Teoría de colas I

Prof. José Niño Mora

Investigación Operativa, Grado en Estadística y Empresa, 2011/12

Esquema

- Teoría de colas
- Ejemplo: un centro de atención telefónica (call center)
- Tasa de llegada y de servicio, # de servidores
- Factor de utilización y estabilidad
- Procesos estocásticos de interés
- Medidas de rendimiento
- La ley de Little

Teoría de colas (TC)

- Estudia **sistemas de flujo**, donde **clientes**, que llegan a lo largo del tiempo, requieren acceso a **recursos** de **servicio**

- La **capacidad limitada** de éstos causa: **congestión**, **retrasos**, y/o **pérdida** de clientes

- En TC se calculan **medidas de rendimiento**, p. ej. el **número medio de clientes en el sistema**, el **tiempo medio de espera**, o la **proporción de clientes perdidos**

- La TC ayuda a **diseñar** sistemas con mejor rendimiento

- Es una herramienta importante para el diseño y análisis de **redes de ordenadores**, sistemas de telecomunicación, centros de atención telefónica (*call centers*), etc.

Ej: Un centro de atención telefónica

- Consideremos un modelo de un centro de atención telefónica a clientes (*call center*)

- **Parámetros** del modelo:

  - K: **# de servidores**/operadores (ej: 10 servidores)

  - λ: **tasa de llegada** de clientes (ej: 5 clientes/minuto)

  - μ: **tasa de servicio** de cada servidor (ej: 1 cliente/minuto)

Factor de utilización

- La **carga media de trabajo por cliente** que llega al sistema es el tiempo medio de servicio: $1/\mu$

- El # medio de clientes que llegan al sistema por unidad de tiempo es la **tasa de llegada**: λ

- Por tanto, la **carga media de trabajo que llega al sistema por unidad de tiempo** es: λ/μ

- La **capacidad del sistema** es la **carga máxima de trabajo que puede procesar por unidad de tiempo**: K

- Definimos el **factor de utilización** del sistema:

$$\rho \triangleq \frac{\text{tasa media de llegada de trabajo}}{\text{capacidad del sistema}} = \frac{\lambda/\mu}{K} = \frac{\lambda}{K\mu}$$

Estabilidad

- Intuitivamente, el sistema debe tener capacidad suficiente para procesar la carga de trabajo que llega: cuando esto ocurre, el sistema es **estable**

- Si el sistema no tiene capacidad suficiente, la congestión aumentará indefinidamente: el sistema es **inestable**

- **Condición de estabilidad**: el sistema es estable si y sólo si $\rho < 1$

- Observación: el sistema es inestable en el caso $\rho = 1$

- Ej: Si $\rho = 0.9$ decimos que estamos utilizando el 90% de la capacidad

- Para que el sistema sea estable, hemos de utilizar menos del 100% de su capacidad

Ej: Dinámica del modelo

- t_n^{lleg}: **tiempo de llegada** del cliente $n = 1, 2, \ldots$

- t_n^{sal}: **tiempo de salida** del cliente $n = 1, 2, \ldots$

- $\tau_n = t_n^{\text{lleg}} - t_{n-1}^{\text{lleg}}$: **tiempo entre las llegadas** $n-1$ y n. $\{\tau_n\}$ es una sucesión de **variables aleatorias** i.i.d. con

  $$\tau_n \sim \text{Exp}(\lambda) : \text{P}\{\tau_n \leq t\} = 1 - e^{-\lambda t}, \text{E}[\tau_n] = 1/\lambda, \text{Var}[\tau_n] = 1/\lambda$$

- ξ_n: **tiempo de servicio** del cliente n. $\{\xi_n\}$ es una sucesión de **variables aleatorias** i.i.d. con

  $$\xi_n \sim \text{Exp}(\mu) : \text{P}\{\xi_n \leq t\} = 1 - e^{-\mu t}, \text{E}[\tau_n] = 1/\mu, \text{Var}[\tau_n] = 1/\mu$$

Ej: Relación en el caso $K = 1$ bajo FIFO

• FIFO: First-In First-Out (disciplina de servicio más utilizada)

