Introdução à rede de filas

Marcos N. Magalhães

Departamento de Estatística
Instituto de Matemática e Estatística
Universidade de São Paulo

Prefácio

Este livro pretende apresentar as ídéias básicas de teoria das filas. A intenção é oferecer um texto que possa ser informativo dos principais resultados da área e ilustrativo de algumas técnicas de probabilidade aplicada que são utilizadas em filas. Acreditamos que o livro possa ser acessível a leitores com formação equivalente a de um curso inicial de processos estocásticos na pós-graduação. Estudantes de graduação com pelo menos dois semestres de processos estocásticos e razoável formação matemática também deverão acompanhar o texto sem maiores dificuldades. No sentido de auxiliar a compreensão do material apresentado e preencher alguma lacuna de formação, os apêndices incluem um resumo dos principais conceitos utilizados no texto.

No capítulo 1, introduzimos o conceito de fila e motivamos o leitor através da apresentação de exemplos em sistemas de manufatura, comunicação de dados e prestação de serviços. Apresentamos ainda as características que definem os diversos modelos e as possíveis perguntas de interesse. O material exposto nesse capítulo tem o caráter informativo e ilustrativo do assunto.

Nos capítulos 2 e 3, descrevemos os modelos básicos de filas. Há um vasto material a ser coberto e o texto não teve a pretensão de ser exaustivo. Além dos modelos clássicos, procuramos apresentar outros que possuem uma maior motivação teórica ou prática. Em vários modelos, resultados sobre fluxo de clientes são discutidos.

Os modelos de redes de filas, introduzidos no capítulo 4, constituem-se nos mais discutidos e utilizados em teoria das filas. A apresentação inclui a descrição da rede e a distribuição estacionária do número de clientes no sistema. A obtenção de soluções em forma de produto é também comentada e, para alguns modelos, o tráfego na rede é discutido, enfatizando as condições para obter processos de Poisson. Como mencionado, três apêndices resumem definições e resultados que auxiliam na leitura e compreensão do texto. A bibliografia inclue livros publicados recentemente sobre teoria das filas, além das referências mencionadas no texto.

Gostaria de agradecer à diretoria da ABE e à comissão coordenadora do 12° SINAPE pela oportunidade de oferecer o minicurso e redigir esse trabalho. Várias pessoas colaboraram nessa realização. O estudante Henry Corazza do IME-USP e os colegas Maria Creusa B. Salles do ICMSC-USP e Paulo Renato de Morais do INPE leram boa parte dos originais, contribuindo com críticas e

sugestões, que foram de grande valia. As figuras foram feitas por Luís Ricardo Câmara da ADUSP. Francisco Miraglia do IME-USP ajudou a clarear alguns conceitos de lógica utilizados na demonstração de teorema do capítulo 3. Maria Cecília C. Magalhães da PUC-SP ajudou na revisão do português e Mauro Marques da UNICAMP colaborou com a "infraestrutura computacional". A todos eles meu sincero agradecimento. Naturalmente, as falhas e outros enganos que ainda serão encontrados, são de minha inteira responsabilidade.

Finalmente, dedico esse trabalho a todos aqueles que lutam para diminuir as mais vergonhosas filas do nosso país, as filas dos que esperam (e quase nunca conseguem) terra, teto, comida, emprego, educação e saúde. Infelizmente a cidadania não chegou para a maior parte da nossa população.

São Paulo, Março de 1996
Marcos. N. Magalhães

Conteúdo

Lista de figuras e tabelas

Figuras

Tabelas

Lista de teoremas e resultados

Capítulo 1

Introdução ao conceito de fila

1.1 O que é uma fila?

A pergunta acima pode parecer uma provocação ao leitor que, para poder chegar a ler essa frase, deve ter passado por várias filas durante toda sua vida. Atendimentos em bancos e supermercados, "check in" em hotéis e simpósios são alguns exemplos da espera por um serviço. Em resumo, nenhum de nós escapa de uma filinha aqui e ali no nosso dia a dia. Certamente essas experiências produzem em todos sentimentos muito fortes. Que injustiça é maior do que furar uma fila? A visão popular da espera produz suas frases de efeito que são repetidas frequentemente com conteúdos nem sempre equivalentes. "Quem espera sempre alcança", "quem sabe faz a hora não espera acontecer" e "espere sentado porque de pé cansa" são algumas dessas frases, que se tornaram síntese de atitudes do cotidiano brasileiro. Elas representam alternativas de ação frente aos casos em que o que desejamos ou precisamos não está a disposição. Cada um de nós provavelmente terá uma situação de fila que o marcou. Rapidez, demora, desistência e privilégio em filas certamente já foram vividos por todos nós.

Um modelo ou sistema de filas pode ser brevemente descrito da seguinte forma: usuários (ou fregueses ou clientes) chegam para receber um certo serviço e, devido à impossibilidade de atendimento imediato, formam uma fila de espera. A figura 1.1 ilustra essa idéia. Os termos usuário e serviço são usados aqui com um sentido amplo. Podemos estar nos referindo a carros que chegam a um posto de pedágio, máquinas que esperam para serem consertadas, peças que seguem uma linha de montagem ou mensagens que são transmitidas pelos canais de comunicação. Uma rede de filas é formada por várias filas que se interconectam entre si de modo que o usuário ao sair de uma fila pode (com uma certa probabilidade) dirigir-se a outra. Nas redes abertas há fluxo de fregueses entrando e saindo do sistema. Por outro lado, nas redes fechadas o número de usuários permanece inalterado, isto é, não há movimentação de usuários para dentro ou para fora do sistema. Um serviço de manutenção de máquinas pode ser visto como uma rede fechada onde M máquinas se alternam entre os centros de manutenção e de operação. Dessas definições básicas, ramificam-se um sem número de outros

modelos de filas adequados às várias áreas, sempre na busca de melhor representar a realidade.

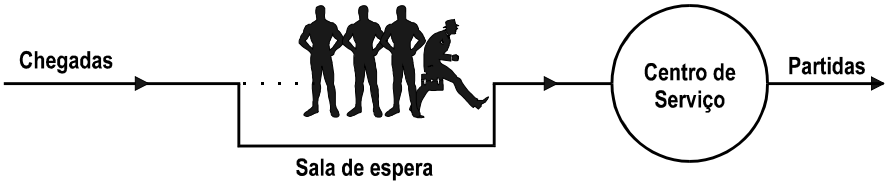

Figura 1.1: Diagrama de uma fila

Em aplicações o estudo dos modelos de filas tem como objetivo a melhoria de desempenho do sistema, entendida, entre outros aspectos, como melhor utilização dos recursos de serviço disponíveis, menor tempo de espera e mais rapidez no atendimento. O pioneiro nesse estudo foi A. K. Erlang que, no começo do século como engenheiro da companhia dinamarquesa de telefones, estudou o problema da congestão de linhas. Telefonia permaneceu a principal aplicação de teoria das filas até por volta de 1950. A partir daí, um grande número de áreas tem utilizado essa ferramenta e a vasta literatura é o melhor indicador dessa expansão. Paralelamente à aplicação, os estudos teóricos têm também se expandido consideravelmente com o auxílio da teoria de probabilidade e processos estocásticos. No momento duas revistas publicam mais enfaticamente artigos em teoria das filas: Queueing Systems-Theory and Applications e Communications on Statistics-Stochastic Models. Entre as revistas internacionais que também publicam com frequência artigos nessa área destacam-se: Operations Research, Management Science, Journal of Applied Probability, Advances in Applied Probability, Naval Research Logistics, IEEE Transactions on Communications e Mathematics of Operations Research. Dentre as revistas nacionais, a Rebrape (ABE) e a Revista de Pesquisa Operacional (SOBRAPO) também publicam artigos teóricos e aplicados em teoria das filas.

No que segue, usaremos os termos em português tentando preservar a idéia e o sentido da palavra de origem. Algumas vezes, apresentaremos em itálico e/ou entre parêntesis o correspondente termo em inglês no sentido de auxiliar a leitura das referências.

1.2 Principais características de uma fila

Algumas das características básicas de uma fila como chegadas, serviço, disciplina de atendimento e capacidade de espera serão descritas a seguir.

a) Chegadas

O processo de chegada (*arrival process*) é a descrição de como os usuários procuram o serviço. Se eles chegam a intervalos fixos de tempo, o processo de chegadas é dito constante ou determinístico. Por outro lado, se as chegadas são aleatórias no tempo, elas formam um processo estocástico e é necessário descrever suas propriedades probabilísticas.

A suposição mais comum é que as chegadas formam um processo de renovação, isto é, os intervalos entre chegadas são independentes e identicamente distribuídos. Em geral, é também assumido a independência em relação ao serviço, mas é possível aplicar um processo de renovação para as chegadas, condicionado à situação do serviço. O processo de Poisson é um processo de renovação com distribuição exponencial e é um dos mais utilizados para modelar as chegadas. Além de descrever, com boa aproximação, diversas situações práticas, os processos de Poisson incorporam facilidades no tratamento matemático proporcionadas pela "falta de memória" da distribuição exponencial. Chamadas telefônicas têm sido modeladas satisfatoriamente dessa forma. As distribuições de Erlang e hiperexponencial são também bastante utilizadas. Modelos mais complicados envolveriam possíveis dependências entre as chegadas como, por exemplo, a situação onde uma chegada de certo tipo aumentaria ou diminuiria a chance de ocorrência de outro tipo de chegada.

As chegadas mencionadas acima podem ser unitárias ou em bloco (*batches*). Neste caso, além do tempo entre chegadas, também o tamanho dos blocos é aleatório. Por exemplo, em aeroportos internacionais, a chegada de passageiros de um certo vôo ao posto alfandegário se dá em bloco, onde o tamanho do bloco é a lotação do avião.

Existem situações em que as chegadas dependem do número de usuários no sistema, podendo até acontecer a situação em que uma chegada não se junta à fila. Isto pode ocorrer por decisão do usuário ou por limitação no espaço para espera. O caso clássico, conhecido como sistema com perda (*loss system*), originou-se do estudo de tráfego telefônico onde o usuário completa a chamada ou obtém o sinal de ocupado e é excluído do sistema. Chegadas com usuários impacientes, que abandonam a fila após alguma espera, podem também ser modeladas.

b) Serviço

Da mesma forma que o processo de chegada, podemos considerar o processo de serviço como sendo determinístico ou aleatório. A distribuição do tempo de serviço pode depender do estado do sistema ou, até mesmo, do tipo do usuário a ser servido. Porém a hipótese mais simples é a de independência, isto é, o serviço é um processo de renovação. Dentre as distribuições mais usadas destacam-se a exponencial, Erlang e hiperexponencial. O número de servidores disponíveis para atendimento a uma mesma fila também deve ser especificado. Nesse caso, é comum mencionar que os servidores estão em paralelo numa referência a estarem atendendo uma mesma fila.

c) Disciplina de atendimento

A disciplina de atendimento se refere à maneira como os usuários serão selecionados para receber serviço. No nosso cotidiano os atendimentos, em geral, se dão pela ordem de chegada. A fila no caixa do supermercado, a retirada de carros do estacionamento e a compra de ingressos para o cinema são exemplos dessa disciplina, que será referida como FCFS (do inglês, *first come first served*). Em aplicações, outras disciplinas podem aparecer. A disciplina LCFS (*last come first served*) pode ser usada em modelos de arquivo ou de busca em discos rígidos. O serviço em ordem aleatória, independente do tempo de chegada, pode servir de modelo para alguns sistemas computacionais. Outra disciplina, que tem aplicação em computação é a do processamento ou tempo compartilhado (*processor or time sharing*) que é definida pela dedicação, a todos os usuários presentes no sistema, de uma pequena quantidade de serviço de cada vez. Assim, em rodadas sucessivas, o usuário vai recebendo sua dose de atendimento até que sua requisição total de serviço seja completada. A disciplina de atendimento pode ainda estabelecer prioridades entre os usuários de modo a atender primeiro os de alta prioridade. Em alguns modelos, o serviço pode até ser interrompido para dar lugar a um usuário de prioridade mais alta. Quando modelos com prioridade são adotados é necessário especificar como se dará a ordem de atendimento dentro da mesma classe de prioridade e, em geral, FCFS é utilizada nesses casos.

d) Capacidade do sistema

É muito comum haver uma limitação física no número de usuários que podem esperar. Se a capacidade total estiver ocupada, o usuário não poderá entrar no sistema e será perdido ou desviado para outro centro de serviço. Essa limitação se relaciona com a chegada mas a decisão de não se juntar à fila não é do usuário e sim do sistema de serviço.

Variando as características *a*-*d* acima, podemos obter um grande número de modelos. O que vamos descrever nesse texto são os modelos básicos sobre os quais poderão ser combinadas e/ou adicionadas restrições e propriedades especiais.

Para descrever uma fila será usada a notação definida por Kendall [1953], que consiste da forma A/B/c/K/Z, onde A descreve a distribuição do tempo entre chegadas, B a distribuição do tempo de serviço, c o número de servidores, K a capacidade da fila de espera (alguns autores definem K como a capacidade total de usuários no sistema) e Z a disciplina de atendimento.

Algumas escolhas para A e B são as seguintes:

M: distribuição exponencial (de *memoryless*)
E_k: distribuição Erlang-k
D: distribuição determinística ou degenerada
U: distribuição uniforme
G: distribuição geral (não especificada)

A omissão de K e Z na representação acima indica que a fila tem capacidade infinita e disciplina FCFS. Por exemplo, a fila M/G/1 tem chegadas exponenciais, serviço com distribuição geral e um servidor, não há limite na sala de espera e o atendimento é em ordem de chegada. Por outro lado, a fila G/E2/3/15 tem chegadas seguindo uma distribuição geral, o serviço segue a distribuição de Erlang-2, existem 3 servidores, a capacidade máxima do sistema é 18 (note que a fila máxima tem comprimento 15) e, como nada foi mencionado, a disciplina de atendimento é FCFS.

1.3 Exemplos e aplicações

No sentido de ilustrar a amplitude das aplicações dos modelos de filas, vamos descrever, brevemente, algumas situações práticas onde eles podem ser utilizados.

Na área comercial são utilizados, entre outros, no atendimento bancário, na passagem pelo caixa de um supermercado, no abastecimento de combustível e em máquinas automáticas de venda de bebidas. Essas são situações onde é natural associar um modelo de filas para estudar o desempenho do sistema.

Em transportes alguns exemplos são o pouso e a decolagem de aviões, carros transitando por ruas e estradas parando aqui e ali em um semáforo ou posto de pedágio, caminhões num terminal de carga esperando sua vez de serem carregados ou descarregados e navios buscando atracação em um porto.

Em sistemas industriais um grande número de sistemas de filas pode ser encontrado. Peças percorrendo uma mesma linha de produção podem ser vistas como um sistema de filas em série. Se diferentes tipos de peças são processados elas podem percorrer diferentes sequências de máquinas até ficarem prontas dando origem assim a redes de filas mais complexas. Uma vez prontas, uma amostra delas é submetida aos inspetores de qualidade e uma nova fila pode ser formada.

Telefonia, redes de computadores e sistemas de comunicação de dados fornecem também vários exemplos de aplicações de sistemas e redes de filas. Mensagens para serem transmitidas de um ponto a outro precisam passar por diversos centros de retransmissão ocasionando atrasos devido a um possível acúmulo de mensagens em um centro intermediário. A CPU (unidade central de processamento) de um computador atende múltiplas tarefas e pode ser vista como um servidor que atende usuários de vários tipos que têm prioridades e tempos de serviço diferentes.

Vamos utilizar os próximos três exemplos para ilustrar algumas quantidades de interesse que podem ser respondidas pela modelagem com filas.

Exemplo 1.1: (Allen [1990])

Uma companhia aérea planeja construir um novo centro telefônico para reserva de passagens. Cada agente terá um terminal de computador e atenderá um cliente típico em 5 minutos. O tempo de serviço é assumido ser exponencial. As chegadas ocorrem aleatoriamente e o sistema tem capacidade ilimitada para permitir que chamadas aguardem a disponibilidade de algum agente. Uma média de 36 chamadas por hora é esperada durante o período de pico do dia. Pretende-se atender três requisitos:

a) A probabilidade de uma chamada encontrar todos os agentes ocupados não poderá ser maior do que 10%;
b) O tempo médio de espera para ser atendido por um agente não excederá 1 minuto.
c) Menos de 5% das chamadas que chegam vão esperar mais que um minuto para serem atendidas.

Quantos agentes deverá ter o centro?

Em linguagem de filas temos um modelo M/M/c (ver figura 1.2) e precisamos determinar o número de servidores c de modo a cumprir os critérios acima, inclusive nos momentos de pico. Note que a aleatoriedade da chegada foi traduzida por assumir que as chegadas formam um processo de Poisson. Os

parâmetros dessas distribuições estão informados e, como veremos nos próximos capítulos, podemos determinar a distribuição estacionária do número de chamadas no sistema num instante arbitrário de tempo (em função de c). Sendo π_n essa distribuição, a condição a) pode ser escrita como:

$$\text{a)} \sum_{n>c} \pi_n \leq 0.1.$$

Para 5 e 6 servidores teremos essa soma de probabilidades igual a 0.236 e 0.09911, respectivamente. Como a somatória dessas probabilidades indica a probabilidade de fila, seu valor decresce com o aumento do número de servidores. Assim, devemos ter pelo menos 6 servidores para satisfazer a condição a).

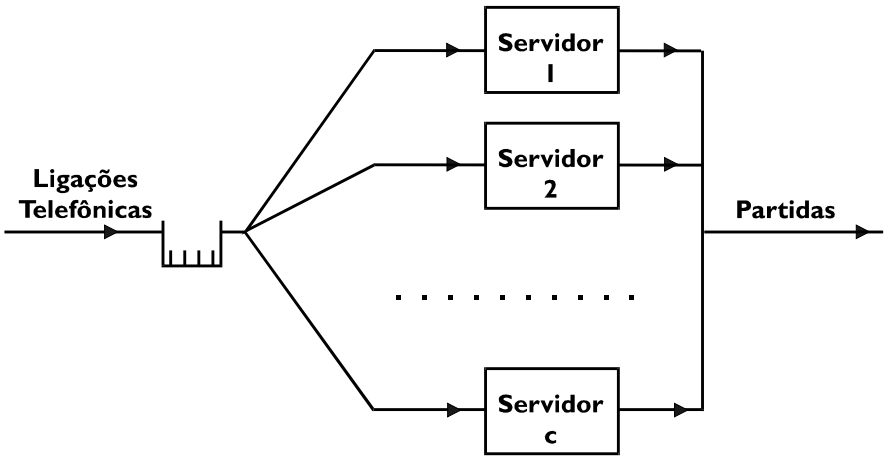

Figura 1.2: Modelagem de uma central de reservas

De π_n pode-se determinar o número médio de chamadas presentes no centro. A fórmula de Little, que será vista mais adiante, relaciona essa média com o tempo médio total gasto no sistema (fila mais atendimento). Como em média o serviço do agente dura 5 minutos, podemos determinar, ainda em função de c, o tempo médio de espera na fila. Seja θ_q o tempo de fila e W_q sua média. As condições b) e c) podem ser reescritas da seguinte forma:

$b)\ W_q \leq 1$ minuto;
$c)\ P(\theta_q > 1) \leq 0.05.$

Para 6 servidores temos $W_q = 1.67$ o que não atende a condição b). É possível verificar que, para 8 agentes, $W_q = 1$ e portanto b) esta satisfeita. Podemos ainda calcular (ver Capítulo 2) a distribuição do tempo de espera na fila, e daí verificar a condição c). Para 8 agentes obtemos $P(\theta_q > 1) = 0.00476$ e assim essa condição também está satisfeita. Concluimos que 8 agentes é a escolha adequada para atender as chamadas, segundo os critérios mencionados. Note que como a escolha foi feita pelo pico de chegadas, alguma folga no atendimento ocorrerá e a performance do sistema deve ser melhor do que os números acima calculados.□

Exemplo 1.2: (Kleinrock [1976])

Um sistema computacional de multiprocessamento é discutido no capítulo 4 de Kleinrock [1976] e consiste em permitir que vários programas (*jobs*) entrem no sistema ao mesmo tempo.

Enquanto a unidade central de processamento (CPU) está ocupada com um *job*, várias unidades auxiliares como impressoras, banco de dados ou leitora de fitas estão atendendo outros programas que já passaram pela CPU. Os *jobs* perfazem ciclos sucessivos entre a CPU e cada unidade auxiliar até completarem seu processamento e deixarem o sistema.

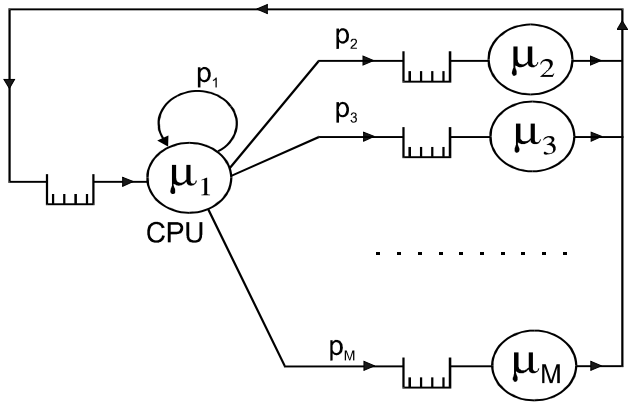

Figura 1.3: Sistema computacional de multiprocessamento

Vamos admitir que há uma limitação no número simultâneo de *jobs* dentro do sistema e uma fila (com ou sem prioridades) regularia a entrada de novos *jobs* no sistema. A modelagem é feita através de uma rede fechada de filas, ver figura 1.3, com k usuários (indicando multiprogramação nível k) e M nós onde o nó 1 é a CPU e os restantes $M-1$ nós são os periféricos. Assumimos que, em todos os nós, a distribuição de tempo de processamento é exponencial.

