, vemos que

$$S_j^*(s) = \frac{p_j X^*(s)}{1-(1-p_j)X^*(s)}.$$

(2) Cuando el proceso de llegadas al sistema es Poisson con parámetro λ, X tiene la distribución exponencial con el mismo parámetro. Podemos entonces usar de nuevo el ejemplo precedente para ver que S_j tiene una distribución exponencial con parámetro λp_j. Podemos entonces concluir que el proceso de llegadas a la cola j es Poisson con parámetro λp_j. Este es un método alternativo de lo que vimos en la Subsección 2.6.2.

Cálculo de la esperanza de $\mathcal{B}_N$

$$E[\mathcal{B}_N] = -\left.\frac{\mathrm{d}\mathcal{B}_N^*(s)}{\mathrm{d}s}\right|_{s=0} = -\left.\frac{\mathrm{d}G_N(X^*(s))}{\mathrm{d}s}\right|_{s=0} = -\left.\left(\left.\frac{\mathrm{d}G_N(z)}{\mathrm{d}z}\right|_{z=X^*(s)} \frac{\mathrm{d}X^*(s)}{\mathrm{d}s}\right)\right|_{s=0}$$

Cuando $s=0$, $X^*(s)=1$, y

$$\left.\frac{\mathrm{d}G_N(z)}{\mathrm{d}z}\right|_{z=1} = E[N].$$

Además,

$$-\left.\frac{\mathrm{d}X^*(s)}{\mathrm{d}s}\right|_{s=0} = E[X].$$

Entonces $E[\mathcal{B}_N] = E[N]E[X]$.

Cálculo del segundo momento de $\mathcal{B}_N$

$$\begin{aligned} E[\mathcal{B}_N^2] &= \left.\frac{\mathrm{d}^2\mathcal{B}_N^*(s)}{\mathrm{d}s^2}\right|_{s=0} = \left.\frac{\mathrm{d}^2G_N(X^*(s))}{\mathrm{d}s^2}\right|_{s=0} = \frac{\mathrm{d}}{\mathrm{d}s}\left.\left(\left.\frac{\mathrm{d}G_N(z)}{\mathrm{d}z}\right|_{z=X^*(s)} \frac{\mathrm{d}X^*(s)}{\mathrm{d}s}\right)\right|_{s=0} \\ &= \left.\left(\left.\frac{\mathrm{d}^2G_N(z)}{\mathrm{d}z^2}\right|_{z=X^*(s)} \left(\frac{\mathrm{d}X^*(s)}{\mathrm{d}s}\right)^2\right)\right|_{s=0} + \left.\left(\left.\frac{\mathrm{d}G_N(z)}{\mathrm{d}z}\right|_{z=X^*(s)} \frac{\mathrm{d}^2X^*(s)}{\mathrm{d}s^2}\right)\right|_{s=0} \end{aligned}$$

Cuando $s=0$, $X^*(s)=1$, y

$$\left.\frac{\mathrm{d}^2G_N(z)}{\mathrm{d}z^2}\right|_{z=1} = E[N^2] - E[N].$$

Además,

$$\left.\frac{\mathrm{d}^2X^*(s)}{\mathrm{d}s^2}\right|_{s=0} = E[X^2].$$

Entonces,

$$E[\mathcal{B}_N^2] = (E[N^2] - E[N])(E[X])^2 + E[N]E[X^2] = E[N^2](E[X])^2 + E[N]var[X] \tag{3.4}$$

3.3 La cola G/M/1/0

Consideramos ahora llegadas a un servidor sin cola. Una llegada que encuentra el servidor ocupado se pierde. Un ejemplo de tal modelo son los redes telefonicos. Alla, el servicio represente una comunicación. Cuando conversamos, otras llamadas no pueden llegar.

El tiempo de servicio tiene la distribución exponencial con parámetro μ. Sea t_n el tiempo de la n-ésima llegada, y T_n el tiempo entre la n y la $n+1$ llegada: $T_n = t_{n+1} - t_n$.

Suponemos que T_n tienen una distribución general y que $\{T_n\}$ es estacionaria, independiente de los tiempos de servicio. Permitimos que T_n no sean independientes entre ellas. Este sistema se llama la cola G/M/1/0 (veremos más tarde que significan cada símbolo).

¿Cual es la probabilidad de bloqueo de una llegada?.

¿Cual es el rendimiento?

Inmediatamente después de t_n el servidor está ocupado sirviendo un cliente. Este cliente corresponde a la n-ésima llegada, si encontró el servidor desocupado, o una llegada precedente, si el

servidor estaba ocupado inmediatamente antes de t_n. Sea R_n el tiempo que le falta al servidor para terminar el servicio de este cliente. R_n tiene la distribución exponencial con parámetro μ.

La $n+1$ llegada encuentra el servidor ocupado si y solamente si $R_n > T_n$. Entonces,

$$P(bloqueo) = P(R_n > T_n) = E[P(R_n > T_n|T_n)] = E[\exp(-\mu T_n)] = T^*(\mu).$$

El rendimiento es $\lambda(1 - T^*(\mu))$, donde $\lambda = 1/E[T_n]$.

Ejemplo 3.3.1 *Suponemos que el proceso de llegadas a la cola G/M/1/0 es Poison con parámetro λ. Entonces,*

$$P(bloqueo) = T^*(\mu) = \frac{\lambda}{\lambda + \mu},$$

y el rendimiento es

$$\lambda(1 - T^*(\mu)) = \frac{\lambda\mu}{\lambda + \mu}.$$

Por exemplo, si $\lambda = \mu$ entonces el rendimiento es $\lambda/2$.

3.4 Número de llegadas en un intervalo

Sea $X(t)$ un proceso estocástico de Poisson con parámetro λ. El número de llegadas $N(Q)$ en el intervalo $[0, Q]$ (donde Q es fijo) es una VA con distribución de Poisson con parámetro λQ. Entonces,

$$G_{N(Q)}(z) = \exp(-\lambda Q(1 - z)).$$

Ahora, cual será la FGP de $N(Q)$ si Q es una VA?

$$\begin{aligned} G_{N(Q)}(z) &= E\left[z^{N(t)}\right] = E\Big[E\Big(\, z^{N(Q)} \,\Big|\, Q\Big)\Big] = E\Big[\exp(-\lambda Q(1 - z))\Big] \\ &= E\Big[\exp(-(\lambda(1 - z))Q)\Big] = Q^*(\lambda(1 - z)). \end{aligned}$$

Cálculo de la esperanza de $N(Q)$

$$E[N(Q)] = \frac{\mathrm{d}Q^*(\lambda(1 - z))}{\mathrm{d}z}\bigg|_{z=1} = \left(\frac{dQ^*(y)}{dy}\bigg|_{y=\lambda(1-z)} \times \frac{\mathrm{d}(\lambda(1 - z))}{\mathrm{d}z}\right)\bigg|_{z=1}.$$

Cuando $z = 1$, $y = \lambda(1 - z) = 0$, y

$$\frac{dQ^*(y)}{dy}\bigg|_{y=0} = -E[Q].$$

Además,

$$\frac{\mathrm{d}(\lambda(1 - z))}{\mathrm{d}z}\bigg|_{z=1} = -\lambda.$$

Entonces, $E[N(Q)] = \lambda E[Q]$.

Cálculo del segundo momento de $N(Q)$

$$\begin{aligned} E[N(Q)^2] - E[N(Q)] &= \left.\frac{\mathrm{d}^2 Q^*(\lambda(1-z))}{\mathrm{d}z^2}\right|_{z=1} = \frac{\mathrm{d}}{\mathrm{d}z}\left(\left.\frac{dQ^*(y)}{dy}\right|_{y=\lambda(1-z)} \times \frac{\mathrm{d}(\lambda(1-z))}{\mathrm{d}z}\right)\Bigg|_{z=1} \\ &= \left(\left.\frac{d^2Q^*(y)}{dy^2}\right|_{y=\lambda(1-z)} \times \left(\frac{\mathrm{d}(\lambda(1-z))}{\mathrm{d}z}\right)^2\right)\Bigg|_{z=1} = E[Q^2]\lambda^2. \end{aligned}$$

Entonces,

$$E[N(Q)^2] = \lambda^2 E[Q^2] + \lambda E[Q]. \tag{3.5}$$

Ejemplo 3.4.1 *Sea Q exponencial con parámetro μ. ¿Cual es la distribución de $N(Q)$?*

Solución:
Recordamos que

$$Q^*(s) = \frac{\mu}{\mu + s}.$$

Entonces,

$$G_{N(Q)}(z) = Q^*(\lambda(1-z)) = \frac{\mu}{\mu + \lambda(1-z)} = \frac{1}{z}\frac{zp}{1-(1-p)z}$$

donde

$$p = \frac{\mu}{\lambda + \mu}.$$

Vemos que $G_{N(Q)}(z)$ es la FGP de $X - 1$ donde X tiene la distribución geometrica con parámetro p. Entonces, $P(N(Q) = n) = (1-p)^n p$.

3.5 Funciones Generadoras de Probabilidad de Vectores Aleatorios

3.5.1 Definición

Sea $\mathbf{X} = (X_1, ..., X_K)$ un vector aleatorio. Definimos $\mathbf{z} = (z_1, ..., z_K)$,

$$\mathbf{G_X}(\mathbf{z}) = E\left(z_1^{X_1} z_2^{X_2} \cdots z_K^{X_K}\right) = \sum_{i_1=0}^{\infty}\sum_{i_2=0}^{\infty} ... \sum_{i_K=0}^{\infty} P(\mathbf{X} = (i_1, ..., i_K)) z_1^{i_1} z_2^{i_2} \cdots z_K^{i_K}.$$

Tenemos la convergencia para todo z t.q. $|\mathbf{z}| \leq 1$.

3.5.2 Propiedades

Sea $\mathbf{W} = \mathbf{Z} + \mathbf{Y}$, donde $\mathbf{Z}$ y $\mathbf{Y}$, son independientes. Entonces

$$\begin{aligned} \mathbf{G_W}(\mathbf{z}) &= E\left[z_1^{X_1+Y_1} z_2^{X_2+Y_2} ... z_K^{X_K+Y_K}\right] = E\left[z_1^{X_1} z_2^{X_2} ... z_K^{X_K} \times z_1^{Y_1} z_2^{Y_2} ... z_K^{Y_K}\right] \\ &= E\left[z_1^{X_1} z_2^{X_2} ... z_K^{X_K}\right] E\left[z_1^{Y_1} z_2^{Y_2} ... z_K^{Y_K}\right] = \mathbf{G_X}(\mathbf{z})\mathbf{G_Y}(\mathbf{z}) \end{aligned}$$

Obtener las probabilidades $\mathbf{G_X}(1,...,1,z,1,...,1) = G_{X_i}(z)$, donde z está en la posición i. Entonces

$$P[X_i = 0] = \mathbf{G_X}(1,...,1,0,1,...,1),$$

donde 0 esta en la posición i. Similarmente,

$$P[X_i = 0, X_j = 0] = \mathbf{G_X}(1,...,1,0,1,...,1,0,1,...,1),$$

donde hay 0 en las posiciones i y j. Las otras probabilidades estan logradas con

$$P(X_i = k, X_j = \ell) = \frac{1}{k!\ell!} \times \frac{\mathbf{G_X}(1,...,1,z,1,...,1,y,1,...,1)}{\partial z^k \partial y^\ell}\bigg|_{z=y=0}$$

donde z y y están en las posiciones k y ℓ.

Cálculo de momentos: $\mathbf{G_X}(1,...,1) = 1$, y

$$E[X_i] = \frac{\partial \mathbf{G_X}(1,...,1,z,1,...,1)}{\partial z}\bigg|_{z=1},$$

$$E[X_i(X_i-1)...(X_i-k+1)] = \frac{\partial^k \mathbf{G_X}(1,...,1,z,1,...,1)}{\partial z^k}\bigg|_{z=1}$$

donde z está en la posición i.

Correlación: Por $j \neq i$ tenemos:

$$E[X_i X_j] = \frac{\partial^2 \mathbf{G_X}(1,...,1,z,1,...,1,y,1,...,1)}{\partial z \partial y}\bigg|_{z=1,y=1}$$

3.5.3 Número de llegadas en un intervalo

Sea $X_i(t)$ un proceso estocástico de Poisson con parámetro λ_i, $i = 1,...,K$. Suponemos que $X_i(t)$ son independientes, $i = 1,...,K$. El número de llegadas $N_i(Q)$ en el intervalo $[0,Q]$ (donde Q es fijo) es una VA con distribución de Poisson con parámetro $\lambda_i Q$.

Ahora, cuál será la FGP de $\mathbf{N}(Q)$ si Q es una VA?

$$\begin{aligned}
\mathbf{G}_{\mathbf{N}(Q)}(\mathbf{z}) &= E\Big(z_1^{N_1(Q)} z_2^{N_2(Q)} \cdots z_K^{N_K(Q)}\Big) \\
&= E\Big[E\Big(\, z_1^{N_1(Q)} z_2^{N_2(Q)} \cdots z_K^{N_K(Q)} \Big|\, Q\Big)\Big] \\
&= E\left(E\left[e^{-\lambda_1 Q(1-z_1)}|Q\right]E\left[e^{-\lambda_2 Q(1-z_2)}|Q\right]...E\left[e^{-\lambda_K Q(1-z_K)}|Q\right]\right) \\
&= E\left[e^{-\lambda_1 Q(1-z_1)} e^{-\lambda_2 Q(1-z_2)} ... e^{-\lambda_K Q(1-z_K)}\right] \\
&= E\Big[\exp\Big(-\Big(\sum_{i=1}^K \lambda_i(1-z_i)\Big)Q\Big)\Big] = Q^*\Big(\sum_{i=1}^K \lambda_i(1-z_i)\Big).
\end{aligned}$$

Correlaciones: Por $j \neq i$ tenemos

$$\begin{aligned} E[N_i(Q)N_j(Q)] &= \left.\frac{\partial^2 \mathbf{G}_{\mathbf{N}(Q)}(1,...,1,z,1,...,1,y,1,...,1)}{\partial z \partial y}\right|_{z=1,y=1} \\ &= \left.\frac{\partial^2 Q^*(\lambda_i(1-z)+\lambda_j(1-y)}{\partial z \partial y}\right|_{z=1,y=1} \\ &= \lambda_i \lambda_j E[Q^2]. \end{aligned}$$

Entonces

$$Cov[N_i(Q)N_j(Q)] = E[N_i(Q)N_j(Q)] - E[N_i(Q)]E[N_j(Q)] = \lambda_i \lambda_j Var[Q].$$

Ejemplo 3.5.1 *Sea Q exponencial con parámetro μ.*
¿Cual es la FGP multi-dimencional de $\mathbf{N}(Q)$ (el nímero de llegadas durante este intervalo)?
¿Cual es la la correlación $E[N_i(Q)N_j(Q)]$, $j \neq i$?

Solución:
Recordamos que

$$Q^*(s) = \frac{\mu}{\mu + s}.$$

Entonces,

$$G_{\mathbf{N}(Q)}(z) = Q^*(\sum_{i=1}^{K} \lambda_i(1-z_i)) = \frac{\mu}{\mu + \sum_i \lambda_i(1-z_i)}.$$

Además, tenemos $E[N_i(Q)N_j(Q)] = \lambda_i \lambda_j E[Q^2]$.

3.6 Transformada de Laplace Stieltjes de Vectores Aleatorios

3.6.1 Definición

Sea $\mathbf{X} = \{X_1, X_2, ..., X_K\}$ un vector aleatorio no-negativo. La **Transformada de Laplace Stieltjes** de $\mathbf{X}$ (o de su distribución $F_{\mathbf{X}}$) está definida como

$$\mathbf{X}^*(\mathbf{s}) = E\left(e^{-(s_1X_1+s_2X_2\cdots s_KX_K)}\right) = \int_0^\infty e^{-\mathbf{s}\cdot\mathbf{x}} dF_{\mathbf{X}}(\mathbf{x})$$

donde $F_{\mathbf{X}}(\mathbf{x})$ es la distribución de probabilidad de $\mathbf{X}$.