Bajo FIFO, en el caso de $K = 1$ servidor, tenemos la relación:

$$t_n^{\text{sal}} = \max(t_n^{\text{lleg}}, t_{n-1}^{\text{sal}}) + \xi_n$$

Ej: Procesos estocásticos de interés

- $L(t) =$ **# de clientes (llamadas) en el sistema** (en espera o en servicio) en el instante $t \geq 0$

- $Q(t) =$ **# de clientes en espera** en el instante t

- $B(t) =$ **# de clientes en servicio** en el instante t

- Relación: $L(t) = Q(t) + B(t)$

- Para cada **realización**, $L(t), Q(t), B(t) \geq 0$ son funciones escalonadas, con saltos de +/- 1, continuas por la izquierda

Ej: Más procesos estocásticos de interés

- $S_n = t_n^{\text{sal}} - t_n^{\text{lleg}} =$ **tiempo total en el sistema** del cliente n

- $W_n =$ **tiempo total de espera** del cliente n

- Relación: $S_n = W_n + \xi_n$

Ej: Medidas de rendimiento

- $\bar{L} \triangleq \mathsf{E}[L] = \lim_{T\to\infty} \frac{1}{T} \int_0^T L(t)\,dt$: **# medio de clientes en el sistema**
- $\bar{Q} \triangleq \mathsf{E}[Q] = \lim_{T\to\infty} \frac{1}{T} \int_0^T Q(t)\,dt$: **# medio de clientes en espera**
- $\bar{B} \triangleq \mathsf{E}[B] = \lim_{T\to\infty} \frac{1}{T} \int_0^T B(t)\,dt$: **# medio de clientes en servicio**
- Relación: $\bar{L} = \bar{Q} + \bar{B}$

Ej: Más medidas de rendimiento

- $\bar{S} \triangleq \mathsf{E}[S] = \lim_{n\to\infty} \frac{1}{n}(S_1 + \cdots + S_n)$: **tiempo medio en el sistema** por cliente

- $\bar{W} \triangleq \mathsf{E}[W] = \lim_{n\to\infty} \frac{1}{n}(W_1 + \cdots + W_n)$: **tiempo medio de espera** por cliente

Ej: Probabilidades

- $p_k \triangleq \mathsf{P}\{L = k\} = \lim_{T\to\infty} \frac{1}{T} \int_0^T 1_{\{L(t)=K\}}\, dt$: **probabilidad de que haya k clientes en el sistema**

- $\sum_{k=0}^{\infty} p_k = 1$

- La probabilidad de que todos los servidores estén ocupados es: $\sum_{k=K}^{\infty} p_k = 1 - \sum_{k=0}^{K-1} p_k$

- Relación: $\bar{L} = \sum_{k=0}^{\infty} k p_k$

Aclaración: Medias "a largo plazo"

- Dado un # inicial de clientes, $L(0) = i$, la evolución del proceso $\{L(t) : t \geq 0\}$ depende de i
- El **# medio de clientes en el instante** t es: $\mathsf{E}[L(t) \mid L(0) = i]$ (depende de $L(0) = i$ y de t)
- Pero: si el sistema es **estable** ($\rho < 1$), entonces

$$\lim_{t\to\infty} \mathsf{E}[L(t) \mid L(0) = i] = \bar{L}$$
$$\lim_{t\to\infty} \mathsf{E}[Q(t) \mid L(0) = i] = \bar{Q}$$
$$\lim_{t\to\infty} \mathsf{E}[B(t) \mid L(0) = i] = \bar{B}$$
$$\lim_{n\to\infty} \mathsf{E}[W_n \mid L(0) = i] = \bar{W}$$
$$\lim_{n\to\infty} \mathsf{E}[S_n \mid L(0) = i] = \bar{S}$$

La ley de Little

- Relaciona tres medidas de rendimiento en un sistema estable: el **# medio en el sistema** $\bar{L}$, la **tasa de llegadas** λ, y el **tiempo medio en el sistema** $\bar{S}$