Nesse modelo assume-se que há um número grande de *jobs* desejando entrar no sistema de modo que, quando uma vaga aparece, ela é imediatamente preenchida. Como na rede fechada não há entrada nem saída de usuários, este aspecto é modelado pelo *feedback* no nó 1. Isto é, com probabilidade p_1, o *job* retorna imediatamente à CPU para um novo serviço (na prática ele é substituído por um novo) e com probabilidade p_j ele se dirige ao nó periférico j $(2 \leq j \leq M)$. Algumas questões de interesse nesse modelo incluem a identificação de gargalos no processamento, nível de utilização dos equipamentos, tempo médio de cada ciclo e fluxo dos *jobs* através do sistema. Pode-se obter, entre outras medidas, a distribuição de probabilidade estacionária dos k *jobs* entre os M nós do sistema e então responder várias perguntas sobre a performance do sistema.□

Exemplo 1.3: Linha de Produção

Um certo sistema de manufatura, apresentado na figura 1.4, consiste de uma sequência de máquinas que processam um único tipo de peça. O serviço em cada máquina é feito em ordem de chegada e não há restrição no tamanho da fila de espera. As máquinas estão sujeitas a quebras com a consequente suspensão temporaria da produção. Existe ainda a possibilidade de alguma peça requerer um novo processamento, por não ter sido aprovada pelo controle de qualidade.

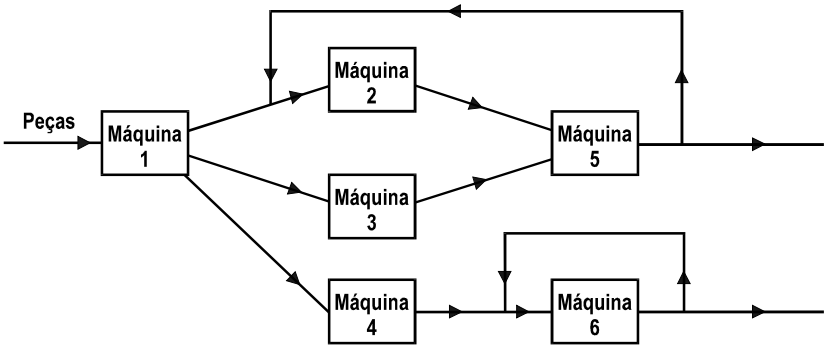

Figura 1.4: Exemplo de uma linha de produção

Deseja-se estudar o comportamento do sistema no sentido de identificar a necessidade de ampliação do número de máquinas bem como as características da saída desse sistema. Note que, em geral, as saídas servirão como entrada para uma outra etapa do processo produtivo. Do ponto de vista de filas, as peças são usuários que passam pelos vários centros recebendo diferentes serviços. O retorno de peças para reprocessamento pode ser modelado através do chamado *feedback*. Em geral, após concluir serviço no centro j, o usuário tem uma certa probabilidade positiva p_j de retornar a j. As quebras de máquinas são consideradas como "férias" (*vacations* em inglês) do servidor e podem ter duração aleatória ou determinística. Dependendo das distribuições de probabilidade envolvidas na chegada, serviços e quebras, a solução analítica desse modelo pode ser bastante díficil. Nesse caso, o uso da técnica de simulação, aproximações ou hipóteses simplificadoras pode ser o caminho para a obtenção das respostas desejadas.□

1.4 Exercícios

1) (Gross & Harris [1985]) Para cada uma das situações abaixo, modele um sistema de filas, indicando o significado das chegadas, serviços e outras características:

a) Pouso de aviões em aeroportos.
b) Caixas de supermercados.
c) Atendimento em postos do correio.
d) Pedágio em rodovias.
e) Posto de gasolina com várias ilhas de atendimento.

f) Centro de lavagem automática de carros.
g) Chamadas telefônicas chegando a um PABX de uma empresa.
h) Pacientes, com hora marcada, chegando num consultório médico.

2) Suponha que uma máquina processa certa peça em um tempo constante de 5 minutos. As peças chegam para processamento em intervalos constantes de 3 minutos. Descreva a situação da máquina no fim do expediente.

3) Na modelagem do atendimento bancário, indique as diferenças entre fila única e o atendimento tradicional em caixas paralelos. O caixa trabalha mais no atendimento de fila única? E quanto ao usuário, quais são os efeitos?

4) Como seria possível modelar, usando teoria das filas, o transporte marítimo de 5 navios entre 3 portos?

5) Num serviço de balsas para automóveis, existe diferença em cobrar a tarifa antes ou depois da travessia? Descreva os modelos de fila correspondentes e faça uma comparação das dificuldades de cada um.

6) Descreva como poderíamos modelar, através de filas, um restaurante com as seguintes características:

a) Rodízio de churrasco (garçon percorre a mesa dos clientes com espetos de diferentes tipos de carne).

b) Buffet de saladas e pratos quentes (clientes se servem sózinhos quantas vezes quiserem).

c) Comida por quilo (os clientes se servem do que desejam comer e pagam pelo peso da comida).

7) Discuta a conveniência de prioridades num serviço de cópias. Suponha que existam dois tipos de trabalho que são solicitados: grandes quantidades de cópias como apostilas, livros e teses e serviços rápidos de poucas páginas. Considere separadamente os casos de uma ou duas máquinas de atendimento e estude o problema levando em conta diversos aspectos como espera pelo serviço e eficiência na utilização das máquinas.

8) Seria possível montar um modelo de filas para a burocracia nas universidades? Estude a questão.

Capítulo 2

A fila M/M/1 e suas variantes

2.1 Introdução

Apresentamos neste capítulo as principais propriedades da fila M/M/1. Essa fila tem chegadas seguindo um processo de Poisson, serviço exponencial e um servidor que atende os usuários em ordem de chegada. Serviços e chegadas são assumidos serem processos independentes. Não há limitação na sala de espera e assim todos os usuários serão atendidos mais cedo ou mais tarde. A figura 2.1 apresenta esquematicamente o funcionamento dessa fila.

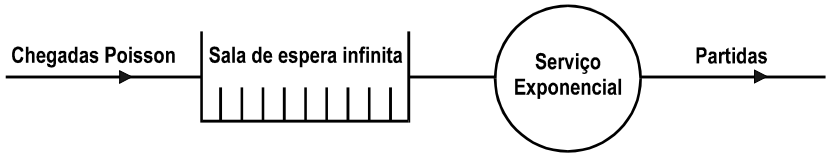

Figura 2.1 Fila M/M/1

Sejam $A(t) = 1 - e^{-\lambda t}$ e $B(t) = 1 - e^{-\mu t}$ as distribuições do intervalo entre chegadas e de duração do serviço, respectivamente. É comum os parâmetros λ e μ serem referidos como taxas, indicando o número médio de usuários que chegam ou são servidos por unidade de tempo.

Denotando por $N(t)$ o número de usuários no sistema no instante t, pode-se demonstrar que $\{N(t); t \geq 0\}$ é um processo de Markov em tempo contínuo com o espaço de estados sendo os inteiros não negativos. $N(t)$ é frequentemente referido como sendo o *estado do sistema* ou *estado da fila.* Na verdade, nesse último caso, abusamos da linguagem para indicar o total de clientes ao invés de considerar somente os que estão na fila propriamente dita.

O tempo de espera para fazer uma transição só depende do estado presente e, se o sistema contém pelo menos um usuário, terá distribuição exponencial de parâmetro $(\lambda + \mu)$ correspondente ao mínimo entre duas

exponenciais independentes (serviço e chegada). No caso de não haver nenhum usuário no sistema, o tempo para uma transição terá distribuição exponencial de parâmetro λ.

O processo de Markov $N(t)$ é de fato um processo de nascimento e morte com taxas independentes do estado do sistema. Para $n > 0$, a transição do estado n a $n+1$ se dá com taxa λ e entre n e $n-1$ com taxa μ. Quando $n = 0$, só pode haver chegadas. Dessa forma, chegadas correspondem a nascimentos e o fim de serviço a uma morte. A figura 2.2 apresenta um diagrama com as taxas de mudança do processo.

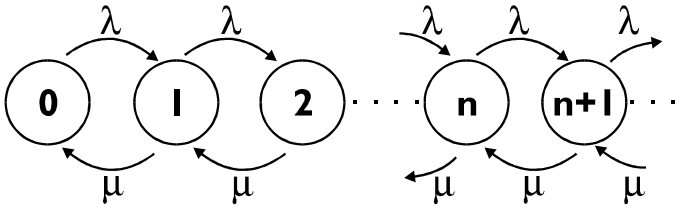

Figura 2.2: Taxas de transição no processo de Markov

Seja $\pi_n(t) = P(N(t) = n)$ a distribuição de probabilidade do número no sistema no instante t; sua respectiva distribuição estacionária será π_n. Sob algumas condições (ver por exemplo Çinlar [1975]), π_n será também a distribuição limite de $\pi_n(t)$, quando t vai a infinito. O gerador infinitesimal do processo $N(t)$ é dado por:

$$\Lambda = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & \cdots & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & \cdots \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ 0 & 0 & \mu & -(\lambda+\mu) & \lambda & \cdots \\ 0 & 0 & 0 & \ddots & \ddots & \ddots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

Como é usual na teoria dos processos de Markov, a distribuição estacionária será calculada resolvendo o sistema $\boldsymbol{\Pi}\Lambda = 0$, com a restrição adicional de $\boldsymbol{\Pi}e = 1$, onde $e = (1, 1, 1, ...)^t$ e $\boldsymbol{\Pi} = (\pi_0, \pi_1, ...)$.

Ao invés de obter $\boldsymbol{\Pi}$ resolvendo o sistema acima, vamos escrever as equações de equilíbrio do sistema usando o procedimento conhecido como balanço de fluxo e, a partir delas, obter a distribuição estacionária.

Definimos o fluxo como o produto da probabilidade estacionária pela taxa de transição. Vamos admitir que, em regime estacionário, o sistema permanece uma fração do tempo π_n no estado n. Dessa forma, se chegam λ usuários por unidade de tempo, a taxa de fluxo médio que entra em $n+1$, proveniente do estado n, é $\lambda\pi_n$. Se condições de equilíbrio existem, então, para cada estado, o fluxo que sai deve ser igual ao fluxo que entra nesse estado. Para determinar o fluxo total que sai do estado n $(n \geq 1)$, notamos que o sistema vai a $n+1$ se uma chegada ocorre e a $n-1$ se acontece um fim de serviço. O fluxo total que entra no estado n vem a partir de $n+1$ com um fim de serviço ou então, a partir de $n-1$, pela ocorrência de uma chegada. Para o estado $n=0$, o fluxo que sai corresponde a uma chegada, enquanto que o fluxo que entra, vem do estado 1 devido a um fim de serviço. As equações de balanço são as seguintes:

$$\lambda\,\pi_n + \mu\,\pi_n = \mu\,\pi_{n+1} + \lambda\,\pi_{n-1} \quad , n \geq 1 \tag{2.1a}$$

$$\lambda\,\pi_0 = \mu\,\pi_1. \tag{2.1b}$$

Esse sistema é denominado de *equações de balanço global* e será resolvido por dois diferentes métodos. O primeiro método usa indução matemática e é o mais fácil de ser aplicado nesse caso. No segundo, a solução é obtida através da aplicação da função geradora de probabilidade, técnica muito útil em modelos mais complexos.

a) Solução através do Método Iterativo

Resolvendo as equações (2.1) para $n = 0, 1, 2, \ldots$ temos,

$$\begin{aligned}
\pi_1 &= (\lambda/\mu)\pi_0 \\
\pi_2 &= (\lambda/\mu)^2\pi_0 \\
\pi_3 &= (\lambda/\mu)^3\pi_0 \\
&\vdots
\end{aligned}$$

A expressão sugerida seria $\pi_k = (\lambda/\mu)^k\pi_0$. Supomos que seja válida para $k = n$ e vamos verificar que também vale no caso $n+1$. De (2.1a) vem

$$(\lambda+\mu)(\lambda/\mu)^n\pi_0 \;= \mu\,\pi_{n+1} + \lambda\,(\lambda/\mu)^{n-1}\pi_0\,.$$

Portanto,
$$\pi_{n+1} = \Big[\frac{\lambda^{n+1} + \mu\lambda^n}{\mu^{n+1}} - \Big(\frac{\lambda}{\mu}\Big)^n\Big]\pi_0\,,$$

$$\pi_{n+1} = \Big(\frac{\lambda^{n+1} + \mu\lambda^n - \mu\lambda^n}{\mu^{n+1}}\Big)\pi_0\,,$$
$$\pi_{n+1} = \Big(\frac{\lambda}{\mu}\Big)^{n+1}\pi_0.$$

Então, pelo príncipio de indução, o resultado é válido para qualquer n finito.

Antes de completarmos o cálculo da distribuição estacionária (falta o valor de π_0), vamos introduzir o conceito de fator de utilização, denotado por ρ e definido por:

$$\rho = \lambda/\mu \tag{2.2}$$

O fator ρ é também chamado de intensidade de tráfego, pois sua expressão pode ser interpretada como o produto do número médio de chegadas (λ) pela média do tempo de serviço ($1/\mu$), indicando uma taxa média de circulação dos usuários através do sistema.

Vamos agora calcular π_0. Utilizaremos a condição de que a soma das probabilidades estacionárias deve ser igual a 1. Temos,

$$\sum_{\mathrm{n=0}}^{\infty}\pi_n = 1 \Rightarrow \sum_{n=0}^{\infty}(\frac{\lambda}{\mu})^n\pi_0 = 1 \Rightarrow \sum_{n=0}^{\infty}\rho^n\pi_0 = 1 \Rightarrow \pi_0 = 1/\sum_{n=0}^{\infty}\rho^n\,.$$

A soma acima é a da série geométrica de razão ρ que convergirá se e só se $|\rho| < 1$. Como λ e μ são estritamente positivos, essa condição equivale a $0 < \rho < 1$. Note que nessa condição, a taxa média de serviço por unidade de tempo (μ) é maior do que a taxa média com que os fregueses estão chegando (λ). Portanto, independente do quanto crescer, a fila se esvaziará de tempos em tempos e o processo de Markov será recorrente positivo com uma única distribuição de equilíbrio. Efetuando a soma da série acima obtemos $\pi_0 = 1 - \rho$ e então

$$\pi_n = \rho^n(1-\rho)\,, \quad n \geq 0,\ 0 < \rho < 1. \tag{2.3}$$

b) Solução através de Função Geradora

A função geradora de probabilidade é calculada por $P(z) = \sum_{n=0}^{\infty}\pi_n z^n$, onde z é um número complexo e $|z| \leq 1$. O procedimento consiste em obter uma expressão fechada para $P(z)$ partindo das equações (2.1) e, de sua expansão em série de potências, determinar π_n através da igualdade dos respectivos coeficientes. O desenvolvimento em série nem sempre é simples e em muitos modelos o cálculo não vai além da expressão de $P(z)$. Mesmo nesses casos, os

momentos da distribuição podem ser determinados usando as propriedades da função geradora.

Utilizando a expressão de ρ, multiplicamos (2.1a) por z^n e, após um rearranjo, obtemos

$$\pi_{n+1}\, z^n = (\rho\ + 1)\pi_n z^n - \ \rho\, \pi_{n-1} z^n.$$

Somando para n de 1 até ∞ e usando a definição de função geradora, vem

$$z^{-1} \sum_{n=1}^{\infty} \pi_{n+1}\, z^{n+1} \ = \ (\rho + 1) \sum_{n=1}^{\infty} \pi_n z^n - \rho z \sum_{n=1}^{\infty} \pi_{n-1} z^{n-1}$$

$$z^{-1} \Big(P(z) - \pi_1 z - \pi_0 \Big) = (\rho + 1) \Big(P(z) - \pi_0 \Big) - \rho z P(z)\ .$$

Com o uso da expressão (2.1b), substituimos o valor de π_1, para obter

$$P(z) = \frac{\pi_0}{1 - z\rho}\ .$$

Para determinar π_0, vamos usar, novamente, que a soma das probabilidades deve ser igual a 1. Note que isto é equivalente a aplicar $P(z)$ em $z = 1$ e igualar o resultado a 1. Assim,

$$P(1) = \sum_{n=0}^{\infty} \pi_n\, 1^n \ \Rightarrow 1 \ = \ \frac{\pi_0}{1 - 1\rho} \ \Rightarrow \ \pi_0 = 1 - \rho \ .$$

Se $\pi_0 = 0$, $P(z)$ seria identicamente zero para todo z. Logo, assumimos $\pi_0 > 0$ e segue que $1 - \rho > 0$ ou ainda, $\rho < 1$. Dessa forma, obtemos,

$$P(z) = \ \frac{1 - \ \rho}{1 - z\rho} \quad , 0 < \rho < 1. \tag{2.4}$$

A expansão de $P(z)$ não é difícil nesse caso. O quociente $1/(1 - z\rho)$ é a soma da série geométrica que tem termo inicial 1 e razão $z\rho$. Então,

$$P(z) = \ \sum_{n=0}^{\infty} (1 - \rho)\rho^n z^n \ \Rightarrow \pi_n = \ (1 - \rho)\rho^n,\ n \geq 0,$$

conforme já obtido anteriormente.

A distribuição obtida pelos dois métodos acima é, frequentemente, denominada de distribuição estacionária em tempo contínuo, em oposição às distribuições imersas em instantes particulares do tempo, as quais serão introduzidas nas próximas seções. A distribuição em tempo contínuo é

interpretada como a fração do tempo em que o sistema permanece em cada estado. Isto é, π_n indica a proporção do tempo total em que o sistema tem n usuários presentes. Uma outra interpretação considera essa distribuição como aquela vista por um observador externo, ou seja, ele vê o sistema no estado n, com probabilidade π_n.

A partir da distribuição estacionária, vamos calcular o número médio de usuários na fila e no sistema. Outras medidas de desempenho serão apresentadas nas próximas seções. Seja N o número total de usuários no sistema em regime estacionário e L sua média. Temos

$$L = E(N) = \sum_{n=0}^{\infty} n\,\pi_n = \sum_{n=0}^{\infty} n\,(1-\rho)\rho^n = (1-\rho)\sum_{n=0}^{\infty} n\,\rho^n.$$

Esta última somatória pode ser reescrita como

$$\sum_{n=0}^{\infty} n\,\rho^n = \rho + 2\rho^2 + 3\rho^3 + ... = \rho\,(1 + 2\rho + ...) = \rho\sum_{n=1}^{\infty} n\,\rho^{n-1}\,.$$

A derivada de ρ^n em relação a ρ é dada por $n\rho^{n-1}$. Assim, admitindo satisfeitas as condições para inverter os sinais de somatória e derivada, obtemos

$$L \;=\; (1-\rho)\,\rho\sum_{n=1}^{\infty} n\,\rho^{n-1} \;= (1-\rho)\rho\sum_{n=0}^{\infty}\frac{d}{d\rho}\,\rho^n =$$

$$=\; (1-\rho)\rho\,\frac{d}{d\rho}\sum_{n=0}^{\infty}\rho^n = (1-\rho)\rho\,\frac{d}{d\rho}\,\Big(1/(1-\rho)\Big).$$

Concluimos então que

$$L = \frac{\rho}{(1-\rho)} = \frac{\lambda}{\mu-\lambda}\;\;. \tag{2.5}$$

Seja N_q o número de usuários na fila e L_q seu valor esperado. Temos

$$L_q = E(N_q) = 0\,\pi_0 + \sum_{n=1}^{\infty}(n-1)\pi_n = \sum_{n=1}^{\infty} n\,\pi_n - \sum_{n=1}^{\infty}\pi_n\;\;,$$

então,
$$L_q = L - (1-\pi_0) = \frac{\rho^2}{(1-\rho)} = \frac{\lambda^2}{\mu\,(\mu-\lambda)} \tag{2.6}$$

2.2 O resultado *Pasta* e o tempo de espera

Vamos introduzir nesta seção um importante resultado, conhecido como *Pasta,* que é uma abreviatura da frase em inglês *Poisson Arrivals See Time Averages.* Esse resultado indica que a distribuição do número de usuários no sistema em tempo contínuo é igual àquela obtida observando-se o sistema nos instantes imediatamente precedentes a uma chegada. Isto é, em regime estacionário, os usuários que chegam "enxergam" o sistema com a distribuição em tempo contínuo. A probabilidade de uma chegada encontrar o sistema num particular estado é uma quantidade de interesse no estudo de filas e, em especial, será necessária para calcular a distribuição do tempo de espera na fila e no sistema.

Seja $\pi_n^a(t)$ a probabilidade de uma chegada no instante t encontrar o sistema no estado n (lembramos que o superescrito a vem de *arrival*, o termo em inglês para chegadas). O usuário, que está chegando em t não é computado no estado do sistema e, por essa razão, $\pi_n^a(t)$ é frequentemente mencionada como a distribuição nos instantes imediatamente anteriores a uma chegada ou distribuição imersa nos instantes de chegada. Vamos designar por π_n^a a correspondente distribuição estacionária. A primeira vista, poderíamos pensar que a distribuição em tempo contínuo π_n deveria sempre ser igual a π_n^a para todos os sistemas de filas. Entretanto, isso não acontece, como ilustra o próximo exemplo.