3.6.2 Propiedades de la TLS

1. Si $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$ y $\mathbf{X}$ y $\mathbf{Y}$ son independientes, entonces

$$\mathbf{Z}^*(\mathbf{s}) = \mathbf{X}^*(\mathbf{s})\mathbf{Y}^*(\mathbf{s}).$$

2. $\mathbf{X}^*(\mathbf{0}) = 1$ y $\mathbf{X}^*(0,...,0,s_i,0,...,0) = X_i^*(s_i)$.

3. **Momentos y correlación:**

$$-\left.\frac{\mathrm{d}\mathbf{X}^*(0,...,0,s_i,0,...,0)}{\mathrm{d}s_i}\right|_{s_i=0} = E[X_i],$$

$$(-1)^m \frac{1}{m!}\left.\frac{\mathrm{d}^m\mathbf{X}^*(0,...,0,s_i,0,...,0)}{\mathrm{d}s_i^m}\right|_{s_i=0} = E[X_i^m], \qquad m>0,$$

y para $j \neq i$,

$$\left.\frac{\mathrm{d}^2\mathbf{X}^*(0,...,0,s_i,0,...,0,s_j,0,...,0)}{\mathrm{d}s_i\mathrm{d}s_j}\right|_{s_i=s_j=0} = E[X_iX_j].$$

3.7 Exercicios

1. Obtener el primero y el segundo momento de las distribuciones siguientes, de sus FGP: (i) Binomial $B(n,p)$, (ii) Poisson(λ), (iii) Geométrica (p),

2. Obtener el primero y el segundo momento de las distribuciones siguientes, de sus TLS: (i) Uniforme(a,b), (ii) Erlang(2).

3. Consideramos la cola G/M/1/0. Sea $\lambda = 1/E[T_n]$. Mostrar que la distribución de T_n que minimisa el probabilidad de bloqueo, por esta λ, es la constanta: $T_n = 1/\lambda$. ¿Cual es la distribución de T_n que maximisa el rendimiento por esta λ?

4. Sea Q, una VA con distribución Erlang(k,μ). Sea $N(t)$ un proceso de Poisson con intensidad λ. ¿Cual es la distribución del número de llegadas durante Q? (Ver la Sección 3.4.)

Capítulo 4

Modelos de colas y cadenas de Markov

4.1 Clasificación y caracterización de colas

Hay dos tipos de colas, o de redes de colas:

- **Sistemas de perdidas:** con colas finitas. Las llegadas son perdidas si la cola está llena.
- **Sistemas de espera:** con colas infinitas (que sirve tambien de modelo para colas finitas pero grandes).

Caracterización de colas

1. según el processo de llegadas,
2. según la distribución del tiempo de servicio
3. según el número de servidores (que pueden trabajar al mismo tiempo). Usamos S (servidor) o C (canales) para este numero. Trabajaremos con $S = \infty$ también.
4. según el tamaño de la cola $N \geq 0$,
5. según el régimen:
   - FIFO (First In First Out) o FCFS (First Come First Served), aquí el orden del servicio sigue el orden de llegadas.
   - LIFO (Last In First Out), aquí el servicio de una tarea se para por cuando hay una llegada de una nueve tarea.
   - PS (Processor Sharing), el servidor divide su capacidad y sirve todos las tareas que esperan simultanamente. El servicio de una tarea empeza inmediatamente cuando llega. Naturalmente, el servicio de una tarea dura más cuando hay más tareas en la cola.
   - RS (Random Selection), el orden de servicio es aleatorio.
   - Prioridades

Para definir la distribución de los tiempos entre llegadas y la distribución del tiempo de servicio, usamos las notaciones:
M - para una distribución exponencial,

D - para tiempos deterministas (constantes),
G - para una distribución general (donde conocemos al menos la esperanza y el segundo momento),
GI - para una distribución general pero independiente (el tiempo de servicio de todos los clientes (paquetes) o el tiempo entre sus llegadas son independientes). A veces se escribe G también.
E_k - para una distribución de Erlang (o de Gamma) de orden k,
P_k - para una "phase type distribution".

Escribimos: $A/B/C/N/R$, donde A represente el tipo de llegadas, B representa la distribución del tiempo de servicio, C - el número de canales, N el tamaño de la cola, y R el régimen.

Por ejemplo: M/M/1/∞/FCFS se usa para un proceso de Poisson de llegadas (los tiempos entre llegadas tienen una distribución exponencial), tiempos de servicio con distribución exponencial, un solo servidor, tamaño infinito de cola, y el régimen de FCFS.

4.2 El paradoja del tiempo de espera

Suponemos que los buses llegan a una parada en los tiempos $\{\tau_i\}_{i=1}^{\infty}$, donde $\tau_0 = 0$. Sea $a_i = \tau_i - \tau_{i-1}$.

Concentremonos en los n primeros que han llegado. Tenemos $T := \tau_n = \sum_{i=1}^{n} a_i$.

Suponemos que un pasajero llega en un tiempo aleatorio t (con distribución uniforme en $[0, T]$). ¿Cual es su tiempo promedio de espera $E[W(t)]$?

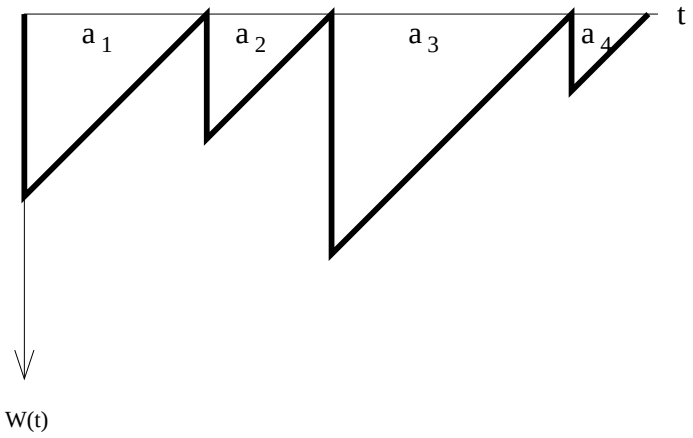

Figure 4.1: Paradoja del tiempo de espera

La probabilidad de que un pasajero llege durante $[\tau_i, \tau_{i+1})$, $i = 0, ..., n-1$ es

$$p_i = \frac{a_i}{T}.$$

Sea J el indice del intervalo donde llega el pasajero. Entonce si llega en $t \in [\tau_i, \tau_{i+1})$ entonces $J = i$. La esperanza de $W(t)$ condicionado que el viajero llegue durante $[\tau_i, \tau_{i+1})$, $i = 0, ..., n-1$ es

$$E[W(t)|t \in [\tau_i, \tau_{i+1})] = E[W(t)|J = i] = \frac{a_i}{2}.$$

Entonces,

$$E[W] = E[E(W|J)] = \sum_{i=1}^{n-1} \frac{a_i}{T} \times \frac{a_i}{2} = \frac{1}{T} \sum_{i=1}^{n-1} \frac{a_i^2}{2}.$$

Podemos escribirlo como

$$E[W] = \frac{\frac{1}{n}\sum_{i=1}^{n-1} a_i^2}{2\frac{1}{n}\sum_{i=1}^{n-1} a_i} = \frac{E[A^2]}{2E[A]}$$

donde A es el tiempo aleatorio entre llegadas del bus.

Conclusiones:

- Si $A = a$ es una constante, entonces $E[A^2] = a^2$ y $E[W] = E[A]/2 = a/2$.
- Si A tiene un distribución exponencial con parámetro μ, entonces $E[A] = \mu^{-1}$, $E[A^2] = 2\mu^{-2}$ y $E[W] = \mu^{-1} = E[A]$.
- En general, como $E[A^2] \geq E[A]^2$, tenemos siempre que $E[W] \geq E[A]/2$, con igualdad solamente cuando A es una constante.
- Aun si a_i tienen todos la misma distribución, a_J no una distribución diferente, y

$$E[a_J] = E[A^2]/E[A] \geq E[A] = E[a_1].$$

El paradoja: Vimos que cuando A tiene una distribución exponencial con parámetro μ, tenemos que esperar en promedio $1/\mu$ hasta el próximo bus. Además, por la misma razón, ya pasado un tiempo promedio de $1/\mu$ desde el último bus que pasó. Entonces el tiempo promedio entre dos buses es $2/\mu$. Pero sabemos que el tiempo entre llegadas tiene una distribución exponencial con parámetro $1/\mu$, entonces su esperanza deberia ser $1/\mu$!

4.3 Tiempos residuales

Suponemos que a_i tienen una distribución general con densidad $f(t)$.

- Tenemos una probabilidad proporcional a t de llegar durante un intervalo de duración t.
- La probabilidad de tener un intervalo de duración entre t y $t+\Delta$ (para Δ pequeña) es aproximadamente $f(t)\Delta$.

Sea $A(t)$ el evento que la longitud de un intervalo es entre t y $t+\Delta$.

Entonces, la probabilidad que un intervalo donde hay una llegada sea de duración entre t y $t+\Delta$ (para Δ pequeña) es

$$\begin{aligned} g(t)\Delta &= P(hayunallegadaduranteelintervalo \cap A(t)) \\ &= P(hayunallegadaduranteelintervalo|A(t))P(A(t)) = (\alpha t)(f(t)\Delta) \end{aligned}$$

Entonces, la densidad de probabilidad de la duración de un intervalo donde hay una llegada aleatoria es

$$g(t) = \alpha t f(t).$$

Para conocer α integramos:

$$1 = \int_0^\infty g(t)dt = \int_0^\infty \alpha t f(t)dt = \alpha \int_0^\infty t f(t)dt = E[A].$$

Entonces $\alpha = 1/E[A]$, y

$$g(t) = \frac{tf(t)}{E[A]}.$$

Logramos

$$E[W] = \int_0^\infty tg(t)dt = \frac{E[A^2]}{E[A]}.$$

La distribución de probabilidad de la duración γ de un intervalo donde hay una llegada aleatoria es

$$P(\gamma \leq x) = \frac{[1 - F(y)]dy}{E[A]},$$

donde $F(y)$ es la distribución de A.

Llamamos el tiempo desde la última llegada de un bus hasta una llegada aleatoria de un pasajero el *tiempo residual pasado* y lo escribimos W^{pas}.

Llamamos el tiempo desde una llegada aleatoria de un pasajero hasta la próxima llegada de un bus el *tiempo residual futuro*, y lo escribimos W^{fut}.

Tenemos

$$E[W^{pas}] = E[W^{fut}] = \frac{E[A^2]}{2E[A]}.$$

4.4 Cadenas de Markov y colas

Una cadena de Markov es un proceso estocástico en tiempo discreto $\{X_n, n = 0, 1, 2, ...\}$ que puede tener un número finito o infinito (enumerable) de valores (estados), y tal que para todos los estados $i_0, i_1, ..., i_{n-1}, i, j$, tenemos

$$P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}, X_n = i) = P(X_{n+1} = j | X_n = i) =: P_{i,j}^{n,n+1}.$$

Si $P_{i,j} = P_{i,j}^{n,n+1}$ no depende de n entonces la cadena tiene probabilidades de transiciones estacionarias. Escribimos estas probabilidades como una matrix:

$$P = \begin{pmatrix} P_{00} & P_{01} & ... \\ P_{10} & P_{11} & ... \\ \vdots & \vdots & \end{pmatrix}$$

La linea i describe la probabilidad de pasar de un estado i a todos los estados j en un paso. Tenemos para todo i, $\sum_{j=0}^\infty P_{i,j} = 1$, y $P_{i,j} \geq 0$ para todos i y j. Si sabemos que $P(X_0 = i) = p_i$, entonces

$$P(X_0 = i, X_1 = i_1, ..., X_n = i_n) = p_i P_{i,i_1} \cdots P_{i_{n-1},i_n}.$$

Ejemplo 4.4.1 *Supongamos que el profesor hace un examen cada semana, y escoje el dia del examen en una manera aleatoria con distribución uniforme sobre los 5 dias de la semana. Supongamos que X_i vale 1 si hay un examen en el dia i y 0 sinon. ¿X_i es una cadena de Markov?*
Non! por ejemplo, si sabemos que en el segundo dia de la semana no habia examen, pero no sabemos si habia o non examen en el primer dia, entonces $P(X_3 = 1 | X_2 = 0) = 1/4$. Pero no es el mismo que si sabemos también lo que pase en el primero dia:

$$P(X_3 = 1 | X_2 = 0, X_1 = 0) = 1/3, \qquad (X_3 = 1 | X_2 = 0, X_1 = 1) = 0.$$

4.4.1 Ejemplo: colas infinitas en tiempo discreto

Consideramos una cola infinita con un servidor. Suponemos que el tiempo de servicio de un paquete es una unidad. Si hay paquetes esperando en el comienzo de una unidad entonces sera servido. Sinon, esperamos hasta la proxima unidad.

En la n-ésima unidad de tiempo llegan un número aleatorio de llegadas ξ_n con una distribución de probabilidad:

$$a_k := P(\xi = k).$$

Por ejemplo, si la distribución de ξ_n es Poisson entonces

$$a_k = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Si es geométrica entonces

$$a_k = (1-p)^{k-1}p.$$

Sea X_n el número de paquetes en el sistema al final del tiempo n. Entonces:

$$X_{n+1} = \begin{cases} X_n - 1 + \xi_{n+1}, & X_n \geq 1, \\ \xi_{n+1}, & X_n = 0. \end{cases}$$

X_n es una cadena de Markov con una matriz de probabilidad:

$$P = \{P_{ij}\} = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & \dots \\ a_0 & a_1 & a_2 & a_3 & \dots \\ 0 & a_0 & a_1 & a_2 & \dots \\ 0 & 0 & a_0 & a_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Esta cola sera estable si $\rho := \sum_{k=0}^{\infty} k a_k < 1$, e inestable si $\rho > 1$.

4.4.2 Ejemplo: la cola M/G/1

Las llegadas siguen un proceso de Poisson con parámetro λ. El servicio es general con una distribución B.

Consideramos la cola en tiempos de salidas: X_n es el número de paquetes inmediatamente despues el n-ésimo fin de servicio, y sea $P_{ij} = P(X_{n+1} = j | X_n = i)$.

Sea ξ_n el número de llegadas durante un servicio y notamos $a_k = P(\xi_n = k)$. Entonces

$$a_k = E[P(\xi_n = k|V = v)] = \int_0^\infty P(\xi_n = k|V = v)dB(v) = \int_0^\infty e^{-\lambda v}\frac{(\lambda v)^k}{k!}dB(v).$$

De nuevo, esta cola será estable si $\rho := \sum_{k=0}^{\infty} k a_k < 1$, e inestable si $\rho > 1$. Aqui, $\rho = \lambda E[V] = \lambda/\mu$ (la esperanza del número de llegadas durante un tiempo V).

Ejemplo 4.4.2 *En una cola M/M/1, $B(v) = P(V \leq v) = 1 - e^{-\mu v}$, entonces*

$$a_i = \int_0^\infty e^{-\lambda v}\frac{(\lambda v)^i}{i!}\mu e^{-\mu v}dv$$

$$\begin{aligned} &= \mu\lambda^i \int_0^\infty e^{-(\lambda+\mu)v} \frac{(v)^i}{i!} dv \\ &= \frac{\mu}{\lambda+\mu} \left(\frac{\lambda}{\lambda+\mu}\right)^i \int_0^\infty (\lambda+\mu)^{i+1} e^{-(\lambda+\mu)v} \frac{(v)^i}{i!} dv \\ &= \frac{\mu}{\lambda+\mu} \left(\frac{\lambda}{\lambda+\mu}\right)^i = \frac{\rho^i}{(1+\rho)^{i+1}}. \end{aligned}$$

La última integral es igual a 1 porque es la integral de una densidad de probabilidad (de la distribución de E_{i+1} con parámetro $\lambda+\mu$). Tenemos

$$a_i = a_{i-1} \frac{\rho}{1+\rho}.$$

Ejemplo 4.4.3 *Consideramos una cola M/D/1 Tenemos*

$$a_0 = a^{-\rho},$$

$$a_i = e^{-\rho} \frac{\rho^i}{i!} = a_{i-1} \frac{\rho}{i}$$

Tenemos la misma dinámica que en el modelo precendente:

$$X_{n+1} = \begin{cases} X_n - 1 + \xi_{n+1}, & X_n \geq 1, \\ \xi_{n+1}, & X_n = 0. \end{cases}$$

En el primer caso, después que un paquete sale, empieza inmediatamente el servicio de otro paquete en el sistema . ξ_{n+1} es el número de llegadas durante este servicio. En el otro caso, no quedan patuetes en el sistema; el próximo paquete a ser servido es también el próximo que sale del sistema, y cuando sale hay ξ_{n+1} paquetes en el sistema que corresponden a los llegados durante su servicio.