- **Ley de Little**: $\bar{L} = \lambda \bar{S}$

- La ley de Little nos permite calcular $\bar{S}$ conociendo λ y $\bar{L}$

- También nos permite comprobar la validez de datos registrados sobre $\bar{L}$, λ y $\bar{S}$

La ley de Little

- Podemos aplicar la ley de Little de forma flexible, variando lo que consideramos como "sistema"

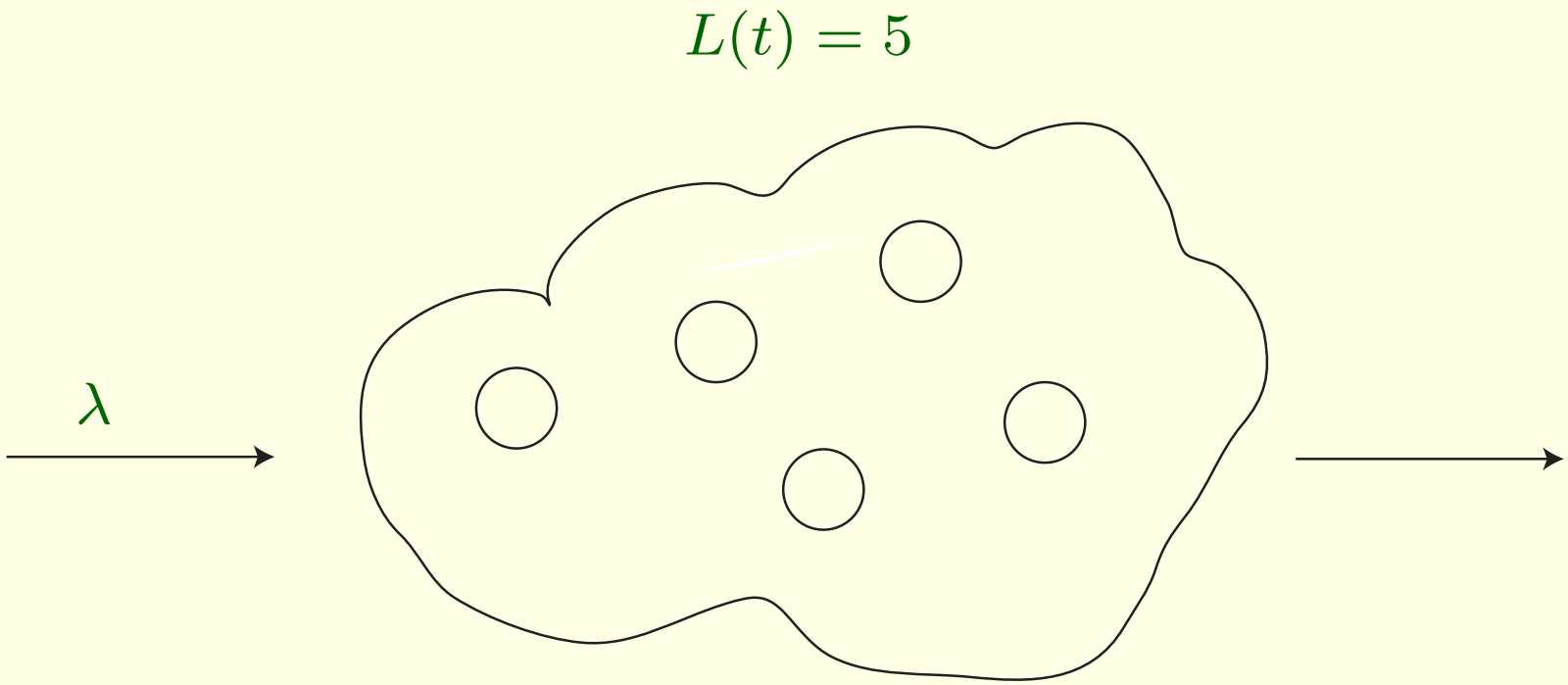
- Así, si el "sistema" se refiere sólo a los clientes en espera, obtenemos la relación: $\bar{Q} = \lambda \bar{W}$

- Además, si el "sistema" se refiere sólo a los clientes en servicio, obtenemos: $\bar{B} = \lambda \frac{1}{\mu} = K\rho$