Exemplo 2.1:

Considere uma fila D/D/1, em que as chegadas e serviços são determinísticos, isto é, as chegadas são espaçadas igualmente a cada x segundos e o serviço tem a duração de exatamente y segundos. Para garantir a estabilidade do sistema vamos considerar que $y < x$, ou seja, o serviço é mais rápido que o intervalo entre chegadas. Dessa forma, após o equilíbrio do sistema ser atingido, o usuário que chega encontra a fila vazia e teremos $\pi_0^a = 1$ e $\pi_k^a = 0$, $\forall\, k \geq 1$. Por outro lado, a fração do tempo que o sistema contém um usuário é y/x. No restante do tempo o sistema estará vazio. Dessa forma, $\pi_0 = 1 - y/x$, $\pi_1 = y/x$ e $\pi_k = 0$, $\forall\, k \geq 2$. Temos assim um exemplo simples em que $\pi_n \neq \pi_n^a$.□

Existe uma larga classe de sistemas de filas em que o resultado *Pasta* é válido, todos eles tendo chegadas Poisson. Uma característica importante é a independência entre o processo de chegada após o instante t e o estado do sistema

em t. É possível provar o resultado *Pasta* mesmo se o modelo adotado não for um processo de Markov ou de nascimento e morte (ver Wolff [1988]).

Teorema 2.1: Resultado *Pasta*

Na fila M/M/1, em regime estacionário, a distribuição do número de usuários no sistema, nos instantes de chegada, coincide com a distribuição em tempo contínuo. Isto é, $\pi_n = \pi_n^a$.

demonstração:

Vamos verificar que, em equilíbrio, $\pi_n(t) = \pi_n^a(t) \; \forall \, t$, implicando que as respectivas distribuições estacionárias são iguais. O evento correspondendo a uma chegada no intervalo $(t, t+\Delta t)$ será representado por $A(t, t+\Delta t)$. Então,

$$\begin{aligned}
\pi_n^a(t) &= \lim_{\Delta t \to 0} P[N(t) = n \mid A(t, t+\Delta t)] \\
&= \lim_{\Delta t \to 0} \frac{P[N(t) = n, A(t, t+\Delta t)]}{P[A(t, t+\Delta t)]} \\
&= \lim_{\Delta t \to 0} \frac{P[A(t, t+\Delta t) \mid N(t) = n] \, P[N(t) = n]}{P[A(t, t+\Delta t)]} \\
&= \lim_{\Delta t \to 0} \frac{P[A(t, t+\Delta t)] \; P[N(t) = n]}{P[A(t, t+\Delta t)]} \\
&= \lim_{\Delta t \to 0} P[N(t) = n] \\
&= \pi_n(t),
\end{aligned}$$

onde aplicamos a independência do evento $A(t, t+\Delta t)$ em relação ao número de usuários no sistema no instante t. Note que isso é uma consequência direta das chegadas formarem um processo de Poisson.□

Vamos agora retomar o cálculo das medidas de desempenho, obtendo o tempo de espera na fila e no sistema. O tempo gasto pelo usuário, no sistema, depende da disciplina de atendimento. No caso da fila M/M/1, a disciplina é FCFS e o tempo de espera de um usuário dependerá de quantos estão presentes no sistema no momento da chegada. Sejam θ_q a variável aleatória que descreve o tempo gasto na fila, $F_{wq}(t)$ sua função distribuição e W_q sua média. Observe que, como vale o resultado *Pasta* para a fila M/M/1, a distribuição nos instantes de chegada é aquela dada pela expressão (2.3).

No cálculo de $F_{wq}(t)$ devemos notar que essa distribuição é do tipo mista com parte discreta e parte contínua. Para $t = 0$ temos,

$$F_{wq}(0) = P[\theta_q \leq 0] = P[\text{Sistema vazio no instante de chegada}] = \pi_0^a = 1 - \rho\,.$$

Considere agora o caso $t > 0$ e suponha que n usuários estão no sistema, no instante de uma chegada. Nesse caso, para que o usuário, ao chegar, não espere mais de t, é necessário que a soma do tempo do serviço em processamento mais o tempo de serviço dos outros $n-1$ usuários, que já estavam na fila, não seja maior que t. Pela propriedade da falta de mémoria da exponencial, não importa quanto tempo decorreu desde o ínicio do serviço até o instante em que ocorre a chegada. Esse tempo terá, também, distribuição exponencial com o mesmo parâmetro μ. Teremos então a convolução de n exponenciais independentes e identicamente distribuídas, o que resulta em uma Erlang-n. Assim,

$$\begin{aligned}
F_{wq}(t) = &\ P[\theta_q \leq t] \\
= &\ F_{wq}(0) + \sum_{n=1}^{\infty}\Big\{P[n \text{ serviços} \leq t \mid \text{encontra } n \text{ no sistema}] \\
&\qquad\qquad P[\text{encontra } n \text{ no sistema}]\Big\} \\
= &\ (1-\rho) + (1-\rho)\sum_{n=1}^{\infty}\rho^n \int_0^t \frac{\mu(\mu x)^{n-1}}{(n-1)!}e^{-\mu x}\,dx \\
= &\ (1-\rho) + (1-\rho)\rho\int_0^t \mu e^{-\mu x}\sum_{n=1}^{\infty}\frac{(\mu x\rho)^{n-1}}{(n-1)!}\,dx \\
= &\ (1-\rho) + (1-\rho)\rho\int_0^t \mu e^{-\mu x(1-\rho)}dx \\
= &\ 1 - \rho e^{-\mu(1-\rho)t}\ , t > 0.
\end{aligned}$$

Então, $$F_{wq}(t) = \begin{cases} 1-\rho & , t = 0; \\ 1-\rho e^{-\mu(1-\rho)t} & , t > 0. \end{cases}$$

Para calcular o valor esperado W_q, precisamos integrar uma função mista e usaremos a integral de Riemann-Stieltjes. Assim,

$$W_q = E(\theta_q) = \int_0^\infty t\, dF_{wq}(t) = 0(1-\rho) + \int_0^\infty t\rho\mu(1-\rho)e^{-\mu(1-\rho)t}\, dt\,.$$

Calculando a integral e substituindo ρ pela sua expressão, obtemos,

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)}\,. \tag{2.7}$$

O tempo total gasto no sistema, representado por θ_{tot}, é a soma de θ_q com o tempo de serviço S, que tem distribuição B. Pela independência dessas variáveis, a distribuição de θ_{tot}, denotada por F_w, será a convolução das respectivas funções de distribuição. Como S tem distribuição exponencial, o tempo gasto no sistema será estritamente positivo e portanto $F_w(0) = 0$. Para $t \geq 0$, temos

$$\begin{aligned} F_w(t) = P(\theta_{tot} \leq t) &= P(\theta_q + S \leq t) = \\ &= \int_0^t [1 - \rho e^{-\mu(1-\rho)(t-s)}]\, dB(s) \\ &= \int_0^t [1 - \rho\, e^{-\mu(1-\rho)(t-s)}]\mu e^{-\mu s} ds \\ &= 1 - e^{-\mu t} - e^{-\mu(1-\rho)t}(1 - \ e^{-\mu\rho t})\,. \end{aligned}$$

Concluimos então que

$$F_w(t) = 1 - e^{-(\mu-\lambda)t}\,.$$

Note que $\mu - \lambda$ é um número positivo por suposição anterior ($\rho < 1$) e, portanto, a distribuição de θ_{tot} é exponencial com parâmetro $\mu - \lambda$. De imediato temos que seu valor esperado, representado por W, será

$$W = E(\theta_{tot}) = \frac{1}{(\mu - \ \lambda)}\,. \tag{2.8}$$

2.3 Distribuição nos instantes de saída

Nesta seção vamos estudar a fila M/M/1 observando o sistema nos instantes de saída (ou partida) dos usuários. Calculamos a distribuição estacionária do número de clientes na fila após uma saída, e verificamos que ela é igual às distribuições em tempo contínuo e imersa nos instantes de chegada. Vamos, também, enunciar e demonstrar um dos mais importantes resultados em teoria das filas, o teorema de Burke (ver Burke [1956]), que caracteriza as saídas da fila M/M/1 como sendo um processo de Poisson.

Sejam T_0, T_1, T_2, ... os sucessivos instantes de fim de serviço, correspondendo às saídas da fila. Denotamos por X_n o número de usuários na fila imediatamente após T_n, isto é, $X_n = N(T_n^+)$.

Vamos verificar que $(X, T) = \{(X_n,T_n); n \geq 0\}$ é um processo de renovação Markoviano. Precisamos mostrar que, para todo $n \geq 1$ e $t \geq 0$, as probabilidades de transição só dependem do estado presente. Isto é,

$$P(X_n, T_n - T_{n-1} \leq t \,|(X_k, T_k), k \leq n-1) = P(X_n, T_n - T_{n-1} \leq t \,|X_{n-1}).$$

Inicialmente, observe que o intervalo de tempo $T_n - T_{n-1}$ será dependente apenas de X_{n-1}, o número deixado no sistema pela $(n-1)$-ésima saída. Se $X_{n-1} > 0$, esse tempo tem distribuição exponencial com parâmetro μ, correspondendo a um serviço completado. Por outro lado, se $X_{n-1} = 0$, o intervalo entre saídas será a soma do tempo de espera para uma chegada mais o tempo de serviço e, portanto, sua distribuição será a convolução de duas exponenciais independentes com parâmetros λ e μ. Quanto a X_n, ele dependerá de quantas chegadas ocorreram desde a última saída. Como as chegadas formam um processo de Poisson, o valor de X_n dependerá apenas de λ e da duração do intervalo $T_n - T_{n-1}$. Dessa forma, o conhecimento de X_{n-1} é suficiente para caracterizar a probabilidade de transição.

A matriz $Q(t)$ é o núcleo do processo (X, T) e seus elementos são:

$$Q_{ij}(t) = P(X_n = j, T_n - T_{n-1} \leq t \,|X_{n-1} = i)\,, t \geq 0, i, j = 0, 1, 2,....$$

Essa expressão refere-se a uma transição entre as saídas $n-1$ e n, porém, pela homogeneidade do processo, ela vale para qualquer outra transição.

Para facilitar o trabalho algébrico, o núcleo de transição será obtido com o uso das transformadas de Laplace-Stieltjes (LS), definidas no apêndice. A transformada de $Q(t)$ será denotada por $\tilde{Q}(s)$ com os correspondentes $Q_{ij}(t)$ representados por $\tilde{Q}_{ij}(s)$. Lembremos, ainda, que a função de distribuição da convolução de variáveis aleatórias independentes é o produto das respectivas transformadas.

Sejam $\tilde{A}(s)$ a transformada da exponencial (λ) correspondente à chegada e $\tilde{H}(s)$ a transformada da exponencial $(\lambda+\mu)$ correspondente ao mínimo entre duas exponenciais independentes (chegadas e serviços). Temos que $\tilde{A}(s)=\lambda/(\lambda+s)$ e $\tilde{H}(s)=(\lambda+\mu)/(\lambda+\mu+s)$. Note que o quociente $\mu/(\lambda+\mu)$ representa a probabilidade do mínimo entre as duas exponenciais ser um fim de serviço. Da mesma forma, teríamos essa probabilidade igual a $\lambda/(\lambda+\mu)$ se o mínimo fosse a chegada.

No cálculo de $Q(t)$ vamos distinguir dois casos:

caso 1: transição $0 \rightarrow j$, $j \geq 0$.

O elemento $Q_{0j}(t)$ corresponde a ocorrência, antes de t, da primeira chegada e de outras j chegadas antes do fim de serviço. Teremos assim que, após a primeira chegada que ocupa o servidor, as próximas j ocorrências serão chegadas. Isto é, o mínimo entre as duas exponenciais, serviço e chegada, é uma chegada em j ocasiões seguidas. Finalmente, a disputa do mínimo é vencida pelo serviço dando origem a transição do processo. Tendo em vista que tratamos com a soma de eventos independentes, teremos o aparecimento da convolução das funções de distribuição. Assim, a transformada LS de $Q_{0j}(t)$ será:

$$\begin{aligned}\tilde{Q}_{0j}(s) &= \tilde{A}(s)\,\frac{\lambda^j}{(\lambda+\mu)^j}\,\tilde{H}^j(s)\frac{\mu}{\lambda+\mu}\,\tilde{H}(s)\\ &= \frac{\lambda}{(\lambda+s)}\,\frac{\lambda^j}{(\lambda+\mu)^j}\,\frac{(\lambda+\mu)^j}{(\lambda+\mu+s)^j}\,\frac{\mu}{(\lambda+\mu)}\,\frac{(\lambda+\mu)}{(\lambda+\mu+s)}\\ &= \frac{\lambda}{(\lambda+s)}\,\frac{\lambda^j}{(\lambda+\mu+s)^j}\,\frac{\mu}{(\lambda+\mu+s)}\,,\ j\geq 0\,.\end{aligned}$$

caso 2: transição $i \rightarrow j$, $j \geq i-1$, $i>0$.

Nesse caso, devem ocorrer $j-i+1$ chegadas antes do fim de serviço:

$$\begin{aligned}\tilde{Q}_{ij}(s) &= \frac{\lambda^{j-i+1}}{(\lambda+\mu)^{j-i+1}}\tilde{H}^{\,j-i+1}(s)\frac{\mu}{\lambda+\mu}\tilde{H}(s)\\ &= \frac{\lambda^{j-i+1}}{(\lambda+\mu)^{j-i+1}}\,\frac{(\lambda+\mu)^{j-i+1}}{(\lambda+\mu+s)^{j-i+1}}\,\frac{\mu}{(\lambda+\mu)}\,\frac{(\lambda+\mu)}{(\lambda+\mu+s)}\\ &= \frac{\lambda^{j-i+1}}{(\lambda+\mu+s)^{j-i+1}}\,\frac{\mu}{(\lambda+\mu+s)}\,,\ \ j\geq i-1, i>0.\end{aligned}$$

Todas as outras transições são nulas pois, como observamos o sistema nos instantes de saída, só poderá haver decréscimos unitários no número de usuários. Por outro lado, os acréscimos podem ser ilimitados. Dessa forma, o núcleo terá valores não nulos somente à direita da sub-diagonal inferior,

$$\tilde{Q}(s) = \begin{bmatrix} \frac{\lambda\mu}{(\lambda+s)(\lambda+\mu+s)} & \frac{\lambda^2\mu}{(\lambda+s)(\lambda+\mu+s)^2} & \frac{\lambda^3\mu}{(\lambda+s)(\lambda+\mu+s)^3} & \cdots \\ \frac{\mu}{(\lambda+\mu+s)} & \frac{\lambda\mu}{(\lambda+\mu+s)^2} & \frac{\lambda^2\mu}{(\lambda+\mu+s)^3} & \cdots \\ 0 & \frac{\mu}{(\lambda+\mu+s)} & \frac{\lambda\mu}{(\lambda+\mu+s)^2} & \cdots \\ \vdots & \ddots & \ddots & \ddots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

A matriz de transição, da cadeia de Markov imersa no processo de renovação Markoviano, é obtida fazendo $s \to 0$ em $\tilde{Q}(s)$, o que equivale a $t \to \infty$ em $Q(t)$. Então,

$$P = \begin{bmatrix} \frac{\mu}{(\lambda+\mu)} & \frac{\lambda\mu}{(\lambda+\mu)^2} & \frac{\lambda^2\mu}{(\lambda+\mu)^3} & \cdots \\ \frac{\mu}{(\lambda+\mu)} & \frac{\lambda\mu}{(\lambda+\mu)^2} & \frac{\lambda^2\mu}{(\lambda+\mu)^3} & \cdots \\ 0 & \frac{\mu}{(\lambda+\mu)} & \frac{\lambda\mu}{(\lambda+\mu)^2} & \cdots \\ \vdots & \ddots & \ddots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

A matriz de transição P é irredutível e aperiódica. Vamos assumir ainda que os parâmetros λ e μ são escolhidos de modo que a cadeia seja recorrente positiva. Segue então, da teoria das cadeias de Markov, que a distribuição estacionária será única. Vamos representa-la por $\Pi^d = (\pi_0^d,\ \pi_1^d, \dots)$, com o superescrito d de *departures* (partidas em inglês).

Proposição 2.2

Em regime estacionário, o usuário deixa a fila M/M/1 com a mesma distribuição vista por um observador externo, isto é $\Pi^d = \Pi$.

demonstração:

Como devemos ter unicidade de solução para Π^d basta verificar que a distribuição Π, dada pela expressão (2.3), satisfaz as equações de equilíbrio:

$$\Pi^d P = \Pi^d \quad \text{e} \quad \Pi^d e = 1.$$

Como Π é também probabilidade, a soma de suas componentes é igual a 1 e só precisamos verificar a primeira dessas equações. Em forma escalar, essa equação vetorial se torna o seguinte sistema:

$$\pi_0^d \frac{\mu}{\lambda+\mu} + \pi_1^d \frac{\mu}{\lambda+\mu} = \pi_0^d \tag{2.9a}$$

$$\pi_0^d \frac{\lambda^n \mu}{(\lambda+\mu)^{n+1}} + \sum_{k=1}^{n+1} \pi_k^d \frac{\lambda^{n-k+1}\mu}{(\lambda+\mu)^{n-k+2}} = \pi_n^d \ , n \geq 1. \tag{2.9b}$$

Pela expressão (2.3), o valor proposto para Π^d será

$$\Pi^d = (1-\rho)(1, \rho, \rho^2, \rho^3, \ldots) ,$$

o que, substituindo em (2.9a), resulta

$$\text{Lado esquerdo} = (1-\rho)\frac{\mu}{\lambda+\mu} + (1-\rho)\,\rho\frac{\mu}{\lambda+\mu} =$$
$$= (1-\rho)\Big(\frac{\mu}{\lambda+\mu} + \frac{\lambda}{\mu}\,\frac{\mu}{\lambda+\mu}\Big) = (1-\rho) = \pi_0^d = \text{Lado direito.}$$

Para a equação (2.9b) temos

$$\text{Lado esquerdo} = (1-\rho)\frac{\lambda^n\mu}{(\lambda+\mu)^{n+1}} + \sum_{k=1}^{n+1}(1-\rho)\rho^k \frac{\lambda^{n-k+1}\mu}{(\lambda+\mu)^{n-k+2}}$$
$$= (1-\rho)\left(\frac{\lambda^n\mu}{(\lambda+\mu)^{n+1}} + \sum_{k=1}^{n+1}\rho^k \frac{\lambda^{n-k+1}\mu}{(\lambda+\mu)^{n-k+2}}\right)$$
$$= (1-\rho)\rho^n\left(\frac{1}{(1+\rho)^{n+1}} + \frac{\rho}{(1+\rho)^{n+2}}\sum_{k=1}^{n+1}\frac{1}{(1+\rho)^{-k}}\right)$$
$$= (1-\rho)\rho^n\left(\frac{1}{(1+\rho)^{n+1}} + \frac{\rho}{(1+\rho)^{n+2}}\frac{(1+\rho)\big((1+\rho)^{n+1}-1\big)}{(1+\rho)-1}\right)$$

$$= (1-\rho)\rho^n \left(\frac{1+(1+\rho)^{n+1}-1}{(1+\rho)^{n+1}} \right)$$

$$= (1-\rho)\rho^n = \pi_n^d = \text{Lado direito.}$$

Dessa forma, o valor proposto para Π^d satisfaz as equações de equilíbrio e, pela unicidade da distribuição estacionária, concluimos que $\Pi^d = \Pi$ e a proposição está provada.□

Vamos estudar, agora, as características do processo pontual de saídas, isto é, estamos interessados no comportamento da sequência $\{T_n;\ n \geq 0\}$. Por abuso de linguagem, essa sequência é frequentemente referida como sendo o processo de saída da fila. O mais correto seria mencionar que ela é o processo marginal dos tempos de saída, referente ao processo *(X, T)*. O resultado, que vamos obter, caracteriza a sequência de tempos de saída para as filas M/M/1 e é conhecido como teorema de Burke, sendo um dos mais importantes resultados em teoria das filas. Em modelos mais complexos, em que as filas se interconectam, a busca de resultados similares ao teorema de Burke é muito importante para o estudo global do sistema.