4.4.3 Propiedades de cadenas de Markov: repaso

Sea X una cadena de Markov con un espacio de estados S discreto, con probabilidades de transiciones estacionaras P. Sea $P_{ij}^n = P(X_{m+n} = j | X_m = i)$. El teorema de Chapman y Kolmogorov dice que

$$P_{ij}^n = \sum_{k \in S} P_{ik}^r P_{kj}^{n-r}, \qquad 0 \leq r \leq n,$$

donde

$$P_{ij}^0 := \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Podemos escribirlo en forma matricial

$$P^n = P^r P^{n-r}.$$

Definición 4.4.1 *Los estados i, j se dicen comunicantes si hay enteros $m \geq 1$ y $n \geq 1$ tal que $P_{ij}^n > 0$ y $P_{ji}^m > 0$.*
Una cadena se llama irreductible si todos los pares de estados son comunicantes.

- Sea f_{ij}^n la probabilidad de llegar de i a j después n pasos por la primera vez.

$$f_{ij}^n = P(X_n = j, X_k \neq j, k = 1, ..., n-1 | X_0 = i).$$

  Tenemos $f_{ii}^0 = f_{ij}^0 = 0$, y $f_{ij}^1 = P_{ij}$.

- Sea $f_{ii} := \sum_{n=1}^{\infty} f_{ii}^n$ la probabilidad de regresar a i si salimos de i.

- Sea m_i la esperanza del tiempo para regresar a i:

$$m_i := \sum_{n=1}^{\infty} n f_{ii}^n.$$

Definición 4.4.2 *Un estado i se llama recurrente si $f_{ii} = 1$, y se llama transitorio si $f_{ii} < 1$.*

Teorema 4.4.1 *El estado i es recurrente si y solamente si $\sum_{n=1}^{\infty} P_{ii}^n = \infty$.*

Definición 4.4.3 *Una cadena de Markov es cíclica si el estado puede presentarse como la union de $k > 2$ conjuntos disuntos de estados $\mathbf{X^1}, \mathbf{X^2}, ..., \mathbf{X^k}$, tal que en tiempo $nk + i$, $n \in N$, la cadena de Markov tiene valores en $\mathbf{X^i}$, $i = 1, ..., k$.*

Teorema 4.4.2 *Sea X una cadena de Markov irreductible y sin ciclos con un espacio discreto S. Entonces todos los estados son recurrentes (y la cadena se llama recurrente), o todos son transitorios (y la cadena se llama transitoria).*

Teorema 4.4.3 *Sea X una cadena de Markov irreductible, sin ciclos y recurrente con un espacio discreto S. Entonces para todo $j \in S$,*

$$\lim_{n\to\infty} P_{ij}^n = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P_{ij}^k = \frac{1}{m_i} =: \pi_i.$$

Si $\pi(i) > 0$ entonces i se llama recurrente positivo. Si $\pi(i) = 0$ entonces i se llama nulo recurrente. Vemos que i es recurrente positivo si $m_i < \infty$, y nulo recurrente si $m_i = \infty$.

Teorema 4.4.4 *Sea X una cadena de Markov irreductible, sin ciclos y recurrente con un espacio discreto S. Si hay i tal que $\pi_i > 0$ entonces $\pi_j > 0$ para todo j.*

Definición 4.4.4 *Sea X una cadena de Markov irreductible, sin ciclos y recurrente con un espacio discreto S. X se llama recurrente positiva si $m_i < \infty$, y nulo recurrente en caso contrario.*

Teorema 4.4.5 *Sea X una cadena de Markov irreductible, sin ciclos y recurrente con un espacio discreto S. Entonces existe un vector $\underline{\pi}$ tal que $\pi_j > 0, \forall j \in S$, con*

$$\lim_{n\to\infty} P_{jj}^n = \pi_j,$$

que satisface

$$\underline{\pi} P = \underline{\pi}, \qquad \sum_{j\in S} \pi_j = 1.$$

$\underline{\pi}$ se llama la probabilidad estacionaria de la cadena. La llamamos estacionaria porque si la cadena tiene la distribución de probabilidad $\underline{\pi}$ en tiempo 0, mantendra esta distribución en todos los tiempos.

4.5 La Cola M/G/1

4.5.1 Cálculo de la probabilidad estacionaria de la cola M/G/1

Recordamos que X_n en una cadena de Markov que significa el número de paquetes inmediatamente después el n-ésimo fin de servicio, y que las probabilidades de transición son

$$P = \{P_{ij}\} = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & \dots \\ a_0 & a_1 & a_2 & a_3 & \dots \\ 0 & a_0 & a_1 & a_2 & \dots \\ 0 & 0 & a_0 & a_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

con

$$a_k = \int_0^\infty e^{-\lambda v} \frac{(\lambda v)^k}{k!} dB(v).$$

Solución directa: Las probabilidades estacionarias son la solución de $\pi P = \pi, \sum_j p_j = 1$. La linea j de $\pi P = \pi$ se escribe como:

$$\pi_j = \pi_0 a_j + \sum_{i=1}^{j+1} \pi_i a_{j-1+1}. \tag{4.1}$$

Tenemos un sistema infinito de ecuaciones que tiene una solución. Definimos

$$G(z) = \sum_{j=1}^{\infty} \pi_j z^j,$$

$$\mathcal{A}(z) = \sum_{j=1}^{\infty} a_j z^j$$

(recordamos que $\mathcal{A}(z) = V^*(\lambda(1-z))$). Multiplicamos la linea j de (4.1) por z^j

$$\pi_j z^j = \pi_0 a_j z^j + \sum_{i=1}^{j+1} \pi_i a_{j-1+1}.z^j,$$

y sumamos sobre todos los j's:

$$\sum_{j=0}^{\infty} \pi_j z^j = \sum_{j=0}^{\infty} \pi_0 a_j z^j + \sum_{j=0}^{\infty} \sum_{i=1}^{j+1} \pi_i a_{j-1+1} z^j,$$

para obtener

$$G(z) = \pi_0 \mathcal{A}(z) + \frac{1}{z} \sum_{j=0}^{\infty} \sum_{i=1}^{j+1} \pi_i a_{j-1+1} z^{j+1} - \frac{1}{z} \sum_{j=0}^{\infty} \sum_{i=1}^{j+1} a_{j+1} z^{j+1}.$$

Entonces,

$$G(z) = \pi_0 \mathcal{A}(z) + \frac{1}{z}\Big(G(z)\mathcal{A}(z) - \pi_0 a_0\Big) - \frac{1}{z}\pi_0\Big(\mathcal{A}(z) - a_0\Big), \tag{4.2}$$

y obtenemos:

$$G(z) = \frac{\mathcal{A}(z)(z-1)\pi_0}{z - \mathcal{A}(z)}.$$

Ahora falta π_0. Podemos obtenerlo de (4.2) con la condición $A(1) = 1$ y $G(1) = 1$:

$$1 = G(1) = \pi_0 \lim_{z\uparrow 1} \frac{z-1}{1-A(z)} = \pi_0 \frac{1}{1-A'(1)}.$$

Entonces $\pi_0 = 1 - A'(1) = 1 - \lambda E[V] = 1 - \rho$. Obtenemos

$$G(z) = (1-\rho)\frac{V^*(\lambda(1-z))(z-1)}{z - V^*(\lambda(1-z))}.$$

Solución con FGP: Vimos que

$$X_{n+1} = \begin{cases} X_n - 1 + \xi_{n+1}, & X_n \geq 1, \\ \xi_{n+1}, & X_n = 0. \end{cases}$$

En el estado estacionario, X_n y X_{n+1} tienen la misma distribución de probabilidad. Entonces:

$$G(z) = E[z^X] = E[z^{X-1\{X>0\}+\xi}] = E[z^{X-1\{X>0\}}]E[z^\xi] = E[z^{X-1\{X>0\}}]\mathcal{A}(z).$$

Tenemos:

$$\begin{aligned} E[z^{X-1\{X>0\}}] &= \sum_{i=0}^{\infty} \pi_i z^{i-1\{i>0\}} = \sum_{i=0}^{\infty} \pi_i z^{i-1+1\{i=0\}} = \sum_{i=1}^{\infty} \pi_i z^{i-1} + \pi_0 \\ &= \sum_{i=0}^{\infty} \pi_i z^{i-1} + \pi_0(1-z^{-1}) = \frac{1}{z}G(z) + \pi_0(1-z^{-1}) \end{aligned}$$

Entonces,

$$G(z) = \pi(0)\frac{\mathcal{A}(z)(z-1)}{z-\mathcal{A}(z)}.$$

En la proxima subsección mostraremos que $\pi(0) = 1 - \rho$.

Hasta aquí estudiamos la distribución o la FPG del proceso del número de paquetes en la cola en tiempos de salida. Vamos usar ahora dos teoremas importantes para lograr la distribución en cualquiera tiempo.

- Primero, usamos el Teorema de Burke [9] que dise que en el estado estacionara de cualquiera cola, la districión del estado de la cola just despues una salida es la misma districión que just antes de una llegada. Este teorema no necesita la distribución de Poisson de las llegadas.

- Como las llegadas siguen un proceso de Poisson, podemos usar la propiedad de PASTA (*Poisson Arrivals See Time Average* en ingles): la distribución del estado de un sistema de colas con un proceso de llegadas de Poison en cualquiero tiempo es el mismo que la distribución en tiempos de llegadas de paquetes

Entonces, la FPG que calculamos y la esperanza del número de paquetes que vamos a calcular son también los que vemos just antes llegadas y tambien en caulquier tiempo.

Prueba de la propiedad de PASTA:
Sean

- $X(t)$= el estado del sistema et tiempo t,
- $P_k(t) = P(X(t) = k)$, la probabilidad de $X(t) = k$ en cualquiera tiempo t fijo,
- $R_k(t)$= la probabilidad que $X(t-) = k$ cuando sabemos que hay una llegada en tiempo t.

Vamos a mostrar que $R_k(t) = R_k(t)$.

Sea $A(t, t + \Delta t)$ el evento que hay una llegada durante el intervalo $(t, t + \Delta t)$. Entonces

$$\begin{aligned} R_k(t) &= \lim_{\Delta t \to 0} P[X(t) = k | A(t, t + \Delta t)] = \frac{P[X(t) = k \cap A(t, t + \Delta t)]}{P[A(t, t + \Delta t)]} \\ &= \frac{P[A(t, t + \Delta t) | X(t) = k] P[X(t) = k]}{P[A(t, t + \Delta t)]} = \frac{P[A(t, t + \Delta t)] P[X(t) = k]}{P[A(t, t + \Delta t)]} = P[X(t) = k] \end{aligned}$$

4.5.2 Cálculo de $\pi(0)$ y la esperanza del número de paquetes

Buscamos $E[X]$. Podemos obtenerlo como

$$E[X] = G'(1).$$

Otra solución más fácil: Ya vimos que la probabilidad de que el servidor esté vacio es $\pi_0 = 1 - \rho$ y que

$$X_{n+1} = X_n - 1 + \delta_x + \xi_{n+1}.$$

donde $\delta_n = 1\{X_n = 0\}$ y tenemos $E[\delta_n] = \pi_0$. Entonces,

$$\begin{aligned} X_{n+1}^2 &= X_n^2 + \delta_n + (1 - \xi_{n+1})^2 + 2(X_n \delta_n + (\xi_{n+1} - 1)\delta_n + X_n(\xi_{n+1} - 1)) \\ &= X_n^2 + \delta_n + \xi_{n+1}(\xi_{n+1} - 1) + (1 - \xi_{n+1}) + 2((\xi_{n+1} - 1)\delta_n + X_n(\xi_{n+1} - 1)) \end{aligned}$$

Con $E[X_n^2] = E[X_{n+1}^2]$ (estado estacionario) obtenemos

$$0 = E[\delta_n] + E[\xi_{n+1}(\xi_{n+1} - 1)] + E[1 - \xi_{n+1}](2E[\delta_n] + E[X_n]).$$

Entonces

$$E[X] = \rho + \frac{E[\xi(\xi - 1)]}{2(1 - \rho)}.$$

Ahora,

$$E[\xi(\xi - 1)] = \mathcal{A}^{(2)}(z)\Big|_{z=1}$$

donde $\mathcal{A}(z) = V^*(\lambda(1 - z))$, entonces $E[\xi(\xi - 1)] = \lambda^2 E[V^2]$. Tenemos la formula de *Pollaczek-Khinchin*:

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

Sea L_q el número de paquetes en la cola, y Y el número en el servidor (1 o 0). Tenemos: $X = L_q + Y$, y $E[Y] = 1 - \pi_0 = \rho$. Entonces

$$E[L_q] = \frac{\lambda^2 E[V^2]}{2(1 - \rho)} = \frac{\rho^2 + \lambda^2 var[V]}{2(1 - \rho)}.$$

Vemos que

- la esperanza del número de paquetes en la cola crece con ρ, y converge a infinito cuando $\rho \to 1$. Vemos de nuevo que la cola no puede ser estable cuando $\rho \geq 1$.
- la esperanza del número de paquetes en la cola crece de una manera lineal con la varianza del tiempo de servicio.

Ejemplo 4.5.1 *En la cola M/M/1, $E[V^2] = 2/\mu^2$. Entonces*

$$E[L_q] = \frac{\lambda^2 E[V^2]}{2(1-\rho)} = \frac{\lambda^2}{\mu^2(1-\rho)} = \frac{\rho^2}{1-\rho} \quad y \quad E[X] = \frac{\rho}{1-\rho}.$$

Ejemplo 4.5.2 *En la cola M/D/1, $E[V^2] = 1/\mu^2$. Entonces*

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

es la mitad que la cola M/M/1!

4.5.3 Tiempo de espera

Sean

- W^q= El tiempo de espera en la cola,
- W= El tiempo del sistema: $W = W_q + V$.

Observación: sea X_n el número de paquetes cuando el paquete n sale, y su FGP en el estado estacionario $G(z)$. Entonces X_n son todas las llegadas durante el tiempo de sistema W_n. Sea W^* la TLS de W_n en el estado estacionario. Sabemos que el FGP del número de llegadas durante W_n es $W^*(\lambda(1-z))$. Entonces,

$$W^*(\lambda(1-z)) = W_q^*(\lambda(1-z))V^*(\lambda(1-z)) = G(z) = (1-\rho)\frac{V^*(\lambda(1-z))(z-1)}{z - V^*(\lambda(1-z)))}.$$

Entonces,

$$W_q^*(\lambda(1-z)) = (1-\rho)\frac{(z-1)}{z - V^*(\lambda(1-z)))}.$$

Con $s = \lambda(1-z)$ tenemos $z = 1 - s/\lambda$, entonces:

$$W_q^*(s) = \frac{(1-\rho)s}{\lambda(V^*(s) - (1-s/\lambda))} = \frac{(1-\rho)s}{\lambda V^*(s) - (\lambda - s)}.$$

Ejemplo 4.5.3 *En la cola M/M/1, $V^*(s) = \mu/(\mu+s)$. Entonces*

$$W_q^*(s) = \frac{(1-\rho)s}{\frac{\lambda\mu}{\mu+s} - (\lambda - s)}, \qquad W(s) = W_q^*(s)\frac{\mu}{\mu+s} = \frac{(1-\rho)s}{\lambda - \frac{\lambda-s}{\mu}(\mu+s)} = \frac{\mu-\lambda}{\mu-\lambda+s}.$$

Entonces, W_n tiene la distribución de probabilidad exponencial con parámetro $\mu - \lambda$!

La esperanza del tiempo de sistema (cola M/G/1):

$$E[W_q] = -W_q^*(0) = \frac{\lambda E[V^2]}{2(1-\rho)}, \qquad E[W] = W[W_q] + E[V] = \frac{\lambda E[V^2]}{2(1-\rho)} + E[V].$$

Recordamos que

$$E[L_q] = \frac{\lambda^2 E[V^2]}{2(1-\rho)}, \qquad E[X] = W[L_q] + E[Y] = \frac{\lambda^2 E[V^2]}{2(1-\rho)} + \rho.$$

Entonces:

$$E[X] = \lambda E[W], \qquad E[L_q] = \lambda E[W_q].$$

4.6 Colas G/G/1

El cargo de trabajo $\mathcal{V}(t)$: Definimos el cargo de trabajo como el tiempo que necesita para que todos los paquetes que estan en la cola salen. Es la suma del tiempo residual en tiempo t y el tiempo de servicio de todos los paquetes en la cola.