- Así, hemos obtenido la relación $\rho = \bar{B}/K$, que nos dice que el **factor de utilización** ρ es la **proporción media de servidores ocupados**

Validez general de la ley de Little

• La ley de Little es válida para cualquier sistema de colas estable: el # medio de clientes en el sistema (dentro del borde en la figura) es igual a la tasa de llegada (que es igual que la de salida) multiplicada por el tiempo medio que un cliente permanece en el sistema



Aplicando la ley de Little

- La ley de Little se puede aplicar a numerosas situaciones
- Ej: Vamos a un local de moda con capacidad para 60 personas, que suele estar lleno; nos dicen que el tiempo medio de estancia es de 3 horas. Supongamos que al llegar encontramos una cola para entrar con 19 personas. ¿Cómo podemos estimar nuestro tiempo de espera?
- Aplicamos la ley de Little al sistema que consiste en el interior del local (no contamos la cola de espera para entrar)
- Sabemos que: $\bar{L} = 60, \bar{S} = 3$; por la ley de Little, la tasa de llegada de clientes es de $\lambda = \bar{L}/\bar{S} = 20$ clientes por hora, igual que la tasa de salida
- Como al llegar encontramos 19 clientes en espera, entraremos cuando hayan salido 20 clientes, lo que ocurrirá, en promedio, dentro de 1 hora

La ley de Little y sistemas informáticos

- Consideremos un sistema compuesto por n usuarios, conectados a un sistema informático con un **tiempo medio de respuesta** de r minutos

- Cada usuario alterna entre dos estados: preparar una tarea informática, lo cual lleva un tiempo medio de z minutos; y, tras completarla y enviarla al sistema informático, esperar la respuesta de éste

- Los usuarios envían al sistema informático, en promedio λ tareas por minuto

- Pregunta: Si conocemos n, z, λ, ¿cómo podemos estimar r?

La ley de Little y sistemas informáticos

- Para aplicar la ley de Little, consideramos que los "clientes" son las "tareas", en preparación o en ejecución

- Por tanto, el # de "clientes" en el sistema permanece constante: $L(t) \equiv n \Longrightarrow \bar{L} = n$

- Imaginamos que, tras ser ejecutadas, las tareas retornan inmediatamente a los usuarios, de manera que éstos comienzan a preparar una nueva tarea

- Por tanto, el tiempo medio que una tarea permanece en el sistema es: $\bar{S} = z + r$

- Por la ley de Little: $n = \lambda(z + r) \Longrightarrow r = n/\lambda - z$

La ley de Little: Justificación

- Una empresa gestiona el sistema (p. ej. un aparcamiento)
- Tarifa: 1 € por cliente y hora en el sistema
- Método de cobro: se registra la hora de llegada y la de salida de cada cliente; cuando sale se le cobra la diferencia entre ambas en Euros
- Si llegan λ clientes por hora, la empresa tarifa, en promedio, $\lambda \bar{S}$ € por hora
- Para comprobar la validez de sus datos, la empresa implementa un sistema de cálculo equivalente: registra, para cada tiempo t, el # $L(t)$ de clientes en el sistema, y calcula su promedio: $\bar{L}$
- Así, la empresa tarifa, en promedio, $\bar{L}$ € por hora
- Por tanto: $\bar{L} = \lambda \bar{S}$

Ej: un sistema de colas con $\rho > 1$

- Si un sistema de colas tiene $\rho \approx 1$ o $\rho > 1$ decimos que está en un régimen de **tráfico pesado**

- Hemos visto que el caso $\rho \geq 1$ es inestable

- Datos de noticia en *El País*, 28-11-2005: "El atasco judicial se perpetúa, con más de dos millones de asuntos en trámite"

- "Se estima que ingresarán durante 2005 cerca de 70.000 asuntos más de los que se resuelven"

Cómo calcular $\bar{L}$

- Para calcular $\bar{L}$:

  1. Calcularemos p_k, para $k = 0, 1, 2, \ldots$

  2. Calcularemos $\bar{L}$ mediante la relación

$$\bar{L} = \sum_{k=0}^{\infty} k\, p_k$$