Teorema 2.3 : Teorema de Burke (via equivalência)

Na fila M/M/1 em equilíbrio, a sequência de saídas forma um processo de Poisson com parâmetro λ.

demonstração:

Vamos mostrar que o processo *(X, T)* é equivalente a um processo de renovação com distribuição exponencial de parâmetro λ. Pela proposição b3, apresentada no apêndice, devemos verificar a igualdade $\Pi^d\, Q(t) = A(t)\, \Pi^d$, ou, com o uso de transformadas, $\Pi^d\, \tilde{Q}(s) = \frac{\lambda}{\lambda+s}\, \Pi^d$. Desenvolvendo o lado esquerdo dessa equação, o elemento genérico desse vetor, para $n \geq 0$, será

$$\begin{aligned}
\left[\Pi^d\,\tilde{Q}(s)\right]_n &= \pi_0^d \frac{\lambda^{n+1}\mu}{(\lambda+s)(\lambda+\mu+s)^{n+1}} + \sum_{k=1}^{n+1} \pi_k^d\, \frac{\lambda^{n-k+1}\mu}{(\lambda+\mu+s)^{n-k+2}} \\
&= (1-\rho)\Big(\frac{\lambda^{n+1}\mu}{(\lambda+s)(\lambda+\mu+s)^{n+1}} + \sum_{k=1}^{n+1} \rho^k\, \frac{\lambda^{n-k+1}\mu}{(\lambda+\mu+s)^{n-k+2}} \Big) \\
&= (1-\rho)\Big(\frac{\lambda\rho^n\mu^{n+1}}{(\lambda+s)(\lambda+\mu+s)^{n+1}} + \sum_{k=1}^{n+1} \rho^k \frac{\rho^{n-k+1}\mu^{n-k+2}}{(\lambda+\mu+s)^{n-k+2}} \Big) \\
&= (1-\rho)\rho^n \frac{\lambda}{\lambda+s}\Big(\frac{\mu^{n+1}}{(\lambda+\mu+s)^{n+1}} + \\
&\qquad \frac{(\lambda+s)}{\lambda} \frac{\rho\mu^{n+2}}{(\lambda+\mu+s)^{n+2}}\ \sum_{k=1}^{n+1} \frac{\mu^{-k}}{(\lambda+\mu+s)^{-k}} \Big) \\
&= (1-\rho)\,\rho^n \frac{\lambda}{\lambda+s}\ \ \frac{\mu^{n+1}}{(\lambda+\mu+s)^{n+1}} \Big(1 + \\
&\qquad \frac{(\lambda+s)}{(\lambda+\mu+s)}\ \sum_{k=1}^{n+1} \frac{\mu^{-k}}{(\lambda+\mu+s)^{-k}} \Big).
\end{aligned}$$

Como,
$$\begin{aligned}
\sum_{k=1}^{n+1} \frac{\mu^{-k}}{(\lambda+\mu+s)^{-\mathrm{k}}} &= \frac{\frac{(\lambda+\mu+s)}{\mu}\left\{\left(\frac{\lambda+\mu+s}{\mu}\right)^{n+1} - 1\right\}}{\frac{(\lambda+\mu+s)}{\mu} - 1} \\
&= \frac{(\lambda+\mu+s)}{(\lambda+s)}\left\{\left(\frac{\lambda+\mu+s}{\mu}\right)^{n+1} - 1\right\}.
\end{aligned}$$

Então,

$$\begin{aligned}
\left[\Pi^d\,\tilde{Q}(s)\right]_n &= (1-\rho)\rho^n \frac{\lambda}{\lambda+s} \frac{\mu^{n+1}}{(\lambda+\mu+s)^{n+1}} \Big(\frac{\lambda+\mu+s}{\mu}\Big)^{n+1} \\
&= (1-\rho)\rho^n \frac{\lambda}{\lambda+s} \\
&= \frac{\lambda}{\lambda+s}\,\pi_n^d \quad , n \geq 0.
\end{aligned}$$

Concluimos, então, que *(X, T)* é equivalente a um processo de renovação com distribuição exponencial de taxa λ, isto é, a sequência das saídas forma um processo de Poisson (λ).□

***Comentário*:** O teorema de Burke foi provado originalmente para a fila M/M/c e a demonstração apresentada aqui pode ser aplicada para esse caso também. Uma outra prova, usando reversibilidade, será vista na seção 2.5.

2.4 Fórmulas de Little

Nesta seção vamos estabelecer algumas relações entre as medidas de desempenho do sistema. Essas expressões, conhecidas como fórmulas de Little, relacionam o número médio de usuários (L ou L_q) com o tempo médio de espera (W ou W_q) e tiveram a sua primeira demonstração formal no trabalho de Little [1961]. Desde então provas alternativas e extensões têm sido apresentadas e sua validade extrapolou os modelos Markovianos e pode ser verificada para sistemas mais gerais (ver Wolff [1989]).

Admitindo a existência de equilíbrio no sistema M/M/1, a relação $L = \lambda W$ pode ser explicada intuitivamente da seguinte forma. Um usuário que chega, espera em média W para sair do sistema. Suponha que, ao sair do serviço, ele olha para trás e observa quantos usuários ficaram no sistema. Lembre que a disciplina de atendimento é FCFS e portanto ele deve "enxergar" em média L usuários presentes, aqueles que chegaram durante seu tempo de serviço e de espera na fila. Cada um desses L usuários levou em média $1/\lambda$ para chegar e, assim, podemos escrever $L(1/\lambda) = W$ ou $L = \lambda\, W$.

O próximo teorema verifica as fórmulas de Little para a fila M/M/1. A demonstração apresentada poderia também ser aplicada para sistemas com outras distribuições de serviço, mantendo, entretanto, as chegadas seguindo um processo de Poisson e a disciplina FCFS.

Teorema 2.4: Fórmulas de Little

Na fila M/M/1 valem as seguintes relações:

$$L = \lambda W \tag{2.10}$$

$$L_q = \lambda W_q \tag{2.11}$$

demonstração:

Tendo em vista o atendimento em ordem de chegada, o seguinte argumento pode ser usado: a distribuição estacionária de n usuários no sistema, no instante de uma saída, é igual à probabilidade de n chegadas, durante o tempo total gasto no sistema por um usuário arbitrário. Aplicando a lei de probabilidade

total podemos escrever, para $n \geq 0$,

$$\pi_n^d = \int_0^\infty P(n \text{ chegadas durante } \theta_{tot} \mid \theta_{tot} = t)\, dF_w(t)\ ,$$

em que, como antes, θ_{tot} é o tempo total de um usuário no sistema e F_w sua função distribuição.

Multiplicando ambos os lados por n e somando para n de 1 ao infinito, obtemos, do lado esquerdo, o valor esperado do número de usuários no sistema nos instantes de saída. Como vale a igualdade entre as distribuições, imersa nos instantes de saída e em tempo contínuo, o valor esperado com respeito à Π^d será também igual a L. Para desenvolver o lado direito, vamos admitir válida a troca entre os sinais de integração e somatória. Temos então

$$L = \sum_{n=1}^{\infty} n \int_0^\infty \frac{(\lambda t)^n e^{-\lambda t}}{n!} dF_w(t) = \int_0^\infty \lambda t\, e^{-\lambda t} \sum_{n=1}^{\infty} \frac{(\lambda t)^{n-1}}{(n-1)!}\, dF_w(t) =$$

$$= \int_0^\infty \lambda t\ e^{-\lambda t} \left(e^{\lambda t}\right) dF_w(t) = \lambda \int_0^\infty t\ dF_w(t) = \lambda\, E(\theta_{tot}) = \lambda\, W.$$

Dessa forma, (2.10) está verificada. Com essa expressão e as relações $L = L_q - 1 + \pi_0$ (ver 2.6) e $W = E(\theta_q) + E(S) = W_q + 1/\mu$, podemos verificar imediatamente que $L_q = \lambda\, W_q$ e o teorema está demonstrado. □

Uma outra expressão de interesse, porém de validade mais restrita, relaciona o tempo médio na fila com o número médio no sistema. Das expressões (2.5) e (2.7) podemos obter

$$W_q = \frac{L}{\mu}\ , \tag{2.12}$$

que pode ser explicada intuitivamente pela seguinte argumentação. Um usuário ao chegar encontra em média L usuários presentes no sistema ($L > 0$). O seu tempo de fila será aquele necessário para o término do serviço em andamento mais $(L-1)$ serviços com média $1/\mu$ cada, referentes aos usuários em fila no instante da sua chegada. Como o serviço é exponencial, o serviço em andamento tem também duração média de $1/\mu$. Dessa forma, para um usuário arbitrário, o tempo médio de espera na fila será o produto de L por $1/\mu$. Note que o serviço

exponencial foi fundamental na argumentação acima e, dessa forma, a validade de (2.12) está restrita às filas M/M/1 e a algumas de suas variantes.

Existem outras expressões relacionando probabilidades, parâmetros e medidas de desempenho, algumas delas, inclusive, também são referidas como fórmulas de Little. O leitor interessado pode consultar Gross & Harris [1985] para outros resultados.

2.5 Reversibilidade

Nesta seção mostraremos que $N(t)$, o estado da fila M/M/1 no tempo t, é um processo reversível e vamos explorar algumas de suas consequências. O conceito de processo reverso, reversibilidade e resultados correlatos são apresentados no apêndice. Nesta seção, o tempo t poderá tomar valores em toda reta real, de modo a acomodar, sem maiores modificações, a definição do processo reverso. Com a extensão do domínio do tempo, os resultados obtidos nas seções anteriores sofreriam ligeiras modificações na formulação. No entanto, nenhuma mudança substancial aconteceria pois eles se referiram, na sua grande maioria, a intervalos e não a escalas de tempo.

Relembramos que $N(t)$ é um processo de Markov com gerador infinitesimal $\Lambda = [\Lambda_{i,j}]$, $i, j \geq 0$. Os elementos de Λ são as taxas instantâneas de transição desse processo (ver seção 2.1) e são dadas por $\Lambda_{0,0} = -\lambda$, $\Lambda_{0,1} = \lambda$ e

para $i > 0$, valem $\Lambda_{i,j} = \begin{cases} \mu, & j = i-1; \\ -(\lambda+\mu), & j = i; \\ \lambda & j = i+1. \end{cases}$

Proposição 2.5 : Reversibilidade na fila M/M/1

Em regime estacionário, o número de usuários no instante t na fila M/M/1 é um processo Markoviano reversível.

demonstração:

Para mostrar reversibilidade, utilizamos Π dada por 2.3 e vamos verificar que estão satisfeitas as equações de balanço detalhado, $\pi_i \Lambda_{i,j} = \pi_j \Lambda_{j,i}$, para todo $i, j \geq 0$.

Para $i = 0$, a única transição possível é para o estado 1 e, portanto,

$$\pi_0 \Lambda_{0,1} = (1-\rho)\lambda = (1-\rho)\lambda \, \frac{\mu}{\mu} = (1-\rho)\rho \, \mu = \pi_1 \Lambda_{1,0}.$$

Para $i > 0$, temos,

$$\pi_i \Lambda_{i,i-1} = (1-\rho)\rho^i \mu = (1-\rho)\rho^{i-1} \frac{\lambda}{\mu} \mu = (1-\rho)\rho^{i-1}\lambda = \pi_{i-1}\Lambda_{i-1,i} \text{ , e}$$

$$\pi_i \Lambda_{i,i+1} = (1-\rho)\rho^i \lambda = (1-\rho)\rho^i \frac{\mu}{\mu} = (1-\rho)\rho^{i+1}\mu = \pi_{i+1}\Lambda_{i+1,i} \text{ .}$$

Todas as outras transições são nulas e resultam em uma identidade $0 = 0$. Concluimos, então, que as equações de balanço detalhado estão satisfeitas, $N(t)$ é reversível e a proposição está provada.□

Pelo resultado acima, o processo reverso $N(-t)$ terá a mesma estrutura probabilística do processo direto $N(t)$. Para auxiliar a descrição e a compreensão das consequências desse fato na fila M/M/1, apresentamos na figura 2.3, uma realização típica do processo $N(t)$. Observando essa figura no sentido direto da evolução do tempo (da esquerda para direita), notamos que os pontos, em que há um salto para cima, correspondem às chegadas na fila, enquanto que as saídas são os saltos para baixo.

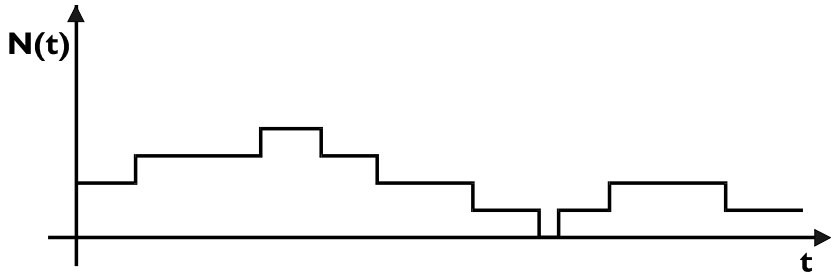

Figura 2.3: Realização típica de $N(t)$

Por outro lado, observando a evolução do processo reverso (da direita para esquerda), os papéis se invertem e as chegadas são os pontos em que, no processo direto, tivemos saídas. Da mesma forma, saídas no reverso correspondem a pontos de chegada no processo direto. Essa será a argumentação básica para provar as próximas proposições. Começamos por demonstrar de novo o teorema de Burke (teorema 2.3), usando agora a reversibilidade.

Teorema 2.6 Teorema de Burke (via reversibilidade)

Na fila M/M/1 em equilíbrio, a sequência de saídas forma um processo de Poisson com parâmetro λ.

demonstração:

Os pontos no tempo, em que a trajetória do processo $N(t)$ salta para cima, formam um processo de Poisson de parâmetro λ. Pela reversibilidade do processo $N(t)$, os pontos de salto para cima, na trajetória do processo reverso $N(-t)$, também precisam formar um processo de Poisson de parâmetro λ.

Note agora que os pontos em que $N(-t)$ salta para cima correspondem aos pontos no tempo em que o processo direto $N(t)$ salta para baixo e, assim, eles também formam um processo de Poisson (λ). Como os tempos de saída da fila M/M/1 correspondem aos pontos de salto para baixo na trajetória de $N(t)$, concluimos que a sequência de saídas forma um processo de Poisson com parâmetro λ. □

Proposição 2.7

Na fila M/M/1 em equilíbrio, o estado do sistema num instante fixado t_0 é independente do processo de saída anterior ao tempo t_0.

demonstração:

Pela reversibilidade de $N(t)$, podemos estabelecer a igualdade entre as seguintes distribuições conjuntas: saídas anteriores a t_0 e estado do sistema em t_0, com chegadas posteriores a $-t_0$ e estado do sistema em $-t_0$. Tendo em vista que as chegadas são Poisson, o processo de chegadas posterior a $-t_0$ é independente do estado do sistema nesse tempo. Dessa forma, o processo de saída até o instante t_0 e o estado do sistema em t_0 são eventos independentes. □

Comentário: O teorema 2.6 e a proposição 2.7 também valem para a fila M/M/c. De modo geral, serão válidos em sistemas de filas com chegadas Poisson e estado do sistema representado por um processo de Markov reversível. Mesmo em situações onde não é possível obter a reversibilidade, a determinação do processo reverso pode ser útil para a caracterização do sistema (ver Kelly [1979]).

2.6 Os modelos M/M/c e M/M/c/K

Nesta seção estudamos algumas variantes da fila M/M/1. Vamos iniciar pela M/M/c, onde c servidores idênticos atendem os usuários que aguardam em fila única para serem atendidos. Nesse caso, dizemos que a fila tem c canais paralelos de serviço. Em seguida, estudamos a fila M/M/c/K que tem c servidores e o limite máximo de K usuários em fila. A figura 2.4 apresenta um diagrama com

essa fila. Os usuários que não conseguem espaço para entrar no sistema constituem o fluxo denominado *overflow* (mantemos o nome em inglês por ausência de uma melhor palavra em português).

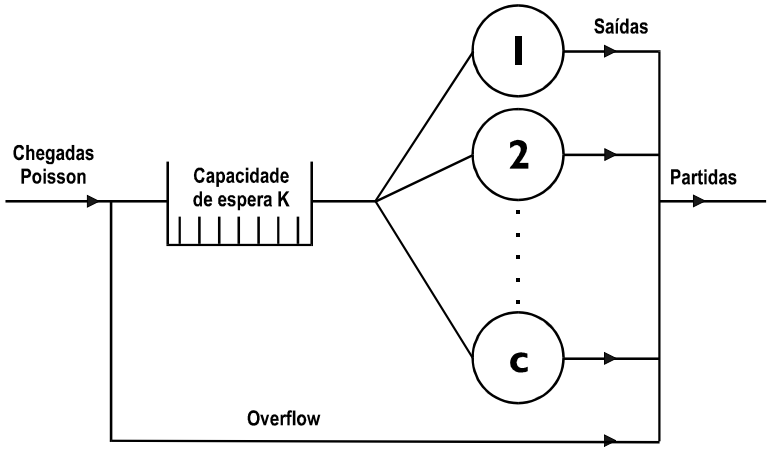

Figura 2.4: Fila M/M/c/K

Como nas seções anteriores, seja $N(t)$ o número de usuários presentes no instante t. O processo $\{N(t);\ t \geq 0\}$ continua sendo Markoviano e em especial um processo de nascimento e morte. Calculamos a distribuição estacionária para o processo de nascimento e morte, para em seguida aplicar o resultado às filas M/M/c e M/M/c/K. Na figura 2.5, as taxas do processo de nascimento e morte são apresentadas.

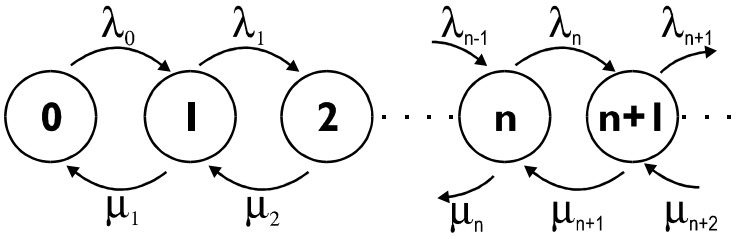

Figura 2.5: Transições no processo de nascimento e morte

As equações de equilíbrio são as seguintes:

$$\lambda_n\pi_n \ + \mu_n\pi_n = \mu_{n+1}\pi_{n+1} + \lambda_{n-1}\pi_{n-1} \ , n \geq 1; \tag{2.13a}$$

$$\lambda_0\pi_0 = \mu_1\pi_1 \,. \tag{2.13b}$$

Aplicando essas equações para $n = 0, 1, 2, \ldots$, obtemos

$$\begin{aligned}
\pi_1 &= \frac{\lambda_0}{\mu_1}\,\pi_0 \\
\pi_2 &= \frac{\lambda_1\lambda_0}{\mu_2\mu_1}\,\pi_0 \\
\pi_3 &= \frac{\lambda_2\lambda_1\lambda_0}{\mu_3\mu_2\mu_1}\pi_0 \\
&\vdots
\end{aligned}$$

A expressão sugerida seria $\pi_n = \pi_0\prod_{i=1}^{n}\frac{\lambda_{i-1}}{\mu_i}$. Supondo essa expressão válida para $n = r$, vamos verificá-la para $n = r + 1$. De (2.13a) vem

$$\begin{aligned}
\pi_{r+1} &= \frac{\lambda_r + \mu_r}{\mu_{r+1}}\pi_0\prod_{i=1}^{r}\frac{\lambda_{i-1}}{\mu_i} - \frac{\lambda_{r-1}}{\mu_{r+1}}\,\pi_0\prod_{i=1}^{r-1}\frac{\lambda_{i-1}}{\mu_i} \\
&= \frac{\pi_0\lambda_r}{\mu_{r+1}}\prod_{i=1}^{r}\frac{\lambda_{i-1}}{\mu_i} + \frac{\mu_r}{\mu_{r+1}}\,\pi_0\prod_{i=1}^{r}\frac{\lambda_{i-1}}{\mu_i} - \frac{\lambda_{r-1}}{\mu_{r+1}}\,\frac{\mu_r}{\mu_r}\,\pi_0\prod_{i=1}^{r-1}\frac{\lambda_{i-1}}{\mu_i} \\
&= \pi_0\prod_{i=1}^{r+1}\frac{\lambda_{i-1}}{\mu_i} + \frac{\mu_r}{\mu_{r+1}}\,\pi_0\prod_{i=1}^{r}\frac{\lambda_{i-1}}{\mu_i} - \frac{\mu_r}{\mu_{r+1}}\,\pi_0\prod_{i=1}^{r}\frac{\lambda_{i-1}}{\mu_i} \\
&= \pi_0\prod_{i=1}^{r+1}\frac{\lambda_{i-1}}{\mu_i}\,.
\end{aligned}$$

Dessa forma, a expressão está verificada para todo $n \geq 1$. Para calcular π_0 impomos a condição da soma das probabilidades igual a 1. Temos então,

$$\pi_0 \ = \ \Big\{\, 1 + \sum_{n=0}^{\infty}\ \prod_{i=1}^{n+1}\ \frac{\lambda_{i-1}}{\mu_i}\ \Big\}^{-1}, \tag{2.14}$$

que será válida se, e só se, a série acima convergir. Em sendo esse o caso, o processo de nascimento e morte tem distribuição estacionária dada por:

$$\pi_n = \pi_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \, , n \geq 1 \, , \tag{2.15}$$

onde π_0 é dada por (2.14).

2.6.1 Resultados para a fila M/M/c

Vamos adaptar as taxas λ_n e μ_n para o caso da fila M/M/c. Teremos $\mu_n = c\mu$ para $n \geq c$, $\mu_n = n\mu$ para $1 \leq n < c$ e ainda $\lambda_n = \lambda$, $\forall n \geq 0$. A condição para existência da distribuição estacionária torna-se então

$$\Big\{\sum_{n=0}^{\infty} \prod_{i=1}^{n+1} \frac{\lambda_{i-1}}{\mu_i} < \infty\Big\} \Leftrightarrow \Big\{\sum_{n=0}^{c-1} \frac{\lambda^n}{n!\,\mu^n} + \sum_{n=c}^{\infty} \frac{\lambda^n}{c^{n-c}\,c!\,\mu^n} < \infty\Big\} \Leftrightarrow$$
$$\Leftrightarrow \Big\{\frac{\lambda^c}{c!\,\mu^c} \sum_{n=c}^{\infty} \frac{\lambda^{n-c}}{c^{n-c}\,\mu^{n-c}} < \infty\Big\} \Leftrightarrow \Big\{\frac{\lambda}{c\mu} < 1\Big\}.$$

A condição é similar àquela obtida para a fila M/M/1. Definimos $\rho = \lambda/c\mu$, mantendo inclusive a mesma interpretação, isto é, ρ é a intensidade de tráfego. A condição $\rho < 1$ indica que, por unidade de tempo, a taxa total de serviço disponível precisa ser maior que a taxa média de chegada.