Sea $\mathcal{V}_n$ el cargo de trabajo just antes la n-ésima llegada. Entonces para colas con régimen FIFO, tenemos la relación:

$$W_q^n = \mathcal{V}_n.$$

Sea V_n el tiempo de servicio de la n-ésima llegada, y S_n el tiempo entre la n-ésima llegada y de la siguiente. Entonces tenemos la recurción que se llama la ecuación de Lindley:

$$\mathcal{V}_{n+1} = \max(\mathcal{V}_n + V_n - S_n, 0). \tag{4.3}$$

Sea $\xi_n = V_n - S_n$. La solución de este recurción es

$$\mathcal{V}_{n+1} = \max\Big(\xi_n, (\xi_n + \xi_{n-1}), ..., (\xi_n + \xi_{n-1} + ... + \xi_1), (\xi_n + \xi_{n-1} + ... + \xi_0 + \mathcal{V}_0), 0\Big). \tag{4.4}$$

Podemos mostrarlo por inducción. Cuando $n = 0$ (4.3) y (4.4) son los mismos. Suponemos que (4.4) es verda para $n-1$. Entonces usando (4.3) con la expresión (4.4) para $\mathcal{V}_n$, logramos (4.4) para $\mathcal{V}_{n+1}$ también.

4.7 Ejercicios

1. Hacen un graf del número de paquetes $L(t)$ en un sistema como función del tiempo t para un sistema con un solo servidor, cuando los tiempos de llegada y servicio son:

Cliente No.	1	2	3	4	5	6	7	8	9	10
tiempo de llegada	1.0	1.4	2.6	3.5	4.1	4.2	4.5	5.3	5.5	6.0
tiempo de servicio	0.8	2.0	1.0	0.7	0.6	0.9	0.2	1.3	0.5	1.1

   El regime de servicio es FIFO.
   Por cada paquete $i = 1, 2, ..., 10$, calcular el tiempo de espera en la cola $W_q(i)$ y el tiempo de sistema $W(i)$. Calcular también los valores promedios y las varianzas.

2. En una cola M/G/1, un paquete que encuentra el servidor ocupado, reune la cola con probabilidad 1-p, y sale sin servicio con probabilidad p.
   ¿Cual es la distribucion estacionar de la probabilidad del numero de paquetes en tiempos de salidas?

3. Hay en un banco dos servidores: a y b. Clientes llegan al banco segun en proceso de Poisson con velocidad λ, Un paquete que llega va primero al servidor a donde recibe un servicio V_1 con distribución exponencial con parámetro μ_1, y despues va al secondo servidor donde recibe un servicio V_2 con distribución exponencial con parámetro μ_2. Un consultador propone que cada servidor reciba la midad de las llegadas (un processo de Poison con velocidad $\lambda/2$) y de un servicio continue distriubido como $V = V_1 + V_2$ a cada paquete.
   ¿Cual es el tiempo promedio de espera en cada método?
   ¿Cual método es mejor?

4. Puequetes llegan segun un proceso de Poisson con parámetro λ a una cola para ser transmidos. El tiempo de transmición de un paquete es una VA V general. La transmición puede fallar con probabilidad p, y entonces tenemos que retransmitir el paquete. ¿Cual es el tiempo de sistema?

Capítulo 5

Invarientes en colas, colas con prioridades y vacaciones

Hay algunas relaciones sensillas en la teoria de cola que pueden usarse con distribuciones cualquieras de servicio y de tiempo de llegada. La ley de Little es una de estes relaciones. Una relación o una expreción que no depende de las distribuciónes de VAs pero solamente de sus esperanzas (o de otro parametro) se llama invariante.

5.1 Ley de Little

La ley de little hace una relación elegante entre una cuantidad del sistema: el número promedio de paquetes en el sistema, y una cuandidad del individo: su tiempo promedio de sistema (el tiempo de espera más el tiempo de servicio). Estas dos cuantitades son proporcionales segun la ley de Little. Una relación similar exsiste entre el número promedio de paquetes en la cola y el tiempo promedio de espera de un paquete.

Empesamos con algunas notaciones.

Consideramos un sistema de espera entre 0 y T. Definimos

- N_0:= el número de paquetes presentes en el tiempo 0,
- $N(T)$:= el número de llegadas durante $(0, T]$.
- A_0:= la suma del tiempo de sistema que necesitan todos los paquetes presentes en el tiempo 0,
- $A^*(T)$:= la suma del tiempo de sistema que necesitan todos los paquetes presentes en el tiempo T,
- $A(T)$:= la suma del tiempo de sistema que necesitan todos los paquetes presentes en el sistema durante $[0, T]$.
- $\lambda(T)$:= el rendimiento promedio de llegadas al sistema:

$$\lambda(T) = \frac{N(T)}{T}.$$

- $\overline{W}(T)$:= el tiempo promedio de sistema de los paquetes que han llegado durante $(0, T]$:

$$\overline{W}(T) = \frac{A(T) - A_0}{N(T)}$$

- $\overline{L}(T)$:= el número promedio de paquetes en el sistema (cola y servidor) durante $(0, T]$:

$$\overline{L}(T) = \frac{A(T) - A^*(T)}{T}.$$

Tenemos por definición de $\overline{W}(T)$:

$$\overline{L}(T) = \frac{A(T) - A^*(T)}{T} = \frac{\overline{W}(T)N(T)}{T} + \frac{A_0}{T} - \frac{A^*}{T} = \overline{\lambda}(T)\overline{W}(T) + \frac{A_0}{T} - \frac{A^*}{T}$$

Eso no depende del número de servidores, ni de la distribución, ni del orden de servicio.

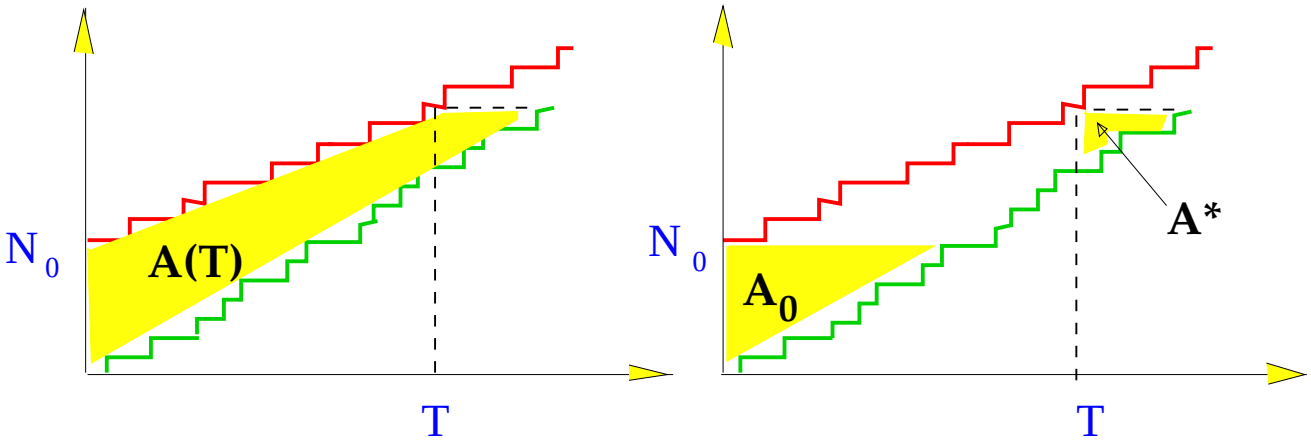

Figure 5.1: Ley de Little

Si existen

$$\lambda := \lim_{T\to\infty} \lambda(T), \qquad W := \lim_{T\to\infty} W(T),$$

y si $\lim_{T\to\infty}(A_0 - A^*)/T = 0$, entonces

$$L = \lambda W.$$

Esto se llama la ley de Little.

En sistemas ergódicos y estacionarios donde las esperanzas de L y W son iguales a los promedios temporales, tenemos:

$$E[L] = \lambda[W],$$

y $\lambda = E[N(t)]/t$ para todo $t > 0$.

Podemos hacer lo mismo para el tiempo de espera promedio W_q (en lugar del tiempo de estancia) y el número de paquetes en la cola L_q (en lugar de L) para obtener:

$$E[L_q] = \lambda[W_q].$$

Para una prueba más detallada, pueden ver [15].

5.2 La ocupación del servidor, cola G/G/.

Consideramos una cola con un servidor. El rendimiento de llegadas al servidor es λ, y el tiempo promedio de servicio es $E[V]$. Sea $\rho = \lambda E[V]$. (Si es una cola G/G/∞, y si la tasa de llegadas es menor que la tasa de salidas, entonces λ es también la tasa de llegadas a la cola).

Cual es la probabilidad p que en un tiempo arbitrario un servidor esté ocupado?

Aplicamos la Ley de Little al sistema que contiene solamente un servidor. Tenemos

$$E[L] = 1 \cdot p + 0 \cdot (1 - p) = p.$$

Además, tenemos $W = V$. Entonces:

$$p = E[L] = \lambda E[V] = \rho.$$

5.3 Otra manera de análisis de la cola M/G/1

Aquí, vamos a utilizar la ley de Little para obtener, de otra manera [8], la esperanza del tiempo de espera y del número de paquetes en la cola.

El tiempo de espera del paquete i es:

$$W_q^i = R_i + \sum_{j=i-N_q^i}^{i-1} V_j$$

donde

- R_i es el tiempo residual de servicio cuando llega el paquete i,
- N_q^i es el número de paquetes en la cola cuando llega el paquete i,
- W_q^i es el tiempo de espera del paquete i.

Entonces:

$$E[W_q] = E[R] + E[V]E[N_q]. \tag{5.1}$$

Sabemos que $E[L_q] = \lambda E[W_q]$. Ahora, como las llegadas siguen un proceso de Poisson, podemos usar la propiedad de PASTA (*Poisson Arrivals See Time Average* en ingles): la distribución del estado de un sistema de colas con un proceso de llegadas de Poison en cualquiero tiempo es el mismo que la distribución en tiempos de llegadas de paquetes y $E[N_q] = E[L_q]$. Entonces $[N_q] = \lambda E[W_q]$. y tenemos con $\rho = \lambda E[V]$:

$$E[W_q] = E[R] + \rho E[W_q].$$

Finalmente,

$$E[W_q] = \frac{E[R]}{1 - \rho}.$$

Sea $M(t)$ el número de servicios terminados durante $[0, t]$. Cuando un nuevo paquete j comienza su servicio, $r(t)$ salta de 0 a V_j, y después $r(t)$ baja linealmente. Entonces P-a.s.

$$E[R] = \lim_{t \to \infty} \frac{1}{t} \int_0^t r(\tau) d\tau,$$

Entonces,

$$E[R] = \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{M(t)} \frac{1}{2} V_i^2 = \frac{1}{2} \lim_{t\to\infty} \frac{M(t)}{t} \lim_{t\to\infty} \frac{\sum_{i=1}^{M(t)} V_i^2}{M(t)} = \frac{1}{2} \lambda V^{(2)}$$

y obtenemos la formula de Pollaczek-Khinchin

$$E[W_q] = \frac{\lambda V^{(2)}}{2(1-\rho)}.$$

5.4 Colas con prioridad

Cuando hay prioridades en la red, podemos ofreser servicios con meno tiempo de espera, meno variación de tiempo de espera, y meno pérdidas para tráfico más prioritario.

Por ejemplo, en redes de ATM, paquetes tienen un bit que indique su prioridad. Paquetes más prioridades pueden ser servidos antes los otros en enrutadores.

Consideramos una cola M/G/1 donde hay K clases de llegadas:

- La clase 1 tiene la prioridad la más alta, y la clase K - la más baja;
- Las llegadas de la clase i formen un proceso de Poisson con parámetro λ_i, $i = 1, ..., K$. Los K procesos de llegadas son independientes.
- Los tiempos de servicios son independientes. La esperanza del tiempo de servicio de clase k es v_k y el segundo momento es $v_k^{(2)}$.

Vamos a ver dos modelos de prioridad.

5.4.1 Prioridades non-preemptivas

En este método, cuando el servidor termina un servicio, escoge un paquete de la clase con la prioridad la más elevada, segun el orden de llegadas en esta clase. El sigue sirviendo este paquete aun si llegan paquetes con más prioridad durante este servicio.

Notación:

- $E[R^{(k)}]$ es la esperanza el tiempo residual de servicio cuando llega un paquete,
- $E[N^{(k)}]$ es la esperanza del número de paquetes de clase k en la cola,
- $E[W_q^{(k)}]$ es la esperanza del tiempo de espera de un paquete de la clase k,
- ρ_k se llama la carga de clase k y $\rho_k = \lambda_k v_k$.

Ahora, consideramos el sistema cuando llega un paquete (el n-ésimo) y spunoemos que somos en el régimen estacionaria. Hay dos objetos cuilla distribución no depende del orden del servicio de los paquetes ya en el sistema (y entonces de sus prioridads):

- La carga de trabajo $\mathcal{V}_n$,
- El tiempo residual R.

Estes catitades no dependen tampoco de la identitad de la llegada (su clase).

Notamos que en el régimen FIFO tuvimos que el tiempo de espera del n-ésimo paquete estuvo el mismo que la carga de trabajo $\mathcal{V}_n$. Pero ne es verdad en sistemas de prioridades!

Para computar $E[R]$, vamos a cambiar el orden de servicios y usar un modelo equivalente de una cola M/G/1 donde hay un solo proceso de llegadas de Poison con parámetro $\lambda = \lambda_1 + ... + \lambda_K$ (pero donde el régimen no es FIFO). Un paquete que llega apertenece a la clase k con probabilidad $p_k = \lambda_k/\lambda$. Entonces, la esperanza y el segundo momento del tiempo de servicio de una llegada cualquiera son

$$v = \sum_{k=1}^{K} \frac{\lambda_k}{\lambda} v_k, \qquad v^{(2)} = \sum_{k=1}^{K} \frac{\lambda_k}{\lambda} v_k^{(2)}.$$

Logramos que

$$E[R] = \frac{\lambda v^{(2)}}{2} = \frac{1}{2} \sum_{k=1}^{K} \lambda_k v_k^{(2)}.$$

La carga equivalente es

$$\rho = \lambda v = \sum_{k=1}^{k} \lambda_k v_k,$$

y la condición de stabilidad de este sistema es $\rho < 1$.

Ahora, como hicimos en (5.1) para lograr la formula de Pollaczek-Khinchin, tenemos aquí

$$E[W_q^{(1)}] = E[R] + v_1 E[N^{(1)}].$$

Usando la ley de Little, tenemos que $E[N^{(1)}] = \lambda_1 E[W_q^{(1)}]$. Entonces logramos:

$$E[W_q^{(1)}] = E[R] + \rho_1 E[W_q^{(1)}],$$

y finalemente

$$E[W_q^{(1)}] = \frac{E[R]}{1 - \rho_1}.$$

Ahora, el tiempo de espera de paquetes de clase 2 es dado por

$$E[W_q^{(2)}] = E[R] + (v_1 E[N^{(1)}] + v_2 E[N^{(2)}]) + (\lambda_1 v_1 E[W_q^{(2)}]).$$

Aquí, la expresión $v_1 E[N^{(1)}] + v_2 E[N^{(2)}]$ coresponde a la esperanza del tiempo de espera de todos los paquetes ya en la cola de clase 1 y 2. La expresión $\lambda_1 v_1 E[W_q^{(2)}]$ coresponde a la esperanza del tiempo de serivicio de todos los paquetes de clase 1 que llegan durante el tiempo de espera de nuestro paquete $C^{(2)}$. Estes paquetes llegan despues la llegada de $C^{(2)}$ pero seran servidos antes de el!