Usando as expressões (2.14) e (2.15), obtemos, após alguma álgebra,

$$\pi_0 = \Big\{\sum_{n=0}^{c-1} \frac{\lambda^n}{n!\,\mu^n} + \frac{\lambda^c}{c!\,\mu^c}\Big(\frac{c\mu}{c\mu - \lambda}\Big)\Big\}^{-1},$$

$$\pi_n = \begin{cases} \frac{\lambda^n}{n!\,\mu^n}\,\pi_0 & , 1 \leq n < c \, ; \\ \frac{\lambda^n}{c^{n-c}c!\,\mu^n}\,\pi_0 & , n \geq c \, . \end{cases}$$

Começamos o cálculo das medidas de desempenho por L_q para, depois, obter L através da fórmula de Little. Ao começar por L_q, nós evitamos ter que trabalhar com as diferentes expressões de π_n, o que seria feito no caso de calcular L diretamente pela definição de valor esperado.

$$L_q = \sum_{n=c}^{\infty}(n-c)\pi_n = \sum_{n=c}^{\infty}(n-c)\frac{\lambda^n}{c^{n-c}\,c!\,\mu^n}\;\;\pi_0 =$$
$$= \frac{\lambda^c}{c!\mu^c}\,\pi_0 \sum_{n=c}^{\infty}(n-c)\frac{\lambda^{n-c}}{c^{n-c}\mu^{n-c}} = \frac{\lambda^c}{c!\,\mu^c}\,\pi_0 \sum_{k=1}^{\infty} k\rho^k =$$
$$= \frac{\lambda^c}{c!\,\mu^c}\,\pi_0\,\rho \sum_{k=1}^{\infty} k\rho^{k-1} = \frac{\lambda^c}{c!\,\mu^c}\,\pi_0\,\rho \sum_{k=1}^{\infty}\frac{d}{d\rho}\,\rho^k =$$
$$= \frac{\lambda^c}{c!\,\mu^c}\,\pi_0\,\rho\;\frac{d}{d\rho}\sum_{k=1}^{\infty}\rho^k \;= \frac{\lambda^c}{c!\,\mu^c}\,\pi_0\,\rho\,\frac{d}{d\rho}\Big(\frac{\rho}{1-\rho}\Big) =$$
$$= \frac{\lambda^c}{c!\,\mu^c}\,\pi_0\,\rho\,\Big(\frac{1}{1-\rho}\Big)^2,$$

onde admitimos válida a troca de ordem entre somatória e derivada. Assim,

$$L_q = \left(\frac{(\lambda/\mu)^c\;\lambda\;\mu}{(c-1)!\;(c\mu-\lambda)^2}\right)\pi_0\,.$$

Partindo de L_q, obtemos W_q e daí W e L. As expressões são:

$$W_q = \frac{L_q}{\lambda} = \left(\frac{(\lambda/\mu)^c\;\mu}{(c-1)!\;(c\mu-\lambda)^2}\right)\pi_0\,,$$

$$W = \frac{1}{\mu} + W_q = \frac{1}{\mu} + \left(\frac{(\lambda/\mu)^c\;\mu}{(c-1)!\;(c\mu-\lambda)^2}\right)\pi_0\,,$$

$$L = \lambda\,W = \frac{\lambda}{\mu} + \left(\frac{(\lambda/\mu)^c\;\lambda\;\mu}{(c-1)!\;(c\mu-\lambda)^2}\right)\pi_0\,.$$

As distribuições do tempo de espera na fila e no sistema podem ser obtidas de modo análogo ao calculado para a fila M/M/1, se bem que com um pouco mais de dificuldade algébrica.

2.6.2 Resultados para a fila M/M/c/K

As taxas de serviço serão idênticas àquelas do modelo M/M/c, porém as taxas de chegadas se modificarão uma vez que o sistema não aceita mais de K na fila de espera. Teremos, então,

$$\mu_n = \begin{cases} n\,\mu & , 0 \le n < c \\ c\mu & , n \ge c \end{cases} \quad \text{e} \quad \lambda_n = \begin{cases} \lambda & , \ 0 \le n < K + c \\ 0 & , \ n \ge K + c \end{cases}$$

e a distribuição estacionária será dada por :

$$\pi_n = \begin{cases} \frac{\lambda^n}{n!\mu^n}\,\pi_0 & , 1 \le n < c\,; \\ \frac{\lambda^n}{c^{n-c}c!\,\mu^n}\,\pi_0 & , \ c \le n \le K + c\,. \end{cases}$$

Note que o espaço de estados está restrito a $0, 1, 2, \ldots, K + c$. Vamos usar, novamente, que a soma das probabilidades deve ser igual a 1 para calcular a expressão de π_0. Dessa forma,

$$\sum_{n=0}^{K+c} \pi_n = \sum_{n=0}^{c-1} \frac{\lambda^n}{n!\,\mu^n}\pi_0 + \sum_{n=c}^{K+c} \frac{\lambda^n}{c^{n-c}\,c!\,\mu^n}\pi_0 = 1\,,$$

e, dependendo de ρ ser igual ou diferente de 1 temos duas expressões distintas para π_0. Como a determinação de π_0 só envolve somas finitas, não há necessidade de condições adicionais para a convergência da somatória e, assim, $\rho = \lambda/c\mu$ não precisa ser menor que 1 nesse caso.

$$\pi_0 = \begin{cases} \left(\sum\limits_{n=0}^{c-1} \frac{\lambda^n}{n!\,\mu^n} + \frac{\lambda^c}{c!\,\mu^c}\,(K - c + 1) \right)^{-1} & , \rho \ = 1; \\ \left(\sum\limits_{n=0}^{c-1} \frac{\lambda^n}{n!\,\mu^n} + \frac{\lambda^c}{c!\,\mu^c} \frac{1-(\lambda/c\mu)^{K+c-1}}{1-\lambda/c\mu} \right)^{-1} & , \rho \ne 1\,. \end{cases}$$

Passamos agora a calcular o número médio de usuários na fila.

$$\begin{aligned} L_q \ &= \sum_{n=c}^{K+c} (n - c)\,\pi_n = \pi_0 \sum_{n=c}^{K+c} (n - c) \frac{\lambda^n}{c^{n-c}\,c!\,\mu^n} \ = \\ &= \pi_0 \frac{\lambda^{c+1}}{c!\,c\,\mu^{c+1}} \sum_{n=c}^{K+c} (n - c) \frac{\lambda^{n-c-1}}{c^{n-c-1}\,\mu^{n-c-1}} \ = \\ &= \pi_0 \frac{\lambda^{c+1}}{c!\,c\,\mu^{c+1}} \sum_{m=1}^{K} m\,\rho^{m-1} = \pi_0 \frac{\lambda^{c+1}}{c!\,c\,\mu^{c+1}} \sum_{m=1}^{K} \frac{d}{d\rho}\rho^m = \\ &= \pi_0 \frac{\lambda^{c+1}}{c!\,c\,\mu^{c+1}} \frac{d}{d\rho} \sum_{m=1}^{K} \rho^m = \pi_0 \frac{\lambda^{c+1}}{c!\,c\,\mu^{c+1}} \frac{d}{d\rho} \Big(\frac{\rho(1-\rho^K)}{1-\rho} \Big)\,; \end{aligned}$$

finalmente obtemos ,

$$L_q = \pi_0 \frac{\lambda^{c+1}}{c!\ c\ \mu^{c+1}} \Big(\frac{1 - \rho^K - K\rho^K(1-\rho)}{(1-\rho)^2} \Big)$$

Para obter outras medidas de desempenho aplicando as fórmulas de Little na fila M/M/c/K, precisamos obter a taxa efetiva de entrada. Note que algumas chegadas podem ser "perdidas" pelo sistema por não haver mais espaço na sala de espera (em inglês essa situação é referida como o usuário tendo *overflowed* o sistema). A probabilidade do sistema permitir entradas é $1 - \pi_{K+c}$ e, assim, a taxa (efetiva) de entradas será $\lambda' = \lambda\,(1 - \pi_{K+c})$. A partir de L_q, calculado acima, podemos obter sucessivamente W_q, W e L como fizemos para a fila M/M/c . Para tanto, devemos apenas substituir λ (da fórmula de Little) por λ'. Para outras medidas de desempenho e variantes da M/M/1, o leitor pode consultar Gross & Harris [1985].

2.7 Exemplos de fluxos em filas

2.7.1 Fluxos na fila M/M/1/0

Vamos descrever os vários processos que caracterizam o movimento dos fregueses nessa fila. O modelo estudado aqui poderia representar o comportamento de um telefone simples em que, ou a ligação é completada ou obtem-se o sinal de ocupado.

Na fila M/M/1/0, quando uma chegada encontra o servidor ocupado, ela sai imediatamente sem receber nenhum serviço e é considerada perdida para o sistema. A disciplina de atendimento é FCFS. Vamos explicitar nossa nomenclatura para os diversos fluxos envolvidos de modo a evitar equívocos de interpretação. O fluxo de chegada compreende os fregueses, que buscam o sistema para serviço, sendo ou não atendidos. O fluxo de entrada (*input*) é a parte das chegadas que efetivamente entra no sistema. O fluxo *overflow* é a parte das chegadas que foram impedidas de entrar no sistema. Os fregueses que deixam o sistema após serem servidos formam o fluxo de saída (*output*). O fluxo de partida será a junção das saídas e do *overflow*. Note que na fila M/M/1, discutida na seção 2.1, as chegadas e entradas representam o mesmo fluxo, o mesmo acontecendo com as partidas e saídas. A figura 2.6 representa a fila M/M/1/0 e os vários fluxos existentes.

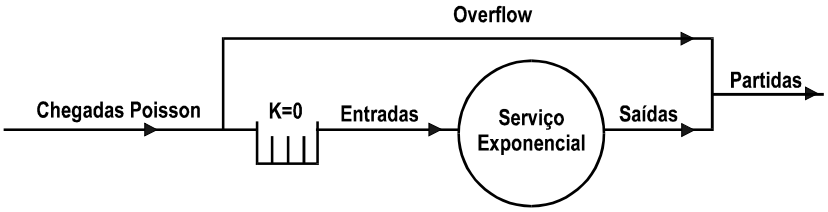

Figura 2.6: Fila M/M/1/0

Seja $N(t)$ o número de fregueses no sistema no instante t. Não é difícil verificar que $\{N(t);\ t \geq 0\}$ é um processo de Markov com espaço de estados $\{0, 1\}$. O gerador infinitesimal é

$$\Lambda = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix},$$

e a distribuição estacionária é dada por $\pi_0 = \mu/(\lambda + \mu)$ e $\pi_1 = \lambda/(\lambda + \mu)$. Pode-se verificar que $N(t)$ é um processo reversível e, dessa forma, como as chegadas formam um processo de Poisson, o processo de partida será também Poisson.

Para estudar os fluxos de saída e *overflow* vamos considerar o estado do sistema nos instantes de partida. Sejam $T_0, T_1, T_2, \ldots,$ os sucessivos instantes de partida do sistema. Denotamos por X_n o número de usuários no sistema imediatamente após T_n, isto é, $X_n = N(T_n^+)$. O processo estocástico definido por $(X, T) = \{(X_n,T_n);\ n \geq 0\}$ é um processo de renovação Markoviano com estados $\{0, 1\}$ e núcleo, na forma de transformadas de L-S, dado por:

$$\tilde{Q}(s) = \begin{bmatrix} \frac{\lambda\mu}{(\lambda+s)(\lambda+\mu+s)} & \frac{\lambda^2}{(\lambda+s)(\lambda+\mu+s)} \\ \frac{\mu}{(\lambda+\mu+s)} & \frac{\lambda}{(\lambda+\mu+s)} \end{bmatrix}$$

A matriz da cadeia de Markov imersa é obtida avaliando-se $\tilde{Q}(s)$ em $s = 0$. A distribuição estacionária coincide com aquela do processo de Markov a parâmetro contínuo.

Para obter os processos de saída e *overflow*, vamos utilizar a técnica de filtragem de processos de renovação Markoviano (ver Çinlar [1975]). Dessa forma, iniciando com o processo de partida filtramos os estados correspondentes aos processos de saída e de *overflow*. Pela simplicidade do nosso modelo, a técnica de filtragem, nesse caso, será similar ao cálculo do tempo de primeiro retorno a um estado em cadeias de Markov.

O núcleo $\widetilde{Q}(s)$ pode ser dividido em blocos representando os estados após uma saída ou *overflow*. Em geral, os blocos são matrizes quadradas cujos elementos representam as transições em uma etapa entre os conjuntos de estados. O caso atual é mais simples e as matrizes se reduzem a escalares, uma vez que temos o estado $\{0\}$ correspondendo a uma saída enquanto que o estado $\{1\}$ indica a ocorrência de *overflow*. As características de transição do processo ficam mantidas após a filtragem e cada um dos processos resultantes será também um processo de renovação Markoviano num sub-conjunto do espaço de estados inicial. Assim a saída e o *overflow* serão processos de renovação Markovianos com apenas um estado, ou seja, serão processos de renovação. Definimos $\widetilde{F}_o(s)$ e $\widetilde{F}_{ov}(s)$ como as transformadas LS da função de distribuição do intervalo entre saídas e entre *overflows*, respectivamente. Podemos verificar que valem

$$\widetilde{F}_o(s) = \frac{\lambda\mu}{(\lambda+s)(\lambda+\mu+s)} + \\ + \frac{\lambda^2}{(\lambda+s)(\lambda+\mu+s)} \sum_{n=0}^{\infty} \Big(\frac{\lambda}{(\lambda+\mu+s)} \Big)^n \frac{\mu}{(\lambda+\mu+s)} \,,$$

$$\widetilde{F}_{ov}(s) = \frac{\lambda}{(\lambda+\mu+s)} + \\ + \frac{\mu}{(\lambda+\mu+s)} \sum_{n=0}^{\infty} \Big(\frac{\lambda\mu}{(\lambda+s)(\lambda+\mu+s)} \Big)^n \frac{\lambda^2}{(\lambda+s)(\lambda+\mu+s)}$$

Após efetuar as operações indicadas, obtemos:

$$\widetilde{F}_o(s) = \frac{\lambda\mu}{(\lambda+s)(\mu+s)} \,;$$

$$\widetilde{F}_{ov}(s) = \frac{\lambda\,(\lambda+s)}{(\lambda+s)(\lambda+\mu+s) - \lambda\mu} \,.$$

Temos aqui um exemplo em que a superposição de dois processos de renovação (saídas e *overflow*) dá origem a um processo de Poisson (partidas). É interessante notar que os processos de saída e de *overflow* não são processos de Poisson seja qual for a escolha dos parâmetros λ e μ (positivos). Da teoria de processos estocásticos, uma proposição conhecida e importante assinala que a superposição de dois processos de renovação independentes produz um processo de Poisson se, e somente se, esses processos são ambos Poisson. O resultado apresentado aqui, aparentemente, contradiz essa proposição. A questão se esclarece, observando que os processos de saída e de *overflow* não são processos

independentes. A correlação assintótica entre eles é apresentada em Disney & Kiessler [1987].

2.7.2 Fluxos na fila M/M/1 com Bernoulli *feedback*

O modelo que vamos estudar consiste numa fila M/M/1 em que os usuários, após receberem atendimento, podem voltar à fila, com probabilidade p, para receber um novo serviço ou partir (sair de vez) do sistema com probabilidade $q = 1 - p$. Admitimos $p < 1$ de modo a permitir que o usuário saia do sistema em algum momento. Temos, assim, a cada fim de serviço, a repetição da decisão de receber mais serviço ou não e, dessa forma, um usuário pode efetuar vários retornos antes de deixar o sistema. Mantivemos o termo inglês *feedback* pois ele já está incorporado à língua portuguesa com o mesmo sentido da palavra original. A denominação *Bernoulli feedback* vem do fato de que a decisão de retornar ou não para mais um serviço se dá, independentemente do estado do sistema e do passado do freguês, com distribuição Bernoulli. Modelos de filas com *feedback* originaram-se de sistemas de produção, em que peças com defeitos não passavam pelo controle de qualidade e eram enviadas de volta à linha de produção. No nosso estudo vamos admitir que o usuário retorna instantaneamente ao final da fila (se ela existir). A figura 2.7 apresenta esse modelo.

Seja $N(t)$ o número de usuários no instante t. Tendo em vista que a ocorrência de *feedback* não altera o número no sistema (ele sai e imediatamente retorna), vamos introduzir uma nova variável $Y(t)$, com valores 0 ou 1, trocando de valor a cada ocorrência de um *feedback*. Essa variável, de auxílio à modelagem, é denominada na literatura de variável *flip-flop*. Dessa forma, o processo definido por $(N, Y) = \{(N(t), Y(t));\ t \geq 0\}$, com espaço de estados $\mathbb{N} \times \{0, 1\}$, será Markoviano.

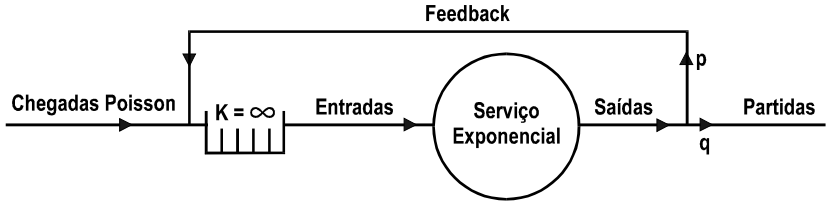

Figura 2.7: Fila M/M/1 com Bernoulli feedback

Com o intuito de aliviar a representação do gerador infinitesimal, vamos criar blocos de estados denotando os conjuntos $\{(0,0),(1,0),(2,0),\dots\}$ e $\{(0,1),(1,1),(2,1),\dots\}$ por $[0]$ e $[1]$, respectivamente. O gerador torna-se então

$$\Lambda = \begin{bmatrix} \Lambda_A & \Lambda_B \\ \Lambda_B & \Lambda_A \end{bmatrix},$$

com Λ_A e Λ_B duas matrizes quadradas de dimensão infinita dadas por

$$\Lambda_A = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & \cdots & \cdots \\ q\mu & -(\lambda+\mu) & \lambda & 0 & \cdots & \cdots \\ 0 & q\mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ \vdots & 0 & q\mu & -(\lambda+\mu) & \lambda & \cdots \\ \vdots & \vdots & 0 & \ddots & \ddots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix},$$

$$\Lambda_B = \begin{bmatrix} 0 & 0 & 0 & \cdots & \cdots & \cdots \\ 0 & p\mu & 0 & 0 & \cdots & \cdots \\ 0 & 0 & p\mu & 0 & 0 & \cdots \\ 0 & 0 & 0 & p\mu & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

A matriz Λ_A representa as transições entre blocos idênticos, ao passo que a matriz Λ_B indica as transições entre blocos diferentes.

Com a condição $\rho = \lambda/q\mu < 1$, a distribuição estacionária será

$$\pi_{(j,l)} = \frac{1}{2}(1-\rho)\,\rho^j\,,\ j \in \mathbb{N} \text{ e } l = 0,1\,.$$

Proposição 2.8 Reversibilidade do processo *(N, Y)*

O processo *(N, Y)* é um processo de Markov reversível.

demonstração:

As equações de balanço detalhado são as seguintes:

(i) $\pi_{(0,0)}\Lambda_{(0,0),(1,0)} = \pi_{(1,0)}\Lambda_{(1,0),(0,0)}$;

(ii) $\pi_{(0,1)}\Lambda_{(0,1),\,(1,1)} = \pi_{(1,1)}\Lambda_{(1,1),(0,1)}$;

(iii) $\pi_{(k,0)}\Lambda_{(k,0),\,(k-1,0)} = \pi_{(k-1,0)}\Lambda_{(k-1,0),(k,0)}$, $k \geq 2$;

(iv) $\pi_{(k,1)}\Lambda_{(k,1),\,(k-1,1)} = \pi_{(k-1,1)}\Lambda_{(k-1,1),(k,1)}$, $k \geq 2$;

(v) $\pi_{(k,0)}\,\Lambda_{(k,0),\,(k,1)} = \pi_{(k,1)}\,\Lambda_{(k,1),(k,0)}$, $k \geq 2$.

Utilizando a distribuição estacionária e o gerador, a verificação dessas equações é imediata e a proposição fica demonstrada.□

Comentário: Uma consequência importante da reversibilidade é a caracterização do processo de partida como sendo Poisson, uma vez que as chegadas formam um processo de Poisson. A verificação desse fato segue a mesma argumentação utilizada na seção 2.5.

Os processos de saída e de *feedback* do modelo M/M/1 com Bernoulli *feedback* foram estudados em Hunter [1985] (ver também Disney & Kiessler [1987]) e são processos de renovação Markovianos com espaço de estados $\{0, 1, 2, \dots\}$ para a saída e $\{1, 2, \dots\}$ para o *feedback*. Da teoria dos processos de renovação Markovianos, a distribuição do intervalo entre ocorrências sucessivas pode ser obtida através do produto $\Pi Q(t)e$ onde Π é a distribuição estacionária, $Q(t)$ é o núcleo e e é um vetor coluna de 1's. Com a mesma notação, a distribuição de dois intervalos consecutivos é dada por $\Pi Q(t_1)Q(t_2)e$. Não vamos apresentar todo o cálculo efetuado, mas apenas as expressões finais para as distribuições de um e a conjunta de dois intervalos para os dois processos. O leitor interessado poderá consultar as referências já mencionadas.

Para o processo de saída, temos

$$F_o(t) = 1 - \frac{q\mu - \lambda}{\mu - \lambda}\, e^{-\lambda t} \; - \frac{p\mu}{\mu - \lambda}\, e^{-\mu t}, t \in \mathbb{R}_+ ;$$

$$F_o(t_1, t_2) = \frac{q(1-\rho)}{(1-\rho q)}\Big\{(1 - e^{-\lambda t_1})(1 - e^{-\lambda t_2}) + \\ + p\, e^{-\lambda t_2}(e^{-\mu t_1} - e^{-\lambda t_1})(1 - e^{-\mu t_2})\Big\} + \\ + \frac{p}{(1-\rho q)}(1 - e^{-\mu t_1})(1 - e^{-\mu t_2}),\ t_1, t_2 \in \mathbb{R}_+$$

Para ser um processo de renovação, será necessário (mas não suficiente) que intervalos consecutivos sejam independentes. Assim, impondo que a conjunta

de dois intervalos seja o produto das marginais, a seguinte condição precisa ser satisfeita:

$$\frac{p\lambda}{\mu-\lambda}e^{-\lambda t_2} - \frac{p\mu}{\mu-\lambda}e^{-\mu t_2} + p\, e^{-(\lambda+\mu)t_2} = 0, \text{ para } \forall\, t_2 \in \mathbb{R}_+ .$$

Assim, para $0 \leq p < 1$, essa condição é satisfeita somente se $p = 0$. Nesse caso, a fila se torna uma M/M/1 sem *feedback* e já sabemos que seu processo de partida é de renovação (de fato um processo de Poisson). Concluimos então que, para $0 < p < 1$, o processo de saída nunca será um processo de renovação.