Usando la Ley de Little, logramos $E[N^{(2)}] = \lambda_2 E[W_q^{(2)}]$. Entonces:

$$E[W_q^{(2)}] = E[R] + \rho_1 E[W_q^{(1)}] + \rho_2 E[W_q^{(2)}] + \rho_1 E[W_q^{(2)}],$$

y finalemente

$$E[W_q^{(2)}] = \frac{E[R] + \rho_1 E[W_q^{(1)}]}{1 - \rho_1 - \rho_2}.$$

Substituando el valor de $E[W_q^{(1)}]$, logramos

$$E[W_q^{(2)}] = \frac{E[R]}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}.$$

De la misma manera podemos calcular la esperanza del tiempo de espera de paquetes de clase k:

$$E[W_q^{(k)}] = \frac{E[R]}{(1-\rho_1-...-\rho_{k-1})(1-\rho_1-...-\rho_k)}.$$

La esperanza del tiempo total de un paquete de clase k en el sistema es $E[W_q^{(k)}] + v_k$.

Ley de conservación

Calculamos La cuantidad $\sum_{k=1}^{K} \rho_k E[W_q^{(k)}]$. Por eso, notamos primero que

$$\frac{1}{a}\left(\frac{1}{x-a} - \frac{1}{x}\right) = \frac{1}{(x-a)(x)}$$

y entonces,

$$\begin{aligned} \rho_k E[W^{(k)}] &= \rho_k \frac{E[R]}{(1-\rho_1-...-\rho_{k-1})(1-\rho_1-...-\rho_k)} \\ &= E[R]\left(\frac{1}{1-\rho_1-...-\rho_{k-1}} - \frac{1}{1-\rho_1-...-\rho_k}\right). \end{aligned}$$

Tomando la suma, logramos

$$\sum_{k=1}^{K} \rho_k E[W_q^{(k)}] = E[R]\left(1 - \frac{1}{1-\rho}\right) = E[R]\frac{\rho}{1-\rho}.$$

Obtuvimos que esta suma es una constante, que depende solamente de $E[R]$ y de ρ, pero no del orden de prioritades.

¿Porque logramos aquí esta constante?

¿Que represente $E\left(\sum_{k=1}^{K} \rho_k E[W_q^{(k)}]\right)$?

La lei de conservación es una consecuencia de que la carga de trabajo $\mathcal{V}_n$ y el tiempo residual no dependen del orden de servicio, y e entonces no dependen del orden de prioridades. Usando la definición de la carga de trabajo y la ley de Little vemos que

$$E[\mathcal{V}_n] = E\left(\sum_{k=1}^{K} v_k E[N^{(k)}]\right) + E[R] = E\left(\sum_{k=1}^{K} \rho_k E[W_q^{(k)}]\right) + E[R]$$

para cualquier orden de servicio. Entonces $E\left(\sum_{k=1}^{K} \rho_k E[W_q^{(k)}]\right)$ no depende de los orden de prioridades. Además, $E[\mathcal{V}_n]$ es el mismo que en la cola equivalente de M/G/1 con el régimen FIFO donde tenemos:

$$E[\mathcal{V}_n] = E[W_q] = \frac{\lambda v^{(2)}}{2(1-\rho)} = \frac{\lambda \sum_{k=1}^{K} \lambda_k v_k^{(2)}}{2(1-\rho)} = \frac{\sum_{k=1}^{K} \lambda_k v_k^{(2)}}{2(1-\rho)}$$

$$E[R] = \frac{1}{2}\sum_{k=1}^{K} \lambda_k v_k^{(2)}.$$

Entonces

$$E\left(\sum_{k=1}^{K} \rho_k E[W_q^{(k)}]\right) = E[\mathcal{V}_n] - E[R] = E[R]\left(\frac{1}{1-\rho} - 1\right) = E[R]\frac{\rho}{1-\rho}.$$

5.4.2 Prioridades preemptivas

En este met método, cuando el servidor termine un servicio, escoje un paquete de la clase con la prioridad la más elevada, segun el orden de llegadas en este clase. Pero no sigue serviendo un paquete C si un paquete más prioridade llega durante este servicio: el interrumpe el servicio y lo recomenza solamente cuando no hay paquetes más prioridades que C en el sistema.

En este método, paquetes de classes $1, ..., k$ no sienten ningun influencia de otros paquetes, y para analyzar el tiempo de espera o de sistema de los clases $1, ..., k$ podemos olvidar la existencia de otras clases.

En prioridades preemptivas aun despues el tiempo de espera de un paquete, su servicio puede ser interpumpido si otro paquete con más prioridad llega. Entonces el tiempo de sistema es una medida más importante aquí que el tiempo de espera.

La esperanza del tiempo de sistema de un paquete C_k de classe k se compose de 3 partes:

1. La esperanza del tiempo hasta que todos los paquetes de clases $1, ..., k$ presente en su llegada salen. Este parte es la misma que la esperanza del tiempo de espera de C_k en una cola M/G/1 sin prioridad donde hay solamente llegadas te clases $1, ..., k$. Podemos escribirlo como

$$\frac{E[R_k]}{1-\rho_1-...-\rho_k}, \qquad \text{donde} \qquad E[R_k] = \frac{1}{2}\sum_{i=1}^{k}\lambda_i v^{(i)}.$$

2. La esperanza del tiempo para servir todos la paquetes de clases más prioidades, $1, ..., k-1$, que llegan durante que C_k esta en el sistema:

$$\sum_{i=1}^{k-1} v_i\lambda_i(E[W^{(k)}] + v_k).$$

3. La esperanza de su proprio tiempo de servicio.

Con este, logramos

$$\begin{aligned} E[W^{(1)}] &= \frac{E[R_1]}{1-\rho_1} + v_1 \\ E[W^{(k)}] &= \frac{E[R_k]}{(1-\rho_1-...-\rho_{k-1})(1-\rho_1-...-\rho_k)} + \frac{v_k}{1-(\rho_1+...+\rho_{k-1})}. \end{aligned}$$

5.5 Colas con vacaciones

5.5.1 Introducción

Hay muchas situaciones donde el servidor de una cola no está disponible todo el tiempo para prestar el servicio. Puede, por ejemplo, necesitar acciones de mantenimiento o reparación (se tiene una falla). A veces un servidor tiene que interrumpir el servicio para tomar café y comer un sandwich.

Para analisar colas con vacaciones tenemos que definir cuando sale el servidor para una vacación. Además, tenemos que definir a quien va a servir cuando regrese de vacaciones. Hay varios regimenes:

1. **El régimen "Gated".** El servidor sirve solamente los paquetes que encuentra cuando regresa de vacaciones. Los paquetes que llegan durante este servicio tienen que esperar hasta el final de

la próxima vacación. Por ejemplo, en el metro de Paris habian puertas que se cerraban cuando llegaba el tren. Los pasajeros en la estación podian subir, pero los que llegaban mientras que el tren estaba en la estación tenian que esperar la partida del tren para que las puertas se abrieran.

2. **El régimen exhaustivo.** En muchos sistemas, el servidor se queda sirviendo hasta que no hay paquetes en la cola. Entonces el servidor sirve todos los que encuentra cuando regresa de vacaciones y tambień todos los que llegan durante este servicio.

Cuando regresa el servidor de vacaciones, si no hay paquetes en la cola puede,

- tomar otra vacación. Este se llama el modelo de vacaciones repetidas.
- esperar hasta que llegue el primer paquete y empezar a servirlo. Este se llama el modelo de vacaciones isolados.

Aquí, vamos a estudiar vacaciones repetidas.

5.5.2 El modelo y primeros resultados

Definiciones:

- **Tiempo de servicio:** Cada llegada necesita un tiempo de servicio distribuido como una VA B con esperanza b y segundo momento $b^{(2)}$.
- **Vacaciones:** El tiempo D_n de la n vacación es una VA con esperanza d y segundo momento $d^{(2)}$.
- El tiempo T_n donde regresa el servidor de la n vacación se llama el n "tiempo de polling".
- Sea J_n el tiempo donde el servidor estaba trabajando antes de tomar la n vacación (permitimos $J_n = 0$, por ejemplo para vacaciones repetidas).
- El período $C_n = T_{n+1} - T_n$ entre dos tiempos de polling se llama el n ciclo. Tenemos $C_n = J_n + D_n$.
- Sea Q_n el número de paquetes esperando en la cola al tiempo T_n.
- $\mathcal{B}(i) := \sum_{j=1}^{i} B_j$ donde B_j son independientes con la distribución como B. Es el tiempo para servir i paquetes.
- $N(T)$:= El número de llegadas en un intervalo de tiempo que dura T.
- $\rho := \lambda b$ se llama la carga del sistema.

Suponemos que (J_n, D_n) son estacionarios y ergódicos. Entonces la proporción de tiempo (y la probabilidad) que el servidor trabaje es $E[J]/E[C]$. Pero sabemos que es también ρ. Entonces

$$\rho = \frac{E[J]}{E[C]} = \frac{E[C] - d}{E[C]}.$$

Entonces,

$$E[C] = \frac{d}{1 - \rho}.$$

Este expresión no depende de las distribuciones de las vacaciones, de los tiepmos entre llegadas y de los tiempos de servicios. Solo las esperanzas importan. Tampoco, no necesitamos independencia.

Ahora suponemos que los paquetes llegan a una cola (infinita) según un proceso de Poisson con parámetro λ, y qye los tiempos entre llegadas, los tiempos de servicio y de vacaciones son todos independientes.

Entonces, en el régimen gated tenemos que el número promedio de llegadas durante un ciclo es

$$E[N(C)] = \lambda \frac{d}{1-\rho}. \tag{5.2}$$

Adémas, es también el número promedio de paquetes presentes en la cola en el principio de ciclo.

En el regimen Exhaustive, el número de paquetes presentes en la cola en el principio de ciclo coresponde a las llegadas durante una vacación, y su esperanza es entonces λv.

5.5.3 El régimen exhaustivo

Ahora mostramos como usar la Ley de Little para obtener la esperanza del tiempo de espera en colas M/G/1 con vacaciones repetidas. Suponemos que a cada vez que el servidor empieza a trabajar, el sigue trabajando hasta que la cola esté vacia. Si no hay paquetes al final de una vacación, el servidor toma otra vacación. El tiempo de la vacación i es D_i, y $\{D_i\}$ son independientes e idénticamente distribuidos.

Sea R el tiempo hasta el principio del próximo servicio. Puede ser un tiempo residual de servicio o de vacación! Por las mismas razones que en el sistema sin vacaciones, obtenemos

$$E[W_q] = E[V]E[N] + E[R]$$

donde N es el número de paquetes en la cola just antes de una llegada. Usamos la Ley de Little y la propriedad de PASTA, logramos

$$E[W_q] = \frac{E[R]}{1-\rho}.$$

Sea $I(t)$ el número de vacaciones terminadas hasta t. Sea $M(t)$ el número de fin de servicios hasta t. Tenemos:

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t r(\tau)d\tau = \lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{M(t)} \frac{1}{2}V_i^2 + \lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{L(t)} \frac{1}{2}D_i^2. \tag{5.3}$$

La esperanza del tiempo de vacaciones en $[0,t]$ es $t(1-\rho)$. Entonces

$$\lim_{t\to\infty} \frac{t(1-\rho)}{L(t)} = d.$$

Entonces la segunda parte de (5.3) converge a $(1-\rho)d^{(2)}/d$. Obtenemos finalmente:

$$E[W_q] = \frac{E[R]}{1-\rho} = \frac{\lambda V^{(2)}}{2(1-\rho)} + \frac{d^{(2)}}{2d}. \tag{5.4}$$

Vemos que la esperanza del tiempo de espera crece de una manera lineal con la varianza del tiempo de servicio y del tiempo de vacación. Adémas, vemos que es finita para $\rho < 1$. $\rho < 1$ es la misma condición de estabilidad que tuvimos en colas sin vacaciones.

5.5.4 El régimen "Gated"

Tenemos la dinámica:

$$Q_{n+1} = N(C_n), \qquad C_n = \mathcal{B}(Q_n) + D_n.$$

Podemos entonces escribir una recursión con Q_n:

$$Q_{n+1} = N(\mathcal{B}(Q_n) + D_n), \tag{5.5}$$

o con C_n:

$$C_n = \mathcal{B}(N(C_{n-1})) + D_n. \tag{5.6}$$

El análisis a partir de Q_n se llama el método de ocupación de cola, y el análisis a partir de C_n se llama el método de tiempo de estación.

Calculo del segundo momento
De (3.5) tenemos que

$$E[N(C_n)^2] = \lambda^2 E[C_n^2] + \lambda E[C_n]$$

De (3.4) vemos que

$$E[\mathcal{B}(N(C_n))^2] = \Big(E[(N(C_n))^2] - E[N(C_n)]\Big)E[V]^2 + E[N(C_n)]V^{(2)}.$$

Entonces, tomando el segundo momento de cada lado de (5.6), logramos (sin escribir los indices)

$$E[C^2] = \rho^2 E[C^2] + \lambda E[C]V^{(2)} + d^{(2)} + 2d\rho E[C].$$

Entonces,

$$E[C^2] = \frac{\lambda d(V^{(2)} + 2dE[V]) + (1-\rho)d^{(2)}}{(1-\rho^2)(1-\rho)}.$$

Tiempo de espera // El tiempo de espera de una llegada cualquiera se composa de dos elementos:

1. C^{fut}, el tiempo residual futuro del ciclo donde habia la llegada,
2. el tiempo de servicio de todos los que llegaban antes este llegada en el mismo ciclo, c.f. durante el ciclo residual pasado, C^{pas}.

Tenemos (Sec. 4.3) que

$$E[C^{fut}] = E[C^{pas}] = \frac{E[C^2]}{2E[C]} = \frac{1}{1-\rho^2}\left((1-\rho)\frac{d^2}{2d} + d\rho + \frac{1}{2}\lambda V^{(2)}\right).$$

Entonces

$$E[W_q(gated)] = E[C^{pas}] + \rho E[C^{fut}] = (1+\rho)\frac{E[C^2]}{2E[C]} = \frac{d^2}{2d} + \frac{1}{1-\rho}\left(d\rho + \frac{1}{2}\lambda V^{(2)}\right).$$

Vemos que

$$E[W_q(gated)] = E[W_q(exhaustive)] + \frac{d\rho}{1-\rho}.$$

5.6 Ejercicios

Usando la recurción (5.5), calcular la esperanza y el secundo momento de Q_n. (Usar las relaciones (3.4) y (3.5)).

Capítulo 6

Redes locales

6.1 Métodos de acceso múltiple en redes locales

En cada red que permite a varios fuentes de conectarse a varios destinaciónes, hay el problema del acceso al red. Al meno de usar la solución costidiosa de tener lazos fisicos entre cada fuente y cada destinación, tenemos que repartir el acceso. Este problema es particulamente importante en comunicaciónes celulares y satellitares porque el mismo canal radio deve ser repartido y no podemos isolar comunicaciónes como lo hacemos en comunicaciónes con fibros opticos, buses o cables.

Hay varios methodos de acceso segun la repartición:

- La repartición estatica: una quantitad fija de recoursos es dedicado a una conección. Las métodos tradiciónales son: TDMA (reparto temporal: *Time Devision Multiple Access* en inglés), FDMA (reparto frequencial: *Frequency Devision Multiple Access* en inglés). CDMA (reparto en códigos: *Devision Multiple Access* en inglés).
- La repartición segun la réclamación: son métodos dinamicas que permiten de dar recursos (tiempo de transmición, frecuencias, etc) a una conección segun su necessidad punctual y segun la disponilidad. Exemplo: los token rings.
- El acceso aleatorio: Acqui vemos métodos de acceso independentes, que pueden entonces se hacer simultanamente. En consecuensia, pueden resultar colisionces (y pérdidas) de paquetes, y hay una necesidad de retransmisión. Ejemplos: ALOHA, Ethernet.

6.2 Metodos de repartición estatica

6.2.1 Reparto frequencial - FDMA

En el FDMA, toda la banda de frecuenza esta trinchada en M partes, y cada fuente tiene su propria banda de frecuenzia, donde puede transmitir independamente de losotros.

En este método hay un problema de intermodulación: hay interferencias de frecuencias vesinas. Por solucionar este problema, dejan zonas de frecuencias non-usadas entre las partes.

Para analyzar el FDMA, suponemos que los paquetes tienen un tamaño constante. Cada paquete de cada fuente necessita M unidades de tiempo para ser transmitido: como dividimos la banda de frequenzia, más la dividimos, más tiempo necessitamos para transmitir un paquete de cada fuente. Hay λ/M paquetes promedio que llegan a cada conección por unidad de tiempo.