Vamos considerar agora o processo de *feedback*. As transformadas da distribuição de um intervalo e a conjunta de dois intervalos consecutivos são dadas por:

$$\widetilde{F}_f(s) = \frac{\mu p}{\mu p + s}\Big\{1 - \frac{\mu q s\,(1-\rho)(1-w)}{(\mu p + s)(\lambda - \lambda w + s)}\Big\},$$

onde $w = \dfrac{1}{2\lambda}\Big\{\lambda + \mu + s - \Big[(\lambda+\mu+s)^2 - 4\mu q\Big]^{1/2}\Big\}$;

$$\begin{aligned}\widetilde{F}_f(s_1,s_2) = \frac{(\mu p)^2}{(\mu p + s_1)(\mu p + s_2)}\Big\{1 - \\ - \frac{\mu q s_1(1-\rho)(1-w_1)}{(\mu p + s_1)(\lambda - \lambda w_1 + s_1)} - \frac{\mu q s_2(1-\rho)(1-w_2)}{(\mu p + s_2)(\lambda - \lambda w_2 + s_2)} + \\ + \frac{s_1 s_2(1-\rho) w_1 w_2}{(1-\rho w_1 w_2)(\lambda - \lambda w_1 + s_1)(\lambda - \lambda w_2 + s_2)}\Big\}\end{aligned}$$

com w_1 e w_2 sendo a expressão w com s substituído por s_1 e s_2, respectivamente.

Impondo-se, como antes, que a conjunta de dois intervalos seja igual ao produto das marginais, pode-se demonstrar que o processo *feedback* não tem intervalos consecutivos independentes e, assim, não pode ser um processo de renovação.

Concluindo, para o modelo M/M/1 com Bernoulli *feedback*, o processo de saída nunca é de renovação (a menos que não exista *feedback*) e, quando separado por uma variável de Bernoulli independente, produz um processo de Poisson (o processo de partida) e um outro processo que nunca será de renovação (o processo *feedback*). Outras relações entre os processos de entrada e saída e entre os processos de chegada e partida podem ser encontradas nas referências citadas acima.

2.8 EXERCÍCIOS

Vários dos exercícios propostos são adaptações de exemplos e exercícios apresentados em Murdoch [1978] (exercícios 1-7 e 11), Allen [1990] (exercício 8), Gross & Harris [1985] (exercícios 9 e 10) e Wolff [1989] (exercício 12).

1) Fregueses chegam ao caixa de uma loja com uma taxa de 5 a cada 30 minutos. O caixa atende cada freguês segundo uma distribuição exponencial com tempo médio de 4.5 minutos. Admitindo um processo de Poisson para as chegadas, pergunta-se:
a) Qual a expectativa do tempo para ser atendido pelo caixa?
b) Qual é a chance da fila exceder 5 clientes?
c) Qual é a probabilidade de, num tempo qualquer, encontrar o caixa ocupado?

2) A que taxa deve trabalhar o único funcionário de uma fotocopiadora de modo a assegurar que, com probabilidade 0.9, ninguém vai esperar mais de 12 minutos pelo atendimento. Admita chegadas e serviços exponenciais, com uma taxa de chegada de 15 clientes por hora.

3) As chegadas de clientes a um caixa eletrônico se dão segundo um processo de Poisson com taxa de 4 clientes por hora. O tempo médio de ocupação da máquina é de 6 minutos (suponha exponencial). Calcule o tamanho médio da fila e a probabilidade de haver mais de 4 pessoas esperando pelo serviço.

4) Um mecânico está sendo selecionado para consertar máquinas, cuja taxa de quebra é de 3 por hora (ocorrendo num tempo aleatório). Cada máquina fora da linha de produção custa a companhia $5 por hora. Dois candidatos se apresentaram para o emprego. O candidato A deseja receber $3 por hora e promete consertar em média 4 máquinas por hora. O candidato B trabalhará a uma taxa média de 6 máquinas por hora e pede $6 por hora de serviço. Admitindo que o tempo de reparo tem distribuição exponencial para os dois candidatos, qual deverá ser escolhido?

5) Navios chegam a um porto aleatoriamente e levam em média um dia para descarregar (tempo exponencial). Considerando uma semana de trabalho de 5 dias e 4 navios chegando por semana, qual será a média e a distribuição do tempo total de permanência neste porto?

6) Uma rede de *fastfood* tem como política que o cliente deverá esperar, em média, somente 2 minutos pelo seu pedido. Se os clientes chegam aleatoriamente a uma taxa de 20 por hora, e o serviço se realiza segundo uma distribuição exponencial com média de 2.2 minutos, qual é o tempo real de espera? Qual

deveria ser a média do tempo de serviço para satisfazer a política prevista pela companhia?

7) Uma clínica é procurada por uma média de 3 pacientes por hora. O médico de plantão leva 15 minutos, em média, em cada consulta. Admitindo-se chegadas e tempo de consulta como tendo distribuição exponencial, pergunta-se:
a) Que proporção do tempo, o médico está sem paciente?
b) Quantos pacientes em média esperam por uma consulta?
c) Qual é a probabilidade da fila para consulta ter mais de 3 pacientes?
d) Qual é o tempo médio que cada paciente permanece na clínica?
e) Qual é a probabilidade de um paciente esperar mais de 1 hora pela consulta?

8) Uma companhia tem um serviço de atendimento ao usuário via telefone. Cada linha tem um plantonista que atende dúvidas, reclamações e sugestões dos consumidores e, se todas as linhas estiverem ocupadas, o usuário precisará ligar de novo, pois o sistema não dá possibilidade de espera. Chegam em média 105 chamadas por hora que têm duração média de 4 minutos. Vamos admitir que as chegadas e a duração da chamada sigam a distribuição exponencial. Quantas linhas precisam estar disponíveis para que a probabilidade de obter um sinal de ocupado não seja superior a 0.005. Qual seria o desempenho se tivessemos 10 linhas?

9) Mostre que, com respeito ao tamanho médio no sistema, temos as seguintes situações:
a) A fila M/M/1 é sempre melhor que a fila M/M/2, se ambas tem o mesmo ρ.
b) O sistema M/M/2 é sempre melhor que o sistema com duas filas M/M/1 independentes, cada uma recebendo metade das chegadas, desde que os dois sistemas tenham a mesma taxa de serviço por servidor.

10) Mostre que, na fila M/M/2, a probabilidade de um servidor, escolhido ao acaso, estar ocupado é $\lambda/2\mu$.

11) Um processo de produção consiste de duas máquinas A e B, operando sequencialmente. A saída da máquina A é Poisson com taxa de 4 por hora e a máquina B processa cada item em tempo exponencial com taxa 5 por hora. O sistema é considerado congestionado se mais de dois itens estão esperando serviço na máquina B.
a) Em regime estacionário, qual é a fração do tempo que o sistema está congestionado?
b) Suponha que a máquina A pára de funcionar toda vez que o sistema está congestionado. Recalcule o item a) e compare as respostas.

12) Verifique que, na fila M/M/1, o processo $\{N_q(t) : t \geq 0\}$, indicando o número de usuários esperando em fila, não é um processo de Markov.

13) No modelo M/M/c/0 a probabilidade estacionária, de todos os servidores estarem ocupados, é denominada *fórmula B de Erlang*. Determine sua expressão.

14) Mostre, através de equivalência, que o processo de partidas da fila M/M/1/0 é um processo de Poisson.

15) Determine o tempo médio entre *overflows* para a fila M/M/1/0.

16) Para a fila M/M/1 com Bernoulli *feedback*, determine o núcleo do processo de renovação Markoviano imerso nos instantes de saídas.

17) Utilizando os resultados apresentados na seção 2.7 para o modelo M/M/1 com Bernoulli *feedback*, determine os tempos médios entre saídas e entre *feedbacks*. Para $\lambda = 1$ e $\mu = 0.8$, represente, num mesmo gráfico, esses tempos médios em função de p (a probabilidade de *feedback*).

18) Verifique se o número de usuários na fila M/M/1/1 é um processo de Markov reversível.

Capítulo 3

A fila M/G/1 e suas variantes

3.1 Introdução

Vamos, neste capítulo, apresentar e discutir as principais propriedades da fila M/G/1. Este modelo se diferencia daquele estudado no capítulo anterior pela possibilidade do serviço se realizar de acordo com uma distribuição geral, sem nenhuma propriedade especial, exceto a independência entre sucessivos serviços e entre serviços e chegadas. Seguindo a notação apresentada no Capítulo 1, a fila M/G/1 tem chegadas seguindo um processo de Poisson (λ), sala de espera de capacidade ilimitada, atendimento em ordem de chegada e um servidor com tempo de serviço S, o qual segue uma distribuição geral B.

Para caracterizar o estado da fila, seja $N(t)$ o número de clientes no sistema no instante t. Precisamos de uma estrutura probabilística para estudar esse processo. Considere, por exemplo, que $N(t) = 2$ para algum t. Isto é, o sistema tem um cliente sendo atendido e um esperando na fila. Como será a transição do processo $N(t)$? O tempo de espera para uma nova chegada, que levaria ao estado 3, é exponencial com parâmetro λ pela falta de memória dessa distribuição. Lembremos que o processo de chegadas é Poisson (λ) e, assim, o intervalo entre chegadas é exponencial. No entanto, o tempo para completar o serviço, que levaria o processo ao estado 1, não pode ser determinado pois a distribuição não tem, necessariamente, a propriedade de falta de mémoria. Dessa forma, a informação disponível de que $N(t) = 2$ não é suficiente para determinar as probabilidades de transição e, consequentemente, $N(t)$ não é um processo de Markov (como foi no caso da fila M/M/1). A alternativa que vamos buscar é encontrar alguns pontos de tempo especiais de tal forma que, a partir desses pontos, possamos ter idéia do comportamento do sistema. Se olharmos o número de clientes nos instantes de partida do sistema, evitaremos a dificuldade de tratar com partes do tempo de serviço. O processo que vamos trabalhar é chamado de processo imerso nos instantes de partidas, que, aqui, é sinônimo de saídas.

Sejam $T_0, T_1, T_2, ...$ os sucessivos instantes de fim de serviço, correspondendo às saídas da fila. Definimos X_n como o número de usuários na fila imediatamente após T_n, isto é, $X_n = N(T_n^+)$. Podemos verificar que o

processo estocástico $(X,T) = \{(X_n,T_n); n \geq 0\}$ é um processo de renovação Markoviano. A argumentação é semelhante à utilizada na seção 2.3. Note que

$$X_n = \begin{cases} X_{n-1} + 1 - A_n & , X_{n-1} \geq 1; \\ A_n & , X_{n-1} = 0, \end{cases} \tag{3.1}$$

onde A_n é o número de chegadas durante o serviço do n-ésimo cliente. Pela expressão acima, X_n depende de X_{n-1} e de quantas chegadas ocorreram desde a última saída, o que é função de λ e da duração do intervalo $T_n - T_{n-1}$. Esse intervalo será dependente apenas de X_{n-1}. Se $X_{n-1} > 0$, esse tempo tem distribuição B, correspondendo a um serviço completado. Por outro lado, se $X_{n-1} = 0$, o intervalo entre saídas será a soma do tempo de espera para uma chegada com o tempo de serviço. Concluimos que o conhecimento de X_{n-1} é suficiente para caracterizar a probabilidade de transição de (X,T) e, portanto, ele é um processo de renovação Markoviano.

Vamos calcular os elementos da matriz $Q(t)$, que representa o núcleo de transição desse processo. Temos dois casos a considerar:

caso 1: transição $0 \rightarrow j$, $j \geq 0$.

O seguinte evento deve ocorrer:

$\Big\{$(primeira chegada) + (j chegadas antes do fim de serviço) $\leq t\Big\}$.

Então,

$$Q_{0j}(t) = \int_0^t \int_0^{t-s} \lambda\, e^{-\lambda s} \frac{(\lambda y)^j e^{-\lambda y}}{j!}\, dB(y)\, ds\,, j \geq 0.$$

caso 2: transição $i \rightarrow j$, $j \geq i-1$, $i > 0$.

O evento será:

$\Big\{$(tempo para $j - i + 1$ chegadas antes do fim de serviço) $\leq t\Big\}$.

Então,

$$Q_{ij}(t) = \int_0^t \frac{(\lambda y)^{j-i+1} e^{-\lambda y}}{(j-i+1)!} dB(y)\,, i > 0, j \geq i-1\,.$$

Para auxiliar a representação do núcleo, vamos definir duas funções $g_n(.)$ e $h_n(.)$, com $n \geq 0$, da seguinte forma:

$$g_n(t) = \int_0^t \frac{(\lambda y)^n e^{-\lambda y}}{n!}\, dB(y)\,, n \geq 0;$$

$$h_n(t) = \int_0^t \lambda\, e^{-\lambda y} g_n(t-y)dy\ ,\ n \geq 0.$$

Então,

$$Q(t) = \begin{bmatrix} h_0(t) & h_1(t) & h_2(t) & \cdots \\ g_0(t) & g_1(t) & g_2(t) & \cdots \\ 0 & g_0(t) & g_1(t) & \cdots \\ 0 & 0 & g_0(t) & \cdots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix},$$

Para determinar a matriz de transição da cadeia imersa, tomamos limite de $Q(t)$ quando t vai ao infinito. Note que, nesse caso, h_n e g_n tornam-se iguais. Temos então,

$$P = \begin{bmatrix} g_0 & g_1 & g_2 & \cdots \\ g_0 & g_1 & g_2 & \cdots \\ 0 & g_0 & g_1 & \cdots \\ 0 & 0 & g_0 & \cdots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix}$$

com $$g_n = \int_0^\infty \frac{(\lambda y)^n e^{-\lambda y}}{n!} dB(y)\ ,\ n \geq 0.$$

A distribuição estacionária da cadeia imersa nos instantes de partida (representada com o superescrito d de *departures*) é a solução do sistema de equações $\mathit{\Pi}^d = \mathit{\Pi}^d P$ e $\mathit{\Pi}^d e = 1$, que pode ser escrito da seguinte forma:

$$\begin{cases} \pi_i^d = \pi_0^d g_i + \sum_{j=1}^{i+1} \pi_j^d g_{i-j+1}\ ,\ i = 0,\ 1, 2, ...; & \text{(3.2a)} \\ \sum_{i=0}^{\infty} \pi_i^d = 1. & \text{(3.2b)} \end{cases}$$

Exceto em casos especiais, a solução analítica do sistema acima não tem uma forma fechada. Soluções numéricas podem ser obtidas com o "adequado" truncamento do espaço de estados. Com o objetivo de calcular algumas medidas de desempenho, vamos obter a função geradora de probabilidade da distribuição estacionária. Por definição, a função geradora é dada por

$$\Phi(z) = \sum_{i=0}^{\infty} \pi_i^d \, z^i \, , \text{para} \, |z| \leq 1.$$

Para auxiliar o cálculo, vamos também definir uma outra função geradora, relativa à probabilidade de i chegadas durante um serviço. Temos,

$$\Upsilon(z) = \sum_{i=0}^{\infty} g_i \, z^i \, , |z| \leq 1.$$

Multiplicando a expressão (3.2a) por z^i e somando para $i \geq 0$, temos

$$\sum_{i=0}^{\infty} \pi_i^d \, z^i = \sum_{i=0}^{\infty} \pi_0^d g_i \, z^i + \sum_{i=0}^{\infty} z^i \sum_{j=1}^{i+1} \pi_j^d \, g_{i-j+1}$$

$$\Phi(z) = \pi_0^d \Upsilon(z) + \sum_{j=1}^{\infty} \sum_{i=j-1}^{\infty} \pi_j^d \, g_{i-j+1} \, z^i$$

$$\Phi(z) = \pi_0^d \Upsilon(z) + \frac{1}{z} \sum_{j=1}^{\infty} \pi_j^d \, z^j \sum_{i=j-1}^{\infty} g_{i-j+1} \, z^{i-j+1}$$

$$\Phi(z) = \pi_0^d \Upsilon(z) + \frac{1}{z} (\Phi(z) - \pi_0^d) \Upsilon(z) \, ,$$

então,

$$\Phi(z) = \frac{\pi_0^d \Upsilon(z)(1-z)}{\Upsilon(z) - z} \, . \tag{3.3}$$

Para obter a expressão da função geradora precisamos determinar π_0^d. Para tal, notamos que $\Phi(1) = 1$ e $\Upsilon(1) = 1$, condições que decorrem da soma de probabilidades ser igual a 1. Com o valor de z substituído por 1, a expressão para $\Phi(z)$ dá uma indeterminação do tipo 0 sobre 0. Dessa forma, será necessário aplicar a regra de L'Hopital, isto é, derivamos o numerador e o denominador do lado direito de (3.3) e depois efetuamos a substituição de z por 1. Temos

$$1 = \left. \frac{\pi_0^d [\Upsilon^{'}(z) \, (1-z) + \Upsilon(z) \, (-1)]}{\Upsilon^{'}(z) - 1} \right|_{z=1} , \tag{3.4}$$

onde $\Upsilon'(z) = \sum_{i=0}^{\infty} i\, g_i\, z^{i-1}$ e, portanto

$$\Upsilon'(1) = \sum_{i=0}^{\infty} i \int_0^{\infty} \frac{(\lambda y)^i e^{-\lambda y}}{i!}\, dB(y) = \int_0^{\infty} \sum_{i=0}^{\infty} i\, \frac{(\lambda y)^i e^{-\lambda y}}{i!}\, dB(y)$$
$$= \int_0^{\infty} \lambda t\, dB(y) = \lambda \int_0^{\infty} t\, dB(y) = \lambda\, E(S).$$

Como antes, definimos a intensidade de tráfego do sistema como sendo $\rho = \lambda E(S)$. Com a substituição do valor de $\Upsilon'(1)$ em (3.4) obtemos $\pi_0^d = 1 - \rho$, o que indica a existência de distribuição estacionária se e só se $\rho \leq 1$. Então,

$$\Phi(z) = \frac{[1 - \lambda\, E(S)]\, \Upsilon(z)\, (1 - z)}{\Upsilon(z) - z}\,. \tag{3.5}$$

Pela definição de $\Upsilon(z)$ temos ainda o seguinte resultado:

$$\Upsilon(z) = \sum_{i=0}^{\infty} z^i \int_0^{\infty} \frac{(\lambda y)^i e^{-\lambda y}}{i!}\, dB(y) = \int_0^{\infty} \sum_{i=0}^{\infty} \frac{(\lambda y z)^i e^{-\lambda y}}{i!}\, dB(y)$$
$$= \int_0^{\infty} e^{-\lambda y} e^{\lambda y z}\, dB(y) = \tilde{B}(\lambda - \lambda z),$$

indicando que $\Upsilon(z)$ pode ser escrita em função da transformada de Laplace-Stieltjes da distribuição do tempo de serviço. Apesar da utilidade deste resultado, mantemos $\Phi(z)$ como dada pela expressão (3.5), para não carregar a notação com duas transformadas diferentes.

Utilizando as propriedades da função geradora podemos obter os momentos da distribuição (isto será feito na próxima seção) e, em casos especiais, através da inversão da função geradora, determinar as probabilidades estacionárias.

Exemplo 3.1: (Kleinrock [1975])

Considere a fila M/H_2/1 onde a distribuição de serviço é uma hiper-exponencial dada por:

$$B(t) = \frac{1}{4}\, \lambda\, e^{-\lambda t} + \frac{3}{4}\, (2\lambda)\, e^{-2\lambda t}\,,\ t \geq 0.$$

O serviço nessa fila se comporta como se existissem dois servidores exponenciais em paralelo com parâmetros λ e 2λ, onde λ é também a taxa de chegada do

processo de Poisson. A média do tempo de serviço fica sendo $E(S) = 5/8\lambda$ e $\rho = 5/8$. Calculando a transformada de $B(t)$, temos

$$\tilde{B}(s) = \frac{1}{4}\,\frac{\lambda}{(\lambda+s)} + \frac{3}{4}\,\frac{2\lambda}{(2\lambda+s)} = \frac{7\lambda s + 8\lambda^2}{4(\lambda+s)(2\lambda+s)},\ \text{com Re}(s) > -\lambda.$$

Portanto,

$$\Upsilon(z) = \tilde{B}(\lambda - \lambda z) = \frac{8 + 7(1-z)}{4(2-z)(3-z)}\ .$$

Substituindo em (3.5), vem

$$\Phi(z) = \frac{(3/8)(1-z)\,[8+7(1-z)]}{8+7(1-z)-4z(2-z)(3-z)}.$$

Expandindo em frações parciais, obtemos

$$\Phi(z) = \frac{3}{8}\Big(\frac{1/4}{1-(2/5)z} + \frac{3/4}{1-(2/3)z}\Big),$$

que após inversão produz

$$\pi_n = \frac{3}{32}\Big(\frac{2}{5}\Big)^n + \frac{9}{32}\Big(\frac{2}{3}\Big)^n,\ n = 0, 1, \cdots\ .$$

A inversão da função geradora não foi difícil nesse caso, sendo uma simples aplicação das tabelas de inversão encontradas em vários textos (por exemplo, Kleinrock [1975]).□

No começo do capítulo, quando iniciamos o estudo do modelo M/G/1, foi necessário escolher momentos especiais de observação do sistema. Escolhemos os instantes de fim de serviço e uma questão a ser indagada é, como relacionar a distribuição estacionária obtida, nesses instantes especiais, com a distribuição em outros instantes? Por exemplo, o que dizer da distribuição em tempo contínuo ou imersa nos instantes de chegadas? O resultado *Pasta* (ver seção 2.2) pode ser aplicado ao modelo M/G/1 e, portanto, a distribuição estacionária vista por um observador externo (tempo contínuo) é igual àquela vista por um usuário ao chegar. A próxima proposição relaciona essas distribuições com a distribuição imersa nos instantes de partida que foi obtida nesta seção.

Proposição 3.1: Igualdade entre distribuições estacionárias

No modelo M/G/1 vale a igualdade entre as distribuições estacionárias em tempo contínuo e as imersas nos instantes de chegadas e partidas. Isto é,

$$\Pi = \Pi^a = \Pi^d \,.$$

demonstração:

Tendo em vista que as chegadas formam um processo de Poisson, o resultado *Pasta* implica $\Pi = \Pi^a$ e, assim, é suficiente provar a igualdade entre as distribuições estacionárias imersas nos instantes de chegada e partida. Vamos admitir a existência do limite das respectivas distribuições e também que $N(0) = 0$ (no instante zero o sistema está vazio). Como precisamos trabalhar com vários instantes de tempo, introduzimos uma notação mais detalhada. Sejam T_n^a o instante da n-ésima chegada e T_n^d o instante da n-ésima saída. Definimos $N_n^a = N(T_n^{a-})$ e $N_n^d = N(T_n^{d+})$, ou seja, N_n^a e N_n^d são, respectivamente, o número no sistema imediatamente antes de T_n^a e após T_n^d.

Vamos verificar que $\{N_{n+k+1}^a \leq k\} \Leftrightarrow \{N_n^d \leq k\}$:

$i)$ $\{N_n^d \leq k\} \Rightarrow \{N_{n+k+1}^a \leq k\}$;

Devemos mostrar que, se o número de usuários no sistema em T_n^{d+} for menor ou igual a k, então o $(n+k+1)$-ésimo usuário ao chegar encontra no máximo k no sistema.

Suponha que $N_n^d = j$, $(j \leq k)$. Portanto, existem $(n+j)$ chegadas antes do instante T_n^d. O usuário que chega em T_{n+j+1}^a precisa encontrar no máximo j presentes no sistema (o máximo j acontece quando o primeiro evento após T_n^d é uma chegada). Assim, $N_{n+j+1}^a \leq j$. O resultado se verifica tomando $j = k$.

$ii)$ $\{N_{n+k+1}^a \leq k\} \Rightarrow \{N_n^d \leq k\}$;

Devemos verificar que, se o número de usuários em T_{n+k+1}^{a-} for menor ou igual a k, então o n-ésimo usuário ao sair deixa no máximo k usuários presentes.

Suponha que $N_{n+k+1}^a = j$, $(j \leq k)$. Precedendo o instante de chegada T_{n+k+1}^a, existiram um total de $n+k$ chegadas das quais j ainda estão no sistema e portanto $(n+k-j)$ usuários já partiram. Isto é, T_{n+k-j}^d é o último instante de partida precedendo T_{n+k+1}^a. Assim, N_{n+k-j}^d será no máximo a diferença entre $(n+k)$ e $(n+k-j)$ e temos então $N_{n+k-j}^d \leq j$ (o máximo aqui corresponde a nenhuma chegada entre T_{n+k-j}^{d+} e T_{n+k+1}^{a-}). Tomando $j = k$, temos $N_n^d \leq k$ e a implicação está demonstrada.

Portanto, vale a igualdade entre os eventos $\{N^a_{n+k+1} \leq k\}$ e $\{N^d_n \leq k\}$. Aplicando probabilidade e passando ao limite temos, para todo $k \geq 0$:

$$P(N^a_{n+k+1} \leq k) = P(N^d_n \leq k)$$
$$\lim_{n\to\infty} P(N^a_{n+k+1} \leq k) = \lim_{n\to\infty} P(N^d_n \leq k)$$
$$\lim_{n\to\infty} P(N^a_n \leq k) = \lim_{n\to\infty} P(N^d_n \leq k)$$
$$\lim_{n\to\infty} P(N(T^{a-}_n) \leq k) = \lim_{n\to\infty} P(N(T^{d+}_n) \leq k)$$

Supondo que a distribuição limite coincide com a distribuição estacionária, concluimos que $\pi^a_k = \pi^d_k$, para todo $k \geq 0$, e a proposição está provada.□

Comentário: A prova acima é uma adaptação daquela apresentada em Cooper [1981] e é válida para outros modelos cujas transições de estado são saltos de tamanho 1 para mais ou menos.

3.2 Medidas de Desempenho

Vamos calcular algumas medidas de desempenho para o modelo M/G/1. Inicialmente, notamos que as fórmulas de Little continuam válidas e vamos utilizá-las para deduzir alguns resultados.

O valor esperado do número de usuários no sistema, nos instantes imediatamente após uma partida, pode ser calculado através da função geradora de probabilidades, dada pela expressão (3.5). Temos,

$$L = \Phi'(z)\big|_{z=1} = \frac{1}{(\Upsilon(z)-z)^2}(1-\rho)$$
$$\Big\{(\Upsilon(z)-z)[\Upsilon'(z)\,(1-z) - \Upsilon(z)] - \Upsilon(z)\,(1-z)(\Upsilon'(z)-1)\Big\}\Big|_{z=1}.$$

A expressão acima produz uma indeterminação quando $z = 1$ e vamos usar a regra de L'Hopital, que nesse caso requer duas derivações. Temos,

$$L = \frac{1}{2(\Upsilon(z)-z)(\Upsilon'(z)-1)}(1-\rho)$$
$$\Big\{(\Upsilon(z)-z)\,[\Upsilon''(z)\,(1-z) - 2\Upsilon'(z)] - \Upsilon(z)\,(1-z)\Upsilon''(z)\Big\}\Big|_{z=1}$$

$$= \frac{1}{2[(\Upsilon'(z)-1)(\Upsilon'(z)-1)+(\Upsilon(z)-z)\Upsilon''(z)]}(1-\rho)$$
$$\Big\{(\Upsilon'(z)-1)[\Upsilon''(z)\,(1-z)-2\Upsilon'(z)]+[\Upsilon'''(z)\,(1-z)-3\Upsilon''(z)]$$
$$(\Upsilon(z)-z)\;-[\Upsilon'(z)\,(1-z)\Upsilon''(z)]+\Upsilon(z)\Upsilon''(z)-\Upsilon(z)\,(1-z)\Upsilon'''(z)\Big\}\Big|_{z=1},$$

Substituindo z por 1 e lembrando que $\Upsilon(1)=1$ e $\Upsilon'(1)=\rho$ vem

$$L=\frac{[2(1-\rho)\rho+\Upsilon''(1)]}{2(1-\rho)}\,.$$

Falta ainda determinar o valor de $\Upsilon''(1)$:

$$\Upsilon''(1)=\sum_{i=0}^{\infty} i(i-1)g_i\;z^{i-2}\Big|_{z=1}=\sum_{i=0}^{\infty} i(i-1)g_i\;=$$
$$=\int_0^{\infty}\sum_{i=0}^{\infty} i(i-1)\,\frac{(\lambda y)^i e^{-\lambda y}}{i!}\;dB(y)=\rho^2+\lambda^2\sigma_S^2\,,$$

onde σ_S^2 é a variância do tempo de serviço S. Finalmente obtemos

$$L=\rho+\frac{\rho^2+\lambda^2\sigma_S^2}{2(1-\rho)} \tag{3.6}$$

A expressão acima é conhecida como fórmula de Pollaczek-Khintchine.

Antes de deduzir outras medidas de desempenho, vamos calcular novamente a expressão de L utilizando um argumento direto. Com o uso de uma variável auxiliar U, reescrevemos (3.1) em uma só equação da seguinte forma:

$$X_n=X_{n-1}+U_{n-1}-A_n \tag{3.7}$$

com
$$U_{n-1}=\begin{cases}1 & ,X_{n-1}\geq 1;\\ 0 & ,X_{n-1}=0.\end{cases}$$

Em regime estacionário, $E(X_n)=E(X_{n-1})=L$ e portanto podemos desprezar o subescrito n, considerando as quantidades envolvidas como referentes a um usuário genérico do sistema. A aplicação do valor esperado nos dois lados de (3.7) fornece

$$L = L - E(U_{n-1}) + E(A_n) \Rightarrow E(U_{n-1}) = E(A_n) = E(A)\,,$$

onde A representa o número de chegadas durante a realização de um serviço (genérico). Relembrando que S denota o tempo de serviço e utilizando o cálculo de $\Upsilon^{'}(1)$ desenvolvido na seção 3.1, temos

$$E(A) = \int_0^\infty E(A|S=t)dB(t) = \int_0^\infty \sum_{i=0}^{\infty} i\ \frac{(\lambda y)^i e^{-\lambda y}}{i!} dB(t) = \lambda E(S)\ .$$

Elevando ao quadrado os dois lados de (3.7), vem

$$X_n^2 = X_{n-1}^2 + U_{n-1}^2 + A_{n-1}^2 + 2(X_{n-1}A_{n-1} - X_{n-1}U_{n-1} - A_{n-1}U_{n-1}).$$

Vamos aplicar valor esperado nos dois lados da igualdade acima. Para tal, note que, em regime estacionário, $E(X_n^2) = E(X_{n-1}^2)$ e, como decorrência da definição de U, vale $U_{n-1}^2 = U_{n-1}$ e $X_{n-1}U_{n-1} = X_{n-1}$. Utilizaremos ainda a independência entre A_{n-1} e X_{n-1}, uma vez que as chegadas não dependem do estado do sistema. Da mesma forma, A_{n-1} e U_{n-1} são independentes. Então

$$\begin{aligned}
0 &= E(U_{n-1}^2) + E(A_{n-1}^2) + 2E(X_{n-1}A_{n-1} - X_{n-1}U_{n-1} - A_{n-1}U_{n-1}) \\
0 &= E(U_{n-1}) + E(A_{n-1}^2) + 2[E(X_{n-1})E(A_{n-1}) - \\
&\qquad\qquad E(X_{n-1}) - E(A_{n-1})E(U_{n-1})] \\
0 &= \rho + E(A_{n-1}^2) + 2[L\rho - L - \rho(1-\rho)].
\end{aligned}$$

Removendo o subescrito (por estarmos em regime estacionário), precisamos obter o valor de $E(A^2)$ para completar o cálculo da fórmula de Pollaczek-Khintchine (já apresentada em (3.6)). Como A representa o número de chegadas durante um serviço genérico (médio), temos

$$\begin{aligned}
E(A^2) &= Var(A) + E^2(A) \\
&= E[Var(A|S)] + Var[E(A|S)] + E^2(A) \\
&= E(\lambda S) + Var(\lambda S) + \rho^2 \\
&= \rho + \lambda^2\sigma_S^2 + \rho^2.
\end{aligned}$$

Conhecendo a expressão de L dada por (3.6), podemos aplicar a fórmula de Little para obter a média do tempo de permanência no sistema. Assim, de $L = \lambda W$, temos,

$$W = E(S) + \frac{\rho\, E(S) + \lambda\, \sigma_S^2}{2(1-\rho)} \tag{3.8}$$

Ainda pelas fórmulas de Litte segue que:

$$W_q = \frac{\rho\, E(S) + \lambda\, \sigma_S^2}{2(1-\rho)} \tag{3.9}$$

$$L_q = \frac{\rho^2 + \lambda^2 \sigma_S^2}{2(1-\rho)} \tag{3.10}$$

Exemplo 3.2:

No exemplo 3.1 calculamos a distribuição estacionária da fila M/H_2/1. Vamos calcular agora algumas medidas de desempenho. Inicialmente da distribuição do tempo de serviço vem $\sigma_S^2 = 31/(64\lambda^2)$. Então, de (3.6) obtemos $L = 1.79$ o qual, quando comparado com o valor 1.66 da M/M/1 com mesmo ρ, indica um pequeno acréscimo no número médio de usuários presentes no sistema. A razão do acréscimo se deve ao aumento do coeficiente de variação do tempo de serviço de 1, da exponencial, para 31/25 no caso da hiperexponencial. Outras medidas de desempenho podem ser calculadas sem dificuldade.□

Devido à complexidade de cálculo, não vamos apresentar outros resultados sobre medidas de desempenho. Em particular, o cálculo da função de distribuição do tempo de espera é bem mais complicado nesse caso pois envolve a distribuição do tempo de serviço residual. Lembramos que esse é o tempo gasto pelo servidor para completar o serviço daquele usuário, que está sendo atendido no instante da chegada do usuário típico. Para esse e outros resultados, o leitor interessado poderá consultar Gross & Harris [1985] ou Cooper [1981].

3.3 O modelo M/G/1/K

Vamos considerar nesta seção a restrição na capacidade de espera. Na notação do Capítulo 1, K indica o número de usuários que podem estar na fila, esperando por serviço. Como antes, os usuários que não conseguem entrar no sistema são perdidos e constituem o fluxo de *overflow* do modelo. Um processo de renovação Markoviano imerso nos instantes de saída será usado como estrutura de análise dessa fila. Note que, apesar do modelo permitir que $K+1$ usuários

estejam presentes, ao observarmos o número no sistema em momentos imediatamente após uma saída, nunca teremos $K+1$ fregueses nesses instantes. Conforme já fizemos na seção 2.6 do capítulo anterior, precisamos chamar a atenção para a distinção entre saídas e partidas. As saídas são o fluxo que passa pelo serviço, enquanto que as partidas são a superposição das saídas e do *overflow*.

Seja $(X^o, T^o) = \{(X_n^o, T_n^o); n \geq 0\}$ o processo de renovação Markoviano, com espaço de estados $\{0, 1, 2, ..., K\}$, representando o número no sistema nos instantes imediatamente após uma saída (onde o superescrito o se refere à palavra em inglês *output*). Apresentamos, a seguir, o núcleo de transição do processo, em que a última coluna incorpora todas as probabilidades de chegadas que encontram o sistema na sua capacidade total.

$$Q^o(t) = \begin{bmatrix} h_0(t) & h_1(t) & h_2(t) & \cdots & h_{K-1}(t) & h_K^*(t) \\ g_0(t) & g_1(t) & g_2(t) & \cdots & g_{K-1}(t) & g_K^*(t) \\ 0 & g_0(t) & g_1(t) & \cdots & g_{K-2}(t) & g_{K-1}^*(t) \\ 0 & 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & g_0(t) & g_1^*(t) \end{bmatrix}, \quad (3.21)$$

onde as expressões de g_n e h_n, já apresentadas na seção 3.1, são

$$g_n(t) = \int_0^t \frac{(\lambda y)^n e^{-\lambda y}}{n!}\, dB(y)\ , n \geq 0,$$

$$h_n(t) = \int_0^t \lambda\ e^{-\lambda y} g_n(t-y) dy\ , n \geq 0,$$

e, ainda,

$$h_n^*(t) = \sum_{i=n}^{\infty} h_i(t)\ ,$$

$$g_n^*(t) = \sum_{i=n}^{\infty} g_i(t),$$

representam a chegada de pelo menos n usuários ao sistema. Tomando o limite de $Q^o(t)$, quando t vai ao infinito, obtemos a cadeia imersa

$$P = \begin{bmatrix} g_0 & g_1 & g_2 & \cdots & g_{K-1} & g_K^* \\ g_0 & g_1 & g_2 & \cdots & g_{K-1} & g_K^* \\ 0 & g_0 & g_1 & \cdots & g_{K-2} & g_{K-1}^* \\ 0 & 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & g_0 & g_1^* \end{bmatrix},$$

onde, como antes, g_n é dada por

$$g_n = \int_0^\infty \frac{(\lambda y)^n e^{-\lambda y}}{n!} dB(y)\ ,\ n \geq 0\ ,$$

e, também, $$g_n^* = \sum_{i=n}^{\infty} g_{\text{i}}\ .$$

A distribuição estacionária, representada por Π^o, é obtida resolvendo o sistema de equações:

$$\begin{cases} \pi_i^o = \pi_0^o g_i + \sum_{j=1}^{i+1} \pi_j^o g_{i-j+1}\ , i = 0,\ 1, 2, ...K-1; & \text{(3.11a)} \\ \sum_{i=1}^{K} \pi_i^o = 1, & \text{(3.11b)} \end{cases}$$

onde desprezamos a última coluna de $\Pi^o = \Pi^o\, P$. As K equações de (3.11a) coincidem com as K primeiras equações de (3.2a) e assim a solução aqui será proporcional à obtida naquele sistema.

Admitindo que Π^o foi determinado, vamos estudar as relações entre as distribuições estacionárias imersas em vários instantes. Lembramos que, devido à capacidade limitada da sala de espera, precisamos distinguir entre chegadas, que receberão serviço (chamadas aqui de entradas), daquelas que farão o *overflow* ao sistema e não receberão atendimento. Sejam Π^a e Π^i as distribuições estacionárias imersas, respectivamente, nos instantes imediatamente anteriores às chegadas e entradas e, como antes, seja Π a distribuição estacionária em tempo contínuo.

A mesma demonstração, utilizada na proposição 3.1, poderia ser adaptada para provar a igualdade entre as distribuições estacionárias imersas nos instantes de entrada e de saída. Assim, vamos admitir $\Pi^i = \Pi^o$.

Como o modelo M/G/1/K tem chegadas Poisson, segue do resultado *Pasta* que $\Pi^a = \Pi$. Note que, como os processos de chegada e entrada têm espaços de estados diferentes, precisamos necessariamente ter $\Pi^i \neq \Pi^a$. Entretanto, todo instante de entrada é também um instante de chegada e assim podemos estabelecer uma relação de proporcionalidade entre elas. Temos

$$\pi_n^a = c\,\pi_n^i = c\,\pi_n^o\,,\; n = 0,\,1,\,...,\,K\;\;,$$

restando determinar π_{K+1}^a e c como função do vetor Π^o.

Em equilíbrio e por unidade de tempo, precisamos ter a taxa efetiva de chegadas igual à taxa efetiva de partidas. Ou seja,

$$\lambda\,(1-\pi_{K+1}^a) = \frac{1-\pi_0^a}{E(S)}\,,$$

então,
$$\pi_{K+1}^a = \frac{\rho-(1-\pi_0^a)}{\rho} = \frac{\rho-(1-c\pi_0^o)}{\rho}\,.$$

Determinamos c impondo que a soma das probabilidades seja 1:

$$\sum_{i=0}^{K+1}\pi_i^a = 1 \;\Rightarrow c\sum_{i=0}^{K}\pi_i^o \;+\; \frac{\rho-(1-c\pi_0^o)}{\rho} = 1 \;\;\Rightarrow\;\; c = \frac{1}{\pi_0^o+\rho}\,.$$

Finalmente,

$$\pi_n^a = \begin{cases} \frac{\pi_n^o}{\pi_0^o+\rho} & n = 0,\,1,\,2,\,...,\,K\,; \\ 1-\frac{1}{\pi_0^o+\rho} & n = K+1. \end{cases}$$

Obtemos assim, em função de $\boldsymbol{\Pi}^o$, as expressões para as distribuições estacionárias em vários instantes do tempo.

3.4 Fluxo de usuários em modelos M/G/1 e variantes

Nesta seção vamos caracterizar o processo de saída para alguns modelos da família M/G/1. A estrutura de análise e notação serão a mesma das seções anteriores, isto é, chegadas formam um processo de Poisson com parâmetro λ, o tempo de serviço tem distribuição B e, exceto quando mencionado, a disciplina de serviço é em ordem de chegada (FCFS). Com X_n^o representando o número no sistema imediatamente após o instante de saída T_n^o, vamos considerar o processo

de renovação Markoviano $(X^o, T^o) = \{(X_n^o, T_n^o); n \geq 0\}$. A distribuição estacionária da cadeia imersa no núcleo $Q^o(t)$ é dada por $\mathit{\Pi}^o = (\pi_0^o, \pi_1^o, \cdots)$. Note que, no caso de sala de espera infinita, as saídas e partidas coincidem.

Para $n \geq 1$, denotamos por D_n a duração do intervalo entre saídas sucessivas, ou seja, $D_n = T_n^o - T_{n-1}^o$. Desejamos estudar as características do processo de saída formado pela sequência $\{D_n; n \geq 1\}$. Em regime estacionário a distribuição entre saídas é calculada pelo produto $\mathit{\Pi}^o Q^o(t)e$ e será representada por $F_o(t)$, onde sua transformada de Laplace-Stieltjes é dada por

$$\tilde{F}_o(s) = \Big(\pi_0^o \frac{\lambda}{\lambda + s} + 1 - \pi_0^o\Big)\tilde{B}(s). \tag{3.12}$$

Uma condição necessária (que não é suficiente) para o processo de saída ser de renovação é que exista independência entre o estado da fila e o intervalo entre saídas. Isto é, para todo estado k e etapa n, vale:

$$P(X_n^o = k | D_n \leq t) = \pi_k^o, \text{ com } n, k \geq 0. \tag{3.13}$$

Observe que D_n depende somente de X_{n-1}^{o}. Se este é zero, D_n será a soma dos tempos de chegada e serviço. Por outro lado, se $X_{n-1}^{\mathrm{o}} > 0$, D_n será apenas o tempo de serviço. Dessa forma, a independência entre X_n^o e D_n é equivalente à independência entre D_n e D_{n+1}.