Cada de M fuentes se comporte como una cola M/D/1, y obtenemos para cada fuente con $\rho = \lambda$, $\mu = 1/M$:

$$E[W_q] = \frac{(\lambda/M)M^2}{2(1-\lambda)} = \frac{\lambda M}{2(1-\lambda)}, \qquad E[W] = M + \frac{\lambda M}{2(1-\lambda)}.$$

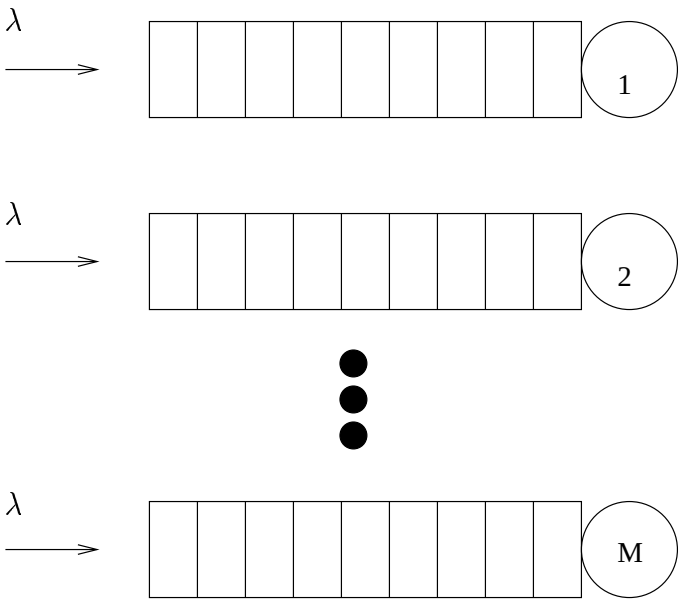

Figure 6.1: Modelisacion del FDMA

6.2.2 Reparto temporal - TDMA

En TDMA, definicimos un siclo trinchado en M periodos temporales. Cada fuente transmite durante un periodo fijo predeterminado. Si las peticiónes de capacidad no estan los mismo, pueden dar algunas periodos en cada ciclo a una misma fuente.

El TDMA se usa mucho en comunicacióncs satelites y mobiles. Por ejemplo, el satelito geostacionar Eutelsat usa este método con periodos de 2ms. A menudo se usan TDMA combinado con FDMA.

Las problemas de TDMA con relación a FDMA son

- TDMA necesita sincronización temporal,
- TDMA mecesita más poder instantáneo de transmición, porque toda la transmición esta consentrada en un periodo.

Para facilitar el analysis de TDMA, concideramos primero una variante de FDMA [7, p. 149]: suponemos que hay M fuentes que usan FDMA, pero suponemos que la transmición de un paquete puede empezar solamente en tiempo $M, 2M, 3M, ...$ Este variante se llama FDMA sincronizada (SFDMA). Para calcular el tiempo de espera promedio, usamos el modelo M/G/1 con vacaciones.

Las vacaciones tienen todos la misma duración M que necesita la transmición de un paquete en FDMA. Entonces $E[D] = M$ y $E[D^2] = M^2$. Usando (5.4), logramos

$$E[W_q(SFDMA)] = \frac{\lambda M}{2(1-\lambda)} + \frac{M}{2}.$$

Regresamos ahora al TDMA, y consideramos la fuente que puede transmitir en los tiempos $M, 2M, 3M, ...$ Vemos que el número de paquetes esperandos de esta fuente es el mismo que obtenimos en SFDMA. Entonces

$$E[W_q(TDMA)] = \frac{\lambda M}{2(1-\lambda)} + \frac{M}{2}, \qquad E[W(TDMA)] = E[W_q(TDMA)] + 1.$$

Vemos que

$$E[W_q(TDMA)] = E[W_q(FDMA)] + \frac{M}{2}, \qquad E[W(TDMA)] = E[W(FDMA)] - (\frac{M}{2} - 1).$$

Entonces, usando TDMA

- el tiempo medio de espera es más largo,
- el tiempo de transmición es más corte,
- el retardo medio de un paquete es más corto.

Obtenimos esto por $\lambda < 1$. Quando λ se aserca a 1, vemos que el tiempo medio de espera y el retardo medio devienen largos, y la diferencia entre FDMA y TDMA deviene insignificante.

6.3 El acceso aleatorio

El acceso aleatorio permite de usar eficientamente el canal porque al contrario de métodos de repartición estatica, el canal esta usado solamente quando lo necesitamos. Estas métodos son usados para transmitir paquetes de información y tambien en conjunto con métodos de repartición estaticas, para hacer una reservación de un parte del canal.

6.3.1 Descripción de ALHOA

ALOHA es el primero método de acceso aleaorio, y estaba dearollada y usada por la primera ves en una red de difusión por radio de paquetes entre las islas de Hawai, en 1970. Este método esta siempre usado en redes de satelites y redes mobiles celulares.

La transmición en ALOHA es completamente decentralizada. Al fin de cada paquete transmisido por cada fuente, un recibo (ACK) regresa a las fuentes para indicar si habia una colisión o si un paquete estaba bien recibido.

Este método tiene dos problemas:

- Hay un riesgo de un regime instable: más hay retransmiciónes, más hay colisiónes, que crean más retransmiciónes. En consequencia, a veces, el rendimiento baja considerablemente: todos las fuentes estan retransmitidos paquetes.

- El rendimiento maximo es de 0.18, que implica que todos las fuentas juntos ne pueden transmitir más que 18% del tiempo. Por ejemplo, si hay 60 fuentes, cada fuente puede transmitir solamente durante 0.3% del tiempo. Si tratamos de transmitir más que 18% del tiempo, el sistema se volve instable, y se crea una congestión grave con demasiado retransmiciónes.

Para amjorar el rendimiento de ALOHA, podemos aniadir un sincronización entre las fuentes. El tiempo esta discretizado: esta divisido en unidades de tiempo que se llaman "slots". La duración de un slot es egual al tiempo de ir y regressar maximal entre dos puntos del red. Las estaciónes son syncronisadas y saben quando empiesa un slot. Una estación puede empesar de transmitir un paquete solamente en el principio de un slot. Este método tiene el mismo problema de estabilidad, pero su rendimiento maximal es 0.36, el doble que ALOHA.

Ahora vamos presentamos un analisis de las dos versiónes de ALOHA, basado sobre [7].

6.3.2 Analisis de ALOHA ranurado

Comesamos a analisar el ALOHA ranurado. Suponemos que

- (1) la duración de cada paquete es de una unidad.
- (2) Las llegadas siguen un processo de Poisson con parametro λ.
- (3) Hay $m = \infty$ de estaciónes, y cada paquete llega a un otra estación.
- (4) Los tiempos de retransmiciónes estan escojidos tal que los tranmiciónes y retransmiciónes formen un proceso de Poisson con parametro $G > \lambda$.

La hipotesis que $m = \infty$ nos da una limite superiora por un modele donde no hay colas en las estaciónes, y donde si un paquete esta transmitido o si espera una retransmición en una estación, otro paquetes no pueden llegar a este estación.

Recordamos que la probabilidad de tener n transmiciónes (de paquetes nuevos o retransmitidos) en una unidad de tiempo esta

$$P(X = n) = G^n \frac{\exp(-G)}{n!}.$$

Entonces, la probabilidad de una buena transmición en una unidad de tiempo es

$$P(X = 1) = G \exp(-G).$$

Esta expreción representa el rendimiento.

Para conocer el rendimiento maximum de ALOHA ranurado, buscamos la condición de equilibrio, definido como la valor de G tal que

$$\text{tasa de salida} = \text{tasa de llegadas } (= \text{el rendimiento }).$$

Obtenemos la condición:

$$\lambda = G \exp(-G).$$

$G \exp(-G)$ como función de G tiene su maximom en $G = 1$, y el valor de λ en este maximo es $1/e = 0.386$.

Por un rendimiento λ inferior a 0.386, hay dos puntos de equilibrio, que estan la intercectión entre la constanta λ y la curba $G \to G \exp(-G)$. Sean G_1 G_2 los valores de G en estes equilibrios, donde

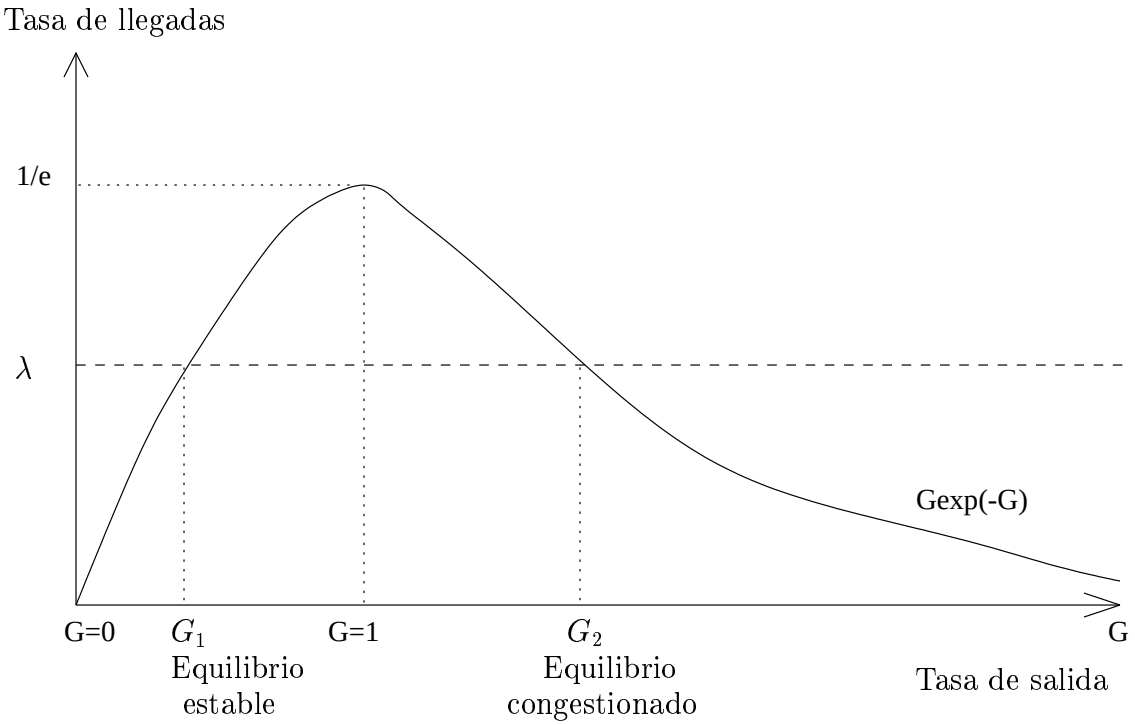

Figure 6.2: Los dos equilibrios en ALOHA ranurado

$G_1 < G_2$. El equilibrio G_2 se llama el equilibrio congestiónado, donde hay más retransmiciónes, y donde el tiempo hasta una retransmición con ecsito es más granda a causa de los colisiónes y retransmiciónes.

Esta analisis esta aproximada, pero un analisis exacte no da tambien dos equilibrio. Un tal analisis nos muenstra que en G_2 no solamente hay más retransmiciónes, pero tambien el rendimiento esta inferior.

6.3.3 Analisis de ALOHA

Usamos priomero el mismo modelo sobre el sistema, salvo que ahora suponemos que no hay sincronización temporal entre las estaciónes.

Suponemos que un paquete esta transmitido en el tiempo t. Definimos el periodo de vulnerabilidad como el interval $[t-1, t+1]$. Si un otro paquete esta transmitido en este interval, abra una colisión.

La probabilidad que no hay otra transmición o retransmición durante $[t-1, t)$ o en $(t, t+1]$ es $\exp(-G)$, entonces la probabilidad que sea una transmición sin colisión es

$$P_{\text{éxito}} = \exp(-2G).$$

El rendimiento es

$$\textit{Rendimiento} = G \times P_{\text{éxito}} = G\exp(-2G).$$

$G\exp(-2G)$ como función de G tiene su maximom en $G = 1/2$, y el valor de λ en este maximo es $1/(2e) = 0.193$.

Por un rendimiento λ inferior a 0.193, hay dos puntos de equilibrio, que estan la intercección entre la constanta λ y la curba $G \to G\exp(-2G)$, con el mismo interpretación que antes.

En este modelo no dicimos como escojer exactamente el tiempo de retransmición. No sabemos si es posibile te escojerlo tal que los tranmiciónes y retransmiciónes formen un proceso de Poisson con parametro $G > \lambda$.

Podemos usar un otro modelo para las retransmiciónes, donde un paquete colisiónado sera retransmitido despues un tiempo τ con distribución exponencial con parametro x: $P(\tau > a) = \exp(-xa)$.

Por una transmición en tiempo t, el periodo de vulnerabilidad es $[t-1, t+1]$. si hay n otras fuentes con paquetes a retransmitir in t, entonces el tiempo hasta la transmición o retransmición proxima (o pasada) tiene una distribución exponencial con parametro

$$G(n) = \lambda + nx.$$

Entonces la probabilidad que en t hay una transmición sin colisión es

$$P_{\text{éxito}} = \exp(-2G(n)).$$

El rendimiento es

$$Rendimiento(n) = G \times P_{\text{éxito}} = G(n)\exp(-2G(n)).$$

Logramos entonces las mismas conclusiónes que antes.

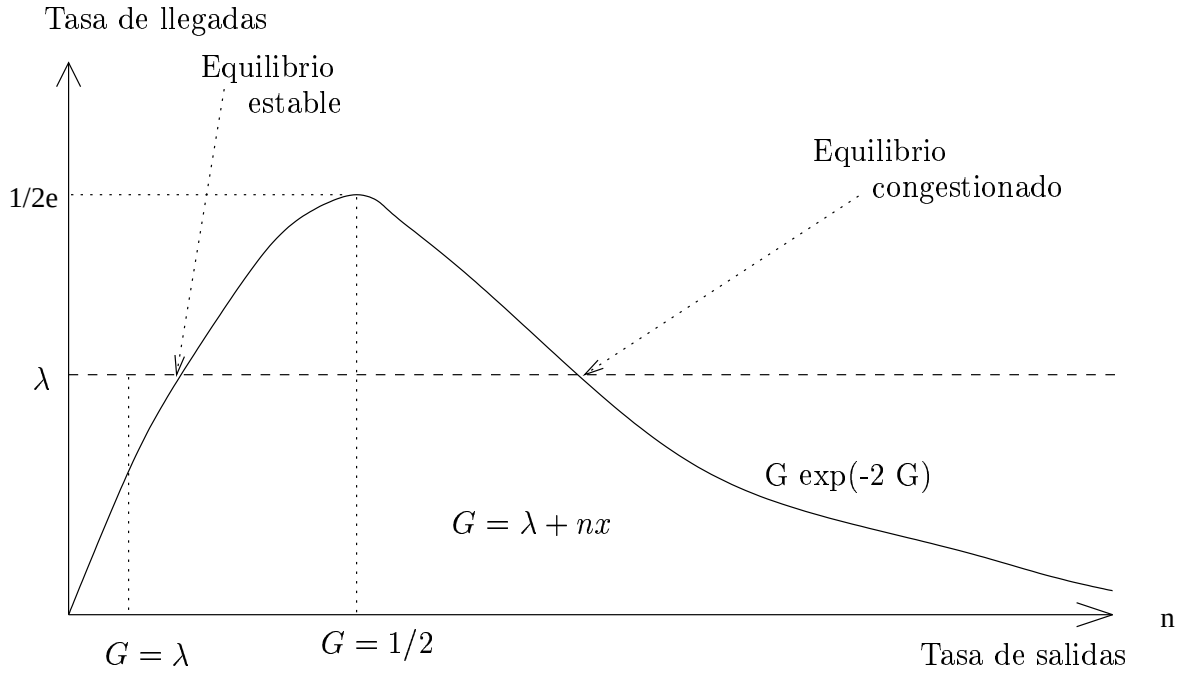

Figure 6.3: Los dos equilibrios en ALOHA

6.4 Ejercicios

1. Consideramos el cambio de los protocoles de Aloha, donde una estación que transmite escoje entre dos niveles de potencia de trancmición con la misma probabilidad con la misma probabilidad. Suponemos que si hay dos estaciones que transmiten en el mismo tiempo con potencias

diferentes, la más puisente no pierde su paquete.
¿Cual es el rendimiento de ALOHA?
¿Cual es el rendimiento de ALOHA ranurado?

Capítulo 7

Análisis de Protocolos de Internet

Nos concentraremos principalmente en el "protocolo de control del transporte" o TCP (*Transport Control Protocol* en inglés) que controla la congestión en la red.

7.1 Descripción general del TCP

7.1.1 Objetivos del TCP

El TCP es un protocolo que tiene tres objetivos:

- Adaptar el rendimiento de la transmisión de paquetes al ancho de banda disponible,
- Evitar la congestión en la red.
- Fiabilizar la comunicación y retransmitir los paquetes que se pierden en la red.

7.1.2 Control por ventana

Para alcanzar estos objetivos, cada paquete transmitido tiene un número de secuencia. Para controlar la velocidad de transmisión, la fuente no puede introducir en la red más que un número determinado de paquetes. Este número se llama *ventana* y la denotamos W. Después que W paquetes son transmitidos, la fuente no puede transmitir más paquetes hasta que sepa que un paquete ha salido de la red y llegado a su destino. Para saberlo, la fuente recibe información del destino que se llaman ACKNOWLEDGEMENT (acuso de recibo).

7.1.3 Acuses de recepción

El ACKNOWLEDGEMENT o ACK tiene dos objetivos:

- Regularizar el ritmo de transmisión del TCP, asegurando que los paquetes pueden ser transmitidos solamente cuando otros paquetes han salido de la red,
- Fiabilizar la transmisión: dar información a la fuente para que ella sepa si tiene que retransmitir algún paquete.

¿Cómo sabe el destino que falta un paquete?

¿Cómo sabemos que un paquete esta perdido?

¿Que información lleva el ACK?

El ACK dice a la fuente cual es el número de secuencia de paquete que espera. Suponemos que los paquetes número 1,2,...,6 llegaron al destino (en orden). Cuando el paquete 6 llega, el destino manda un ACK para decir que espera el paquete 7. Si el paquete 7 llega despues, dice que espera el 8. Ahora, si el paquete 8 está perdido, entonces el paquete siguiente que llega es el 9. En este instante, el destino manda un ACK diciendo que sigue esperando el paquete 8. Un tal ACK se llama "ACK repetido", porque informa de la recepción del mismo paquete (8).

El método siguiente se llama "ACK implícito". Es un método robusto a pérdidas de ACKs. Por ejemplo, suponemos que un ACK diciendo que esperamos el paquete 5 se pierde y despues llega el siguiente ACK que dice que esperamos el paquete 6. Entonces la fuente sabe que el paquete 5 llego también.

Un paquete del TCP es considerado perdido si

- Tres ACKs repetidos por el mismo paquete llegan a la fuente, o
- Cuando un paquete es transmitido, hay un temporizador que empieza a contar. Si su ACK no llega durante un período T_0, hay un "Time-Out" y el paquete se considera perdido.

¿Cómo escoger T_0? La fuente tiene una estimación del promedio del tiempo RTT, que es el tiempo que necesita un paquete para llegar al destino y para que su ACK regrese a la fuente, así que su variabilidad. T_0 esta derminado por estas estimaciones:

$$T_0 = \overline{RTT} + 4D$$

Donde $\overline{RTT}$ es la estimación del RTT, y D es la estimación de la variabilidad del RTT. Para estimar el RTT, medimos la diferencia M entre el tiempo de transmisión de un paquete y el tiempo que tarda su ACK. Despues calculamos:

$$\overline{RTT} \leftarrow a \times \overline{RTT} + (1-a)M,$$

$$D \leftarrow aD + (1-a)|\overline{RTT} - M|.$$

Para disminuir el número de ACKs en el sistema, el TCP usa frecuentemente el método de "ACK retrasado" donde solamente un ACK es transmitido después de cada dos paquetes que llegan.

7.1.4 Ventana dinámica

Desde principios de los años 80, durante muchas años el TCP operó con una ventana fija. Las redes eran inestables, y tenian períodos largos de congestión muy duros, donde los rendimientos bajaban mucho y habia muchas retransmisiones. Para solucionar este problema, Van Jacobson propuso [12] usar una ventana dinámica: su tamaño puede variar segun el estado de la red. Cuando la ventana es pequeña puede crecer rapidamente, y cuando es grande tiene que crecer muy lentamente. Cuando hay congestión, el tamaño de la ventana tendra que disminuir mucho. Así podemos solucionar rapidamente la congestión y al mismo tiempo usar bien los recursos del sistema.

Más precisamente, definimos un unbral W_{th} que se llama "slow start threshold" que representa nuestra estimación de la capacidad de la red. La ventana empieza con el valor de uno. Con cada

ACK que llega, la ventana crece de uno. Así transmitimos primero un solo paquete. Cuando su ACK llega, podemos transmitir dos paquetes. Cuando dos ACKs llegan, la ventana crece a 4, y podemos transmitir 4 paquetes. Vemos que hay un crecimiento exponencial. Este período de crecimiento se llama "slow start". Se llama así porque a pesar que el crecimiento es rápido, es más lento que si hubiese empezado directamente con $W = W_{th}$.

Cuando $W = W_{th}$, pasamos a un período que se llama "congestion avoidance", donde la ventana W crece de $1/\lfloor W \rfloor$ con cada ACK que llega. Entonces despues de transmitir W paquetes, W crece de 1. Si transmitimos los W paquetes en t, entonces en $t + RTT$ transmitimos $W + 1$, y en $t + 2RTT$ transmitimos $W + 2$, etc... Vemos que el crecimiento es lineal.

7.1.5 Perdidas y umbral W_{th} dinámico

No solamente W es dinámica, W_{th} también lo es. Fijamos W_{th} a la mitad del valor de W cuando hay una pérdida de paquete.

Hay algunas variantes del TCP: En la primera variante que se llama "Tahoe", si una pérdida es detectada, la ventana se reduce siempre a 1 y empieza un período de slow-start. Es una caida extrema de rendimiento.

En las variantes que se usan hoy, llamadas Reno o New Reno, la ventana baja a 1 solamente si la pérdida es detectada por un time-out. Si no, baja a la mitad del valor de la ventana, y no se inicia el "slow-start"; quedamos en "congestion avoidance".

7.2 Modelado del TCP

Los objetivos del modelado del TCP son:

- Dimensionar la red: para obtener un rendimiento dado, y cuales son los elementos de red que tenemos que usar (cables de transmisión, enrutadores, etc).
- Desarrollar otros protocolos: el TCP se usa para transmitir datos. Hay un gran número de nuevos protocolos que estan desarrollado para envio de voz o de video, por ejemplo la telefonia sobre Internet. Como la mayoría del tráfico en la Internet es TCP, hay una presión para que los nuevos protocolos no tomen demasiados recursos y que sea repartido el ancho de banda de la misma manera que TCP lo hace. Por ello, tenemos que obtener ecuaciones simples del rendimiento del TCP.

La mayoría de los métodos de modelisación del TCP usan un enfoque "fluido" (en continuo). Muchos métodos desprecian el tiempo de espera en las colas en la red. Estos métodos dan soluciones más simples porque el RTT permanece constante.

7.2.1 Modelo fluido de TCP con un enrutador congestionado

Consideramos una sola conexión TCP. Suponemos que hay un enrutador que es el más lento y donde la congestión ocurre.

Para analizar este modelo con un enfoque fluido escribimos

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \frac{\mathrm{d}W}{\mathrm{d}ack(t)} \times \frac{\mathrm{d}ack(t)}{\mathrm{d}t} = \frac{\mathrm{d}W}{\mathrm{d}ack} \times thp_{out}$$

donde thp_{out} es el rendimiento de la conexión TCP a la salida del enrutador, y $ack(t)$ es el número de ACKs que han llegado hasta el tiempo t. Tenemos

$$\frac{\mathrm{d}W}{\mathrm{d}ack} = \begin{cases} 1 & \text{si } W < W_{th}, \\ W^{-1} & \text{si } W \geq W_{th}. \end{cases}$$

Suponemos que no hay una cola en el enrutador, y que el RTT es constante. Entonces la velocidad de transmisión de paquetes en el tiempo t es $thp_{in} = thp_{out} = W(t)/RTT$.

Entonces:

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \begin{cases} \frac{W}{RTT} & \text{si } W < W_{th}, \\ \frac{1}{RTT} & \text{si } W \geq W_{th}. \end{cases}$$

Entonces la evolución de $W(t)$ para $W < W_{th}$ es la solución de

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \frac{W}{RTT}, \qquad W(0) = 1,$$

que es

$$W(t) = \exp\left(\frac{t}{RTT}\right).$$

Para $W \geq W_{th}$ la solúcion es

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \frac{1}{RTT}, W(t_0) = W_{th}$$

que es

$$W(t) = \frac{t - t_0}{RTT} + W_{th}. \tag{7.1}$$

Suponemos que la velocidad maxima de transmisión de paquetes del enrutador es μ. Habra pérdidas cuando

$$\frac{W(t)}{RTT} \geq \mu.$$

Entonces, el tamaño de la ventana cuando hay pérdidas es

$$W_{\max} = RTT\mu,$$

y $W_{th} = RTT\mu/2$.

La cantitad $RTT\mu$ corresponde al número de paquetes que pueden estar en la red antes de tener una pérdida. Entonces, la red tiene una capacidad de contener paquetes a pesar que no haya colas!

Detección por ACKs duplicados en Reno y New-Reno

Si todas las pérdidas son detectadas por ACKs duplicados, entonces habrá un régimen periódico donde la ventana crece de W_{th} hasta $W_{\max}$ de una manera lineal y regresa (con un salto) a W_{th}. El período dura

$$T = \frac{RTT\mu}{2RTT^{-1}} = \frac{RTT^2\mu}{2}$$

El tamaño promedio de la ventana es

$$\overline{W} = \frac{\int_0^T (W(t))dt}{T} = \frac{1}{T}\left(\frac{T^2}{2RTT} + W_{th}T\right) = \frac{T}{2RTT} + W_{th} = \frac{3RTT\mu}{4}.$$

El rendimiento promedio es

$$\overline{thp} = \frac{\overline{W}}{RTT} = \frac{3}{4}\mu.$$

Vemos que no depende de RTT!

Detección para Timeout, o la versión Tahoe
Si la versión Tahoe del TCP es usada, o si las pérdidas causan Time-out, entonces habrá un régimen periódico con una hase de Slow-Start y una de Congestion-Avoidance.

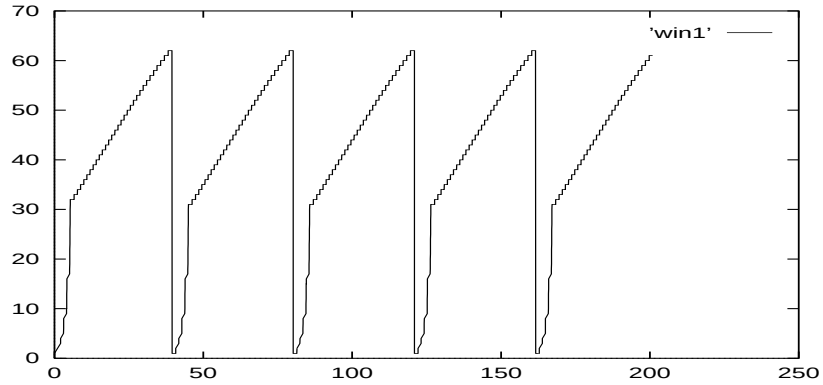

Figure 7.1: Simulación de la evolución de la ventana del TCP en Tahoe

Sean la duración de Slow-Start es T_{ss}, y la duración de Congestion-Avoidance es $T_{ca} = RTT^2\mu/2$. Entonces

$$W_{th} = \exp\left(\frac{T_{ss}}{RTT}\right),$$

y

$$T_{ss} = RTT \log W_{th} = RTT \log\left(\frac{RTT\mu}{2}\right).$$

La cuandidad Q de paquetes que transmitimos durante un período es

$$Q = \int_0^{T_{ss}} W(t)dt = \int_0^{T_{ss}} \exp\left(\frac{t}{RTT}\right)dt = RTT\left[\exp\left(\frac{T_{ss}}{RTT}\right) - 1\right] = RTT(\frac{RTT\mu}{2} - 1),$$

Como W_{th} no puede ser inferior a la ventana minima de 1, tenemos $RTT\mu/2 > 1$, y entonces $Q > 0$. Logramos

$$\begin{aligned}\overline{W} &= \frac{Q}{T} = \frac{1}{T_{ss}+T_{ca}}\left(RTT(\frac{RTT\mu}{2} - 1) + \frac{T_{ca}^2}{2RTT} + W_{th}T_{ca}\right) \\ &= \frac{1}{T_{ss}+T_{ca}}\left(RTT(\frac{RTT\mu}{2} - 1) + \frac{3RTT^3\mu^2}{8}\right).\end{aligned}$$

El rendimiento promedio es $\overline{thp} = \overline{W}/RTT$.

Con los mismos métodos que vimos aquí podemos también calcular el modelo cuando RTT no es fijo, y cuando hay otros flujos no-controlados que usan el mismo enrutador. Para más información pueden leer [5].

7.2.2 Modelo fluido de TCP con pérdidas aleatorias independientes

Cuando hay una red grande con muchas conexiones, la contribución de una conexión a la congestión es insignificante. La congestión es causada por el efecto agregado de todas las otras conexiones. Esto se representa como un proceso aleatorio. Sea T_n el tiempo de la n-ésima pérdida y sea $S_n = T_{n+1} - T_n$ el tiempo entre pérdidas de paquetes de una conexión TCP. Suponemos que

- S_n son independientes con la misma distribución. Sea $s = E[S_n]$ y $s^{(2)} = E[S_n^2]$.
- Usamos la versión Reno o New-Reno del TCP, y todos las pérdidas son detectadas por ACKs duplicados.
- RTT es constante.

Así se considera solamente la fase de Congestion-Avoidance donde la ventana crece linealmente. Sea X_n la ventana justo antes la n-ésima pérdida. Entonces

$$X_{n+1} = \nu X_n + \alpha S_n, \tag{7.2}$$

donde $\nu = 1/2$, y donde $\alpha = RTT^{-1}$.

Sea $x = E[X_n]$ y $x^{(2)} = E[X_n^2]$. En el régimen estacionario, x no depende de n. Tomamos la esperanza de (7.2). Logramos: $x = \nu x + \alpha s$, y entonces

$$x = \frac{\alpha s}{1-\nu}.$$

No es el tamaño promedio de la ventana, pero solamente su tamaño antes de las pérdidas.

Ahora, tomamos la esperanza del cuadrado de cada lado de (7.2). Logramos

$$x^{(2)} = \nu^2 x^{(2)} + \alpha^2 s^{(2)} + 2\nu\alpha x s$$

entonces

$$x^{(2)} = \frac{\alpha^2 s^{(2)} + 2\nu\alpha x s}{1-\nu^2} = \frac{\alpha^2}{1-\nu^2}\left(s^{(2)} + \frac{2\nu s^2}{1-\nu}\right)$$

El tamaño promedio de la ventana se logra por

$$\begin{aligned} \overline{W} &= \frac{1}{E[S_1]} E\left[\int_{T_0}^{T_1} X(t)dt\right] = \frac{1}{s} E\left[\int_{T_0}^{T_1} (\nu X_0 + \alpha t)dt\right] \\ &= \frac{1}{s} E[\nu X_0 S_0 + \frac{\alpha}{2} S_0^2] = \alpha\left(\frac{\nu s}{1-\nu} + \frac{s^{(2)}}{2s}\right) \end{aligned} \tag{7.3}$$

Usamos aquí la independencia de X_0 y S_0, que da $E[X_0 S_0] = E[X_0]E[S_0]$. Tenemos esta independencia porque X_0 es una función de S_i con $i < 0$, y no de $i \geq 0$.

Vemos que si s es un constante, entonces $\overline{W}$ crece con $s^{(2)}$.

7.2.3 Modelo fluido de TCP con pérdidas aleatorias generales

Medidas sobre la red han mostrado que en conexiones largas las pérdidas son independientes. Pero en conexiones sobre distancias de algunos kilómetros los tiempos entre pérdidas son dependientes [1]. Aquí mostramos como se generaliza la solución, usamos el método de [1].

(7.2) es una ecuación de diferencia, y su solución es

$$X_n = \alpha \sum_{k=0}^{\infty} \nu^k S_{n-k-1}. \tag{7.4}$$

Como S_n es estacionaria, podemos suponer que X_n es estacionaria tambien. Tomando la esperanza, podemos ver que la solución de $E[X_n]$ no cambia.

Sin embargo, el cálculo de $\bar{W}$ cambia porque ahora $E[X_0 S_0] \neq E[X_0]E[S_0]$. Podemos usar de nuevo (7.3) donde solo $E[X_0 S_0]$ cambia. Tenemos de (7.4):

$$E[X_0 S_0] = \alpha \sum_{k=0}^{\infty} \nu^k E[S_{-k-1} S_0] = \alpha \sum_{k=0}^{\infty} \nu^k R(k+1)$$

donde definimos $R(k) = E[X_n X_{n+k}]$. $R(k)$ no depende de n porque X_n es estacionaria. Usando (7.3) con esta expresión, logramos

$$\overline{W} = \frac{\alpha}{s}\left[\frac{R(0)}{2} + \sum_{k=1}^{\infty} \nu^k R(k)\right].$$

La probabilidad de perder un paquete es

$$p = \frac{RTT}{s\overline{W}} = \frac{1}{\alpha s \overline{W}}$$

porque la tasa de pérdidas es $1/s$ y el rendimiento es $\overline{W}/RTT$. Denotamos las correlaciones normalidadas $\hat{R}(k) = R(k)/s^2$. Entonces

$$\overline{W} = \alpha s \left[\frac{\hat{R}(0)}{2} + \sum_{k=1}^{\infty} \nu^k \hat{R}(k)\right] = \frac{1}{p\overline{W}}\left[\frac{\hat{R}(0)}{2} + \sum_{k=1}^{\infty} \nu^k \hat{R}(k)\right].$$

Entonces,

$$\overline{W} = \frac{1}{\sqrt{p}}\sqrt{\frac{\hat{R}(0)}{2} + \sum_{k=1}^{\infty} \nu^k \hat{R}(k)},$$

y el rendimiento es

$$Thp = \frac{\overline{W}}{RTT} = \frac{1}{RTT\sqrt{p}}\sqrt{\frac{\hat{R}(0)}{2} + \sum_{k=1}^{\infty} \nu^k \hat{R}(k)},$$

Vemos que

- El rendimiento promedio de TCP es inversamente proporcional al RTT.
- El rendimiento promedio de TCP es inversamente proporcional al $\sqrt{p}$.

Notamos la función de covarianza normalizada:

$$\hat{C}(k) = \hat{R}(k) - 1 = \frac{R(k) - s^2}{s^2}.$$

Entonces

$$Thp = \frac{\overline{W}}{RTT} = \frac{1}{RTT\sqrt{p}}\sqrt{\frac{1+\nu}{2(1-\nu)} + \frac{\hat{C}(0)}{2} + \sum_{k=1}^{\infty} \nu^k \hat{C}(k)}.$$

Si S_n son independientes entonces $\hat{C}_k = 0$ para todo $k \neq 0$, entonces

$$Thp = \frac{\overline{W}}{RTT} = \frac{1}{RTT\sqrt{p}}\sqrt{\frac{1+\nu}{2(1-\nu)} + \frac{\hat{C}(0)}{2}}.$$

Si S_n son constantes entonces $\hat{C}_k = 0$ para todo k.

7.2.4 Modelo de red

El último modelo ha representado toda la red por un proceso de pérdidas que tiene una conexión. Ahora preguntamos la pregunta si dado una arquitectura de red, es posible predecir los procesos de pérdidas y el rendimiento. Vamos a usar un enfoque de punto fijo [2] para hacerlo.

Sea $G = (V, L)$ una red donde V son los nodos y L los enlaces. Hay un conjunto I de clases de conexiones de TCP.

- Una conexión de clase i tiene la fuente S_i y el destino D_i, y
- un camino $\{v_1^i, ..., v_{n(i)}\}$, donde
  v_1^i es el primero nodo despues la fuente de i, y
  $v_{n(i)}$ es el último nodo antes el destino D_i.
  Sea $\pi_i(u) = \{v_1^i, ..., v_{n(i)}^i\}$ el camino desde la fuente S_i hasta u.
- Sea $M = \{\mu_1, ..., \mu_{|V|}\}$ el vector de capacidades donde μ_v es la capacidad del nodo v.
- Sea $\Gamma = \{\gamma_{iv}, i \in I, v \in V\}$ una matriz donde $\gamma_{iv} = 1$ si conexiónes de clase i pasan por el nodo v, y zero sino.
- Sea $\mathbf{p} = (p_1, ..., p_{|V|})$ el vector de probabilidades de pérdidas; p_v corresponde a la probabilidad que un paquete sea perdida en el nodo v.
- Suponemos que las pérdidas en varios nodos son independientes. Entonces la probabilidad de pérdida de un paquete de la conexión i es dado por

$$\kappa_i = \sum_{v \in \pi_i} p_v \prod_{u \in \pi_i(v) \setminus v} (1 - p_u). \tag{7.5}$$

- $T = (T_1, ..., T_{|I|})$ es el vector de tasas de transmisión, donde T_i es la tasa de transmisión de sesión de tipo i.
- N_i, $i \in I$ es el número de conexiones de clase i. Sea $[NT] = (N_1T_1, ..., N_{|I|}T_{|I|})$ el vector donde N_iT_i es la suma de las tasas de transmisiones de conexiones de clase i.

Ahora, tenemos las restricciones de capacidad:

$$\sum_{i \in I} \gamma_{iv} \left(\prod_{u \in \pi_i(v)} (1 - p_u) \right) N_i T_i(\kappa_i) \leq \mu_v, \quad v \in V. \tag{7.6}$$

Entonces, tenemos $|V|$ desigualdades con $|V|$ variables para determinar: $p_1, ..., p_{|V|}$. El conjunto de soluciones de (7.6) no es nulo. Por ejemplo, $p_v = 1, \forall v \in V$ es una solución.

Pero sabemos también (de experimentos y simulaciones) que hay una probabilidad significante de pérdidas en un nodo si y solo si este nodo esta congestionado: la tasa de transmisión en este nodo es igual a su capacidad. Entonces logramos la relaciones

$$p_v \left(\mu_v - \sum_{i \in I} \gamma_{iv} \left(\prod_{u \in \pi_i(v)} (1 - p_u) \right) N_i T_i(\kappa_i) \right) = 0, \tag{7.7}$$

para todo $v \in V$. Además, tenemos

$$0 \leq p_v \leq 1, \quad v \in V. \tag{7.8}$$

Llamamos a (7.6)-(7.8) PC (Problema de Complementaridad).

Sea

$$\Delta_v = \mu_v - \sum_{i \in I} \gamma_{iv} (\prod_{u \in \pi_i(v)} (1 - p_u)) N_i T_i(\mathbf{p}))$$

Lemma 7.2.1 *PC es equivalent al problema PF (punto fijo) siguiente:*

$$p_v = Pr_{[0,1]} \Big\{ p_v - \alpha \Delta_v \Big\}, \tag{7.9}$$

donde $\alpha > 0$ y $Pr_{[0,1]}\{x\}$ es la proyección sobre el intervalo $[0, 1]$:

$$Pr_{[0,1]}\{x\} = \begin{cases} 0, & x < 0, \\ x, & 0 \leq x \leq 1, \\ 1, & 1 < x. \end{cases}$$

Prueba Sea $\mathbf{p}$ una solución de PC, y escogemos $v \in V$ cualquiera. Notamos que $\Delta_v \geq 0$. Si la desigualdad (7.6) es estricta, entonces $p_v = 0$. Entonces tenemos

$$0 = \mathrm{Pr}_{[0,1]} \Big\{ 0 - \alpha \Delta_v \Big\},$$

y logramos (7.9). Si $p_v > 0$ entonces $\Delta_v = 0$ y logramos (7.9) (gracias a (7.8)).

Ahora sea $\mathbf{p}$ una solución de PF, y escogemos $v \in V$ cualquiera. Logramos (7.6) de la proyección Pr. Mostramos la desigualdad (7.6). Suponemos al contrario que $\Delta_v < 0$. Entonces vemos de (7.9) que $p_v = 1$. Pero si $p_v = 1$, entonces $\Delta_v = \mu_v > 0$, y logramos una contradicción. Mostramos la relación (7.7). Tenemos que mostrar que no podemos tener simultanamente $p_v > 0$ y $\Delta_v > 0$. Suponemos lo contrario. La desigualdad $\Delta_v > 0$ implica $p_v - \alpha \Delta_v < 1$. Entonces, obtenemos de (7.9),

$$p_v = p_v - \alpha \Delta_v,$$

que es una contracicción. Entonces (7.7) se satisface también.

Teorema 7.2.1 *El modelo de la red de TCP (7.6), (7.7) y (7.8) tiene una solución si las funciones $T_i(\kappa_i)$ son continuas.*

La prueba sigue el Teorema de punto fijo de Brower.

En varias topologias, es posible mostrar que la solución de (7.6), (7.7) y (7.8) es única. Además, podemos usar iteraciones

$$p_v^{(k+1)} = \mathrm{Pr}_{[0,1]} \Big\{ p_v^{(k)} - \alpha \Delta_v(\mathbf{p}^{(k)}) \Big\}$$

$v \in V$ para calcular soluciones.

7.3 Ejercicios

En el modelo de pérdidas aleatorias de TCP de la sección 7.2.2, calcular el tamaño promedio de la ventana cuando S_n (i) tiene la distribución exponencial con parámetro λ, (ii) es constante, (iii) S_n tiene la distribución uniforme (a, b).

Capítulo 8

Soluciones de Ejercicios del Curso de Teoria de Colas

8.1 Capítulo 2

8.2 Capítulo 3

(3) Usamos la desigualdad de Jensen para obtener:

$$P(bloqueo) = T^*(\mu) = E[\exp(-\mu T_n)] \geq \exp(-\mu E[T_n])] = \exp(-\mu/\lambda)].$$

Ahora, cuando $T_n = 1/\lambda$ por todo n, tenemos igualdad, y logramos el Entonces el mínimo de $P(bloqueo)$. Por el rendimiento tenemos:

$$Thp = \lambda(1 - T^*(\mu)) \leq \lambda(1 - \exp(-\mu/\lambda)])$$

con igualdad cuando $T_n = 1/\lambda$ por todo n.

(4) $N(t)$ tiene la distribución de la suma de k copias independientes de Y definido en Ejemplo 3.4.1. Entonces,

$$P(Z = n) = \sum_{y_1+\ldots+y_n=k} \frac{k!}{y_1!y_2!\ldots y_n!} p^n (1-p)^k.$$

8.3 Capítulo 4

8.4 Capítulo 5

Cálculo de Esperanzas de Q_n
Usamos aqui el método de ocupación de cola que se base sobre el proceso Q_n.

Recordamos que $E[N(T)] = \lambda E[T]$, y que $E\Big[\sum_{j=1}^{N} B_j\Big] = bE[N]$. Entonces

$$E[Q_{n+1}] = E[N(\mathcal{B}(Q_n)) + D_n)] = \lambda E[\mathcal{B}(Q_n) + D_n)] = \lambda(bE[Q_n] + d).$$

En el estado estacionario, $E[Q_{n+1}] = E[Q_n] =: \overline{Q}$ tenemos:

$$\overline{Q} = \lambda(b\overline{Q} + d).$$

Entonces

$$\overline{Q} = \lambda \frac{d}{1-\rho}.$$

Vemos que logramos la misma espresión que en (5.2), porque en el régimen "gated", N_C tiene la misma distribución que Q_n: los clientes que encontramos al principio del ciclo n son los clientes que habian llegado durante el ciclo $n-1$!

Cálculo del segundo momento de Q_n
Recordamos de (3.4) que

$$E([\mathcal{B}(Q_n)]^2) = (E[Q_n^2] - E[Q_n])b^2 + E[Q_n]b^{(2)}.$$

De (3.5) vemos que

$$E[N(\mathcal{B}(Q_n)))^2] = \lambda^2 E([\mathcal{B}(Q_n)]^2) + \lambda E[\mathcal{B}(Q_n)],$$

y también

$$E[N(D_n)^2] = \lambda^2 d^{(2)} + \lambda d.$$

Entonces,

$$\begin{aligned} E[Q_{n+1}^2] &= E[N(\mathcal{B}(Q_n)) + D_n)^2] \\ &= \lambda^2 E([\mathcal{B}(Q_n)]^2) + \lambda E[\mathcal{B}(Q_n)] + \lambda^2 d^{(2)} + \lambda d + 2\lambda^2 db E[Q_n] \\ &= \lambda^2 \Big[(E[Q_n^2] - E[Q_n])b^2 + E[Q_n]b^{(2)} \Big] + \lambda b E[Q_n] + \lambda^2 d^{(2)} + \lambda d + 2\lambda^2 db E[Q_n]. \end{aligned}$$

Como $E[Q_{n+1}] = E[Q_n] = \overline{Q} = \lambda d/(1-\rho)$ donde $\rho = \lambda b$, logramos

$$\begin{aligned} \overline{Q^{(2)}}(1-\rho^2) &= \overline{Q}(\lambda^2(b^{(2)} - b^2) + \rho + 2\rho\lambda d) + \lambda^2 d^{(2)} + \lambda d \\ &= (1+\rho)\lambda d + \frac{\lambda^3 d b^{(2)} + 2\rho\lambda^2 d^2}{1-\rho} + \lambda^2 d^{(2)} \end{aligned}$$

Entonces,

$$\overline{Q^{(2)}} = \frac{\lambda d}{1-\rho} + \frac{\lambda^3 d b^{(2)} + 2\rho\lambda^2 d^2}{(1-\rho)^2(1+\rho)} + \frac{\lambda^2 d^{(2)}}{(1+\rho)(1-\rho)}$$

Indice

Bibliografia

[1] E. Altman, K. Avrachenkov, and C. Barakat. A stochastic model of TCP/IP with stationary random losses. In *SIGCOMM*, pages 231–242, 2000.

[2] E. Altman, K. Avrachenkov, and C. Barakat. TCP network calculus: The case of large delay-bandwidth product. In *borrador*, 2001.

[3] E. Altman, K. Avratchenkov, C. Barakat, and R. Nunez-Queija. State-dependent M/G/1 type queueing analysis for congestion control in data networks. In *Proceedings of the IEEE Infocom 2001 Conference*, Anchorage, Alaska USA, April 2001.

[4] E. Altman, K. Avratchenkov, C. Barakat, and R. Nunez-Queija. TCP modeling in the presence of nonlinear window growth. In *Proceedings of ITC-17*, Salvador da Bahia, Brazil, Dec. 2001.

[5] E. Altman, J. Bolot, P. Nain, D. Elouadghiri, M. Erramdani, P. Brown, and D. Collange. Performance modeling of TCP/IP in a wide-area network. In *IEEE Conference on Decision and Control*, Dec. 1995.

[6] F. Baccelli and P. Bremaud. *Elements of Queueing Theory.* Springer-Verlag, New York, 1994.

[7] D. Bertsekas and R. Gallager. *Data Networks.* Prentice Hall, Englewood Cliffs, New Jersey, 1987.

[8] D.P. Bertsekas and R.G. Gallager. *Data Networks.* Prentice-Hall, 1987.

[9] P. Burke. The output of a queueing system. *Operations Research*, 4:699–704, 1956.

[10] J. R. Jackson. Networks of waiting lines. *Operations Research*, 5:518–521, 1957.

[11] J. R. Jackson. Jobshop-like queueing systems. *Management Science*, 10:131–142, 1963.

[12] V. Jacobson. Congestion avoidance and control. In *ACM SIGCOMM 88*, pages 273–288, 1988.

[13] L. Kleinrock. *Queueing systems.* John Wiley, New York, 1976.

[14] R. Rom and M. Sidi. *Multiple Access Protocols.* Springer-Verlag, 1990.

[15] S. Stidham. A last word on $L = \lambda E[W]$. *Operations Research*, 22:417–421, 1974.