Em regime estacionário e pela definição de $Q^o(t)$, temos

$$\Big[\mathit{\Pi}^o Q^o(t)\Big]_k = P(X_n^o = k | D_n \leq t)\, F_o(t)\,, \forall\, n \geq 0.$$

Utilizando a notação apresentada no ínicio do capítulo para o núcleo de transição e com o uso da condição (3.13), a expressão acima com $k = 0$ e 1, pode ser escrita (no espaço das transformadas LS) como

$$\pi_0^o\, \tilde{h}_0(s) + \pi_1^o\, \tilde{g}_0(s) = \pi_0^o\, \tilde{F}_o(s), \text{ para } k = 0;$$

e

$$\pi_0^o\, \tilde{h}_1(s) + \pi_1^o\, \tilde{g}_1(s) + \pi_2^o\, \tilde{g}_0(s) = \pi_1^o\, \tilde{F}_o(s)\,, \text{ para } k = 1.$$

Após alguma álgebra, essas condições tornam-se respectivamente iguais a:

$$\Big(\pi_0^o \frac{\lambda}{\lambda + s} + \pi_1^o\Big)\tilde{B}(\lambda + s) = \pi_0^o\Big(\pi_0^o \frac{\lambda}{\lambda + s} + 1 - \pi_0^o\Big)\tilde{B}(s); \tag{3.14}$$

$$-\lambda\Big(\pi_0^o\frac{\lambda}{\lambda+s}+\pi_1^o\Big)\widetilde{B}'(\lambda+s)+\pi_2^o\widetilde{B}(\lambda+s)=\pi_1^o\Big(\pi_0^o\frac{\lambda}{\lambda+s}+1-\pi_0^o\Big)\widetilde{B}(s), \qquad (3.15)$$

onde $\widetilde{B}'(\lambda+s)$ denota a derivada de $\widetilde{B}(\lambda+s)$ em relação a s. Essas condições serão utilizadas nas demonstração do teorema 3.3 (a seguir). Antes, porém, precisamos de um resultado preliminar.

Lema 3.2

O sistema M/G/1/K tem as saídas formando um processo de renovação se, e somente se, X_1^o e D_1 são independentes (e assim X_n^o e D_n para todo n).

demonstração:

Ver teorema 3.2 em Disney, Fa0rrell e de Morais [1973].□

Teorema 3.3: Fluxo de saída no modelo M/G/1/K

Para a fila M/G/1/K em equilíbrio considere as seguintes condições:

$C_1)$ Tempos de serviço são todos zero com probabilidade 1;
$C_2)$ $K=0$;
$C_3)$ $K=1$ e o serviço tem duração constante $b\geq 0$;
$C_4)$ $K=\infty$ e os tempos de serviço são exponenciais.

Temos então que:

$i)$ Cada uma das condições C_1 a C_4 implica que o processo de saída é de renovação.

$ii)$ Se as saídas seguem um processo de renovação, então uma das condições C_1 a C_4 se verifica. Nesse caso, a distribuição entre as saídas é dada por:

$$F_o(t)=\begin{cases} A(t)\,, & \text{para } C_1;\\ A*B(t), & \text{para } C_2;\\ \pi_0^o A*B_b(t)+(1-\pi_0^o)B_b(t), & \text{para } C_3;\\ A(t), & \text{para } C_4;\end{cases}$$

com $\pi_0^o=e^{-\lambda b}$ e as funções de distribuição $A(t)$, $B(t)$ e $B_b(t)$ correspondendo respectivamente às distribuições, exponencial (λ), geral e determinística de comprimento b. O símbolo $*$ representa a convolução entre funções distribuições.

demonstração:

Tendo em vista a maior complexidade da demonstração, envolvendo algumas regras de lógica nem sempre familiares ao leitor, vamos antes de demonstrar o teorema, apresentar o esquema lógico que será usado.

Supondo que C_0 represente a condição do processo de saída ser de renovação, precisamos verificar em $i)$ que:

$$(I)\ (C_1 \Rightarrow C_0) \wedge (C_2 \Rightarrow C_0) \wedge (C_3 \Rightarrow C_0) \wedge (C_4 \Rightarrow C_0).$$

Em $ii)$ é preciso demonstrar a implicação $C_0 \Rightarrow C_1 \vee C_2 \vee C_3 \vee C_4$, a qual é equivalente a

$$(II)\ C_0 \wedge \overline{C}_1 \wedge \overline{C}_2 \wedge \overline{C}_3 \Rightarrow C_4\,,$$

onde $\overline{C}_i$ representa a negação de C_i.

Parte 1: Demonstração de (I)

A implicação $(C_4 \Rightarrow C_0)$ é o teorema de Burke, apresentado e demonstrado no capítulo 2 (teorema 2.3). As implicações $(C_1 \Rightarrow C_0)$ e $(C_2 \Rightarrow C_0)$ são imediatas e dispensam prova. Resta $(C_3 \Rightarrow C_0)$, ou seja, $K = 1$ e serviço determinístico implicam que as saídas formam um processo de renovação.

Suponha que o serviço determinístico tenha distribuição

$$B_b(t) = \begin{cases} 0, & t < b; \\ 1, & t \geq b. \end{cases}$$

Se $K = 1$, o sistema, observado nos instantes de saída, só pode estar em 0 ou 1 e a distribuição estacionária será $\pi_0^o = e^{-\lambda b}$ e $\pi_1^o = 1 - e^{-\lambda b}$. Vale, então, a seguinte igualdade

$$\pi_0^o \tilde{B}_b(s) = \tilde{B}_b(\lambda + s),$$

que coincide com a igualdade (3.14) aplicada neste caso. Portanto, como só temos dois estados, (3.15) também vale e temos, assim, a independência entre X_1^o e D_1. Segue do Lema 3.2 que o processo de saída é de renovação. Convém notar que, pela unicidade da inversão das transformadas de Laplace-Stieltjes, na presença de $K = 1$, saídas serem renovação e serviço determinístico são equivalentes.

Parte 2: Demonstração de (II)

Observamos inicialmente que C_3 é composta das partes $\{K = 1\}$ e {serviço constante} que representaremos por C_{3a} e C_{3b}, respectivamente. Da parte 1 desta demonstração já verificamos que, quando $K = 1$, vale $C_0 \Leftrightarrow C_{3b}$ de modo que (II) é equivalente a

$$(II')\ \ C_0 \wedge \bar{C}_1 \wedge \bar{C}_2 \wedge \bar{C}_{3a} \Rightarrow C_4\,.$$

Nossa hipótese fica sendo: processo de saída é de renovação, $\pi_0^o \neq 1$ e $K > 1$. Dividindo (3.15) por (3.14) vem

$$\frac{-\,\lambda\,\tilde{B}'(\lambda+s)}{\tilde{B}(\lambda+s)} + \frac{c}{a+\,\lambda/(\lambda+s)} = a\,,$$

com $a = \pi_1^o/\pi_0^o$ e $c = \pi_2^o/\pi_0^o$. Resolvendo a equação diferencial temos

$$\tilde{B}(\nu) = \tilde{B}(\lambda)\Big(\frac{\lambda+\lambda/a}{\nu+\lambda/a}\Big)^{c/a^2} exp\Big(\frac{c-a^2}{a\lambda}(\nu-\lambda)\Big). \tag{3.16}$$

Substituindo ν por $\lambda + s$ e depois por s e dividindo uma expressão pela outra, obtemos

$$\frac{\tilde{B}(\lambda+s)}{\tilde{B}(s)} = exp\Big(\frac{c-a^2}{a\lambda}\Big)\Big(1-\frac{\lambda}{s+\,\lambda+\lambda/a}\Big)^{c/a^2}. \tag{3.17}$$

Por outro lado, após alguma manipulação algébrica, de (3.14) vem

$$\frac{\tilde{B}(\lambda+s)}{\tilde{B}(s)} = \pi_0^o + \frac{1-\pi_0^o-\pi_1^o}{a} - \frac{1-\pi_0^o-\pi_1^o}{a^2}\,\frac{\lambda}{s+\,\lambda+\lambda/a}\,. \tag{3.18}$$

Uma das nossas hipóteses é o processo de saída ser de renovação e, portanto, as igualdades (3.14) e (3.15) precisam valer. Como elas foram desenvolvidas para produzir (3.17) e (3.18), essas últimas precisam ser iguais. Impondo a igualdade entre o lado direito de (3.17) e de (3.18), verificamos, após a análise dos respectivos polinômios em s, que precisamos ter $c = a^2$. Em seguida, com uso dessa condição, concluimos que também é necessário que $\pi_1^o = \pi_0^o(1-\pi_0^o)$.

A transformação inversa de (3.16), com $c = a^2$, resulta em distribuição exponencial para B. Finalmente, na família das filas M/M/1/K, a igualdade

$\pi_1^o = \pi_0^o(1 - \pi_0^o)$ acontece se, e somente se, $K = \infty$. Concluimos, então, que vale a condição C_4, a implicação (II') foi verificada e o teorema está provado.□

Comentário: Uma conclusão imediata do teorema anterior é que, dentre as filas M/G/1/K, os únicos casos que produzem um processo de saída Poisson são aqueles das condições 1 e 4.

Os resultados apresentados no teorema 3.3 se referem à disciplina FCFS e um servidor. Conforme veremos no próximo teorema, é possível obter saídas Poisson em outras condições.

Teorema 3.4

No modelo M/G/c em equilíbrio, o processo de saída é Poisson se o sistema tem uma das seguintes condições:
$i)$ $c = \infty$;
$ii)$ a disciplina de serviço é a do tempo compartilhado (*processor sharing*);
$iii)$ a disciplina de serviço é LCFS com interrupção (também vale para $K < \infty$);
$iv)$ G = M (serviço exponencial).

demonstração:

A prova será omitida. Consulte Disney & König [1985], Melamed [1979] e as referências lá citadas.□

3.5 Exercícios

1) Para a fila M/G/1 com taxa de chegada $\lambda = 1$ e serviço U$[0, 4]$, calcule as medidas de desempenho L, L_q, W e W_q.

2) Considere a fila M/G/1/1 que tem taxa de chegada $\lambda = 1$ e distribuição de serviço Erlang com parâmetros $k = 2$ e $\mu = 2$. Apresente o núcleo de transição e determine a distribuição estacionária da cadeia imersa nos instantes de saída.

3) Obtenha, para a fila M/G/1 com taxa de chegada λ e serviço U$[0, 4\lambda]$, o núcleo de transição para o processo imerso nos instantes de partida.

4) Para o modelo do exercício 3, calcule o valor esperado do número médio de usuários encontrados por uma chegada (arbitrária).

5) Seja $\lambda = 1$ a taxa de chegada no modelo M/G/1/2 com distribuição de serviço hiperexponencial dada por $B(x) = \alpha\, F_1(x) + (1 - \alpha)\, F_2(x)$ com $0 \leq \alpha \leq 1$ e

$F_1(x)$ e $F_2(x)$ sendo, respectivamente, distribuições exponenciais de parâmetros 2 e 4. Determine a distribuição estacionária do modelo em função de α. Verifique sua resposta colocando $\alpha = 0$ (ou 1) e comparando com as respectivas expressões do modelo M/M/1/2.

6) (Gross & Harris [1985]) Mostre que na fila M/G/1 o valor esperado do tempo requerido para completar o atendimento, do cliente que está sendo servido no momento de uma chegada (arbitrária), é dado por $\frac{\lambda}{2}\mathrm{E}[S^2]$.

7) Determine a distribuição do intervalo entre saídas da fila M/H_2/1, apresentada no exemplo 3.1.

8) Em regime estacionário, qual será a correlação entre dois intervalos consecutivos do processo de partida da fila M/G/1/0?

9) Determine o valor esperado do tempo entre saídas para a fila M/G/1/1 com serviço constante de duração 2 e taxa de chegada igual a 1.

10) Obtenha a distribuição do intervalo entre partidas para a fila M/G/1 com serviço determinístico de duração b.

11) Para a fila M/G/1/1 com serviço U$[0, 4]$ e chegadas com taxa 1, determine o núcleo de transição do processo de renovação Markoviano associado com os instantes de saída da fila.

12) Para a fila descrita no exercício 11, obtenha o número médio de usuários, em regime estacionário, num instante arbitrário de tempo.

Capítulo 4

Redes de Filas

4.1 Introdução

Neste capítulo, vamos apresentar alguns modelos de redes de filas que como o próprio nome indica, são formadas de várias filas interconectadas entre si com usuários deslocando-se entre elas para receber serviço. É usual as filas individuais serem referidas como *nós*, *centros de serviço* ou *estações* e o caminho interno dos usuários como *rotas*. A rede será *aberta* ou *fechada*, dependendo de poder enviar e receber ou não usuários de fora da rede. Isto é, em uma rede fechada o número total de usuários não se altera, havendo apenas uma permutação nas suas posições. Nas redes abertas, por outro lado, a presença total varia pela chegada ou saída externa de usuários.

Redes de filas com serviço exponencial, roteamento Markoviano e apenas um tipo de freguês constituem o modelo mais simples de rede e têm sido bastante estudadas. Nela se incluem os trabalhos pioneiros na área realizados por Jackson [1957 e 1963] e Gordon & Newell [1967]. Essa redes são denominadas redes de Jackson e suas principais características são discutidas na seção 4.2. Em Kelly [1979], redes com serviços exponenciais, porém com vários tipos de fregueses e rotas pré-determinadas, são estudadas. Essa rede é conhecida como rede de Kelly e e é discutida na seção 4.3. Na seção 4.4, introduzimos uma nova rede que estende os modelos anteriores, buscando maior flexibilidade no atendimento. Nossa exposição será baseada no trabalho de Baskett, Chandy, Muntz e Palacios [1975], que originaram o nome com o qual esse modelo é usualmente referido, rede BCMP. Nessa rede os serviços podem ser não exponenciais e a rede pode ser mista, fechada para alguns e aberta para outros tipos. A disciplina de atendimento tem outras opções além da FCFS. O conceito de quase-reversibilidade é definido na seção 4.5. A construção de uma rede de estações quase-reversíveis é apresentada e suas principais propriedades são mencionadas. Em particular, as redes de Jackson, BCMP e Kelly são quase-reversíveis. Há uma vasta literatura explorando a relação entre quase-reversibilidade e a forma da distribuição estacionária e referências serão indicadas ao leitor interessado.

Como veremos ao longo desse capítulo, as chegadas, atendimento e rotas dos usuários são, entre outras, as características que definem os diversos tipos de redes. Os modelos que optamos por apresentar, dão uma idéia da complexidade e variedade de opções que podem ser utilizadas num problema aplicado. É importante mencionar, que o leitor deverá enfrentar agora uma notação mais carregada e nem sempre intuitiva à primeira vista. Esperamos que isso não impeça o reconhecimento pelo leitor das possibilidades de aplicação dos modelos discutidos.

4.2 O modelo de Jackson

4.2.1 Redes abertas

Uma rede aberta de Jackson, exemplificada na figura 4.1, consiste de um conjunto de J nós $\mathcal{J} = \{1, 2, \cdots, J\}$, com as seguintes características (ver Jackson [1957]):

a) As chegadas externas a cada estação $i \in \mathcal{J}$ ocorrem de acordo com um processo de Poisson de parâmetro λ_i, independente das outras chegadas e do serviço nas estações;

b) Para cada nó $i \in \mathcal{J}$, o atendimento é em ordem de chegada segundo uma distribuição exponencial de parâmetro $\mu_i(n_i)$, $\mu_i(0) = 0$. A dependência da taxa ao número de usuários no centro de serviço permite, entre outras opções, que vários servidores em paralelo atendam com a mesma taxa;

c) Após o usuário completar seu serviço na estação i, ele move-se para mais serviço na estação j com probabilidade $r_{i,j}$, onde $i, j \in \mathcal{J}$ (note que é possível retornos à mesma estação). Por outro lado, se o usuário já completou sua demanda de serviço na rede, ele deixa o sistema após ser atendido na estação i, com probabilidade $r_{i,J+1} = 1 - \sum_{k=1}^{J} r_{i,k}$, onde $J+1$ é um nó artificial representando o exterior da rede. Todas as transferências são assumidas serem instantâneas e assim o tempo gasto no sistema é proveniente de esperas e atendimentos nas estações.

Seja R a matriz quadrada de ordem $J+1$ formada pelas probabilidades de mudança das estações (incluindo o nó externo). Para completar sua definição atribuimos a seguinte probabilidade para as entradas externas através do nó i :

$$r_{J+1,i} = \lambda_i / \sum_{k=1}^{J} \lambda_k \, , \forall \, i \in \mathcal{J} \;\; \text{e} \;\; r_{J+1,J+1} = 0.$$

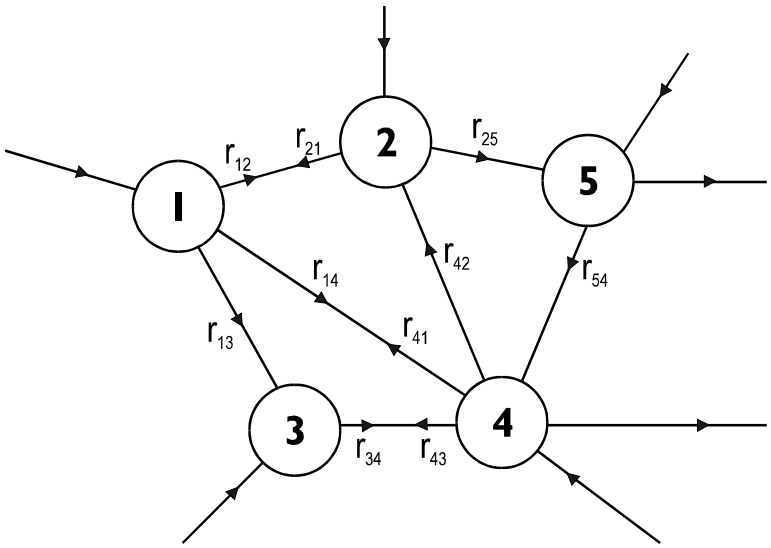

Figura 4.1: Rede de Jackson com 5 nós

Dessa forma, R fica sendo uma matriz estocástica e portanto representando as transições para alguma cadeia de Markov. É usual R ser denominada como a *matriz de roteamento* da rede de Jackson.

Da teoria das cadeias de Markov, vamos relembrar a definição de comunicação entre estados. Dizemos que i se comunica com j ($i \rightarrow j$) quando é possível, em algum número de etapas, partir de i e alcançar j. Isto é, existe uma sequência de estações $i_0 = i, i_1, i_2, \ldots, i_k = j,$ de modo que, para algum índice $k > 0,$ temos

$$r_{i,i_1} r_{i_1,i_2} r_{i_2,i_3} \cdots r_{i_{k-1},i_k} r_{i_k,j} > 0.$$

Vamos acrescentar mais duas condições, além das mencionadas acima, para garantir que todos os centros de serviço possam receber usuários e que os fregueses que entram, tenham probabilidade positiva de sair em algum momento:

d) Todos os nós da rede poderão receber usuários externos de forma direta ou indireta. Isto é, $\forall \, j \in \mathcal{J}$ *ou* $\lambda_j > 0$ *ou existe um* i *tal que* $\lambda_i > 0$ *e* $i \rightarrow j$.

e) O usuário servido em i *tem probabilidade positiva de sair da rede em uma ou mais etapas. Ou seja,* $\forall\, i \in \mathcal{J}$ *ou* $r_{i,J+1} > 0$ *ou existe um* $j \in \mathcal{J}$ *tal que* $r_{j,J+1} > 0$ *e* $i \to j$.

Com as condições acima, a cadeia de Markov associada com a matriz R será recorrente não nula e as *equações de tráfego* abaixo

$$\gamma_m = \lambda_m + \sum_{k=1}^{J} r_{i,k}\,\gamma_k \ , m \in \mathcal{J}\ ,$$

terão solução única. Note que as equações de tráfego correspondem às m primeiras equações na igualdade $\gamma R = \gamma$ com γ sendo um vetor de ordem $J+1$ e $\gamma_{J+1} = \sum_{k=1}^{J} \lambda_k$. Podemos interpretar γ_m como a taxa total de entradas ao nó $m \in \mathcal{J}$, sendo resultado da composição das chegadas (externas) com as saídas de estações que se movimentam para o centro m.

Seja N o processo estocástico de dimensão J representando, em tempo contínuo, o número de usuários em cada centro de serviço da rede, isto é, para cada $t \in \mathbb{R}_+, N(t) = (N_1(t), N_2(t), \cdots, N_J(t))$. Não é difícil verificar que esse processo é Markoviano com espaço de estados dado pelo produto cartesiano dos números naturais, $\{0, 1, \cdots\}^{\times J}$. Note que as transições nesse processo são de três tipos: chegadas externas, fim de serviço com mudança de nó (que acarreta uma chegada interna) e fim de serviço com saída da rede. Se a rede não está vazia, o tempo de espera para uma transição é o mínimo entre todos os serviços que estão sendo realizados e as chegadas externas. Como todos esses tempos são exponenciais independentes, segue que a espera para uma transição será também exponencial com parâmetro dependendo do estado do sistema antes da transição. Se a rede está vazia, será necessário esperar pela primeira chegada, que será o mínimo entre os tempos de chegadas externas e, portanto, com distribuição exponencial.

Um teorema muito importante que foi provado por Jackson afirma que a distribuição estacionária do número de usuários presentes no sistema pode ser calculada por um produto de distribuições estacionárias marginais (referentes a cada centro). Este resultado é extraordinário pois indica que o sistema se comporta como se cada centro de serviço $i \in \mathcal{J}$ fosse uma fila com chegadas e serviços exponenciais com taxas γ_i e $\mu_i(n_i)$, respectivamente, independente dos demais centros. Assim, no caso de c_i servidores idênticos, teríamos o sistema se comportando como se fosse uma coleção independente de filas M/M/c_i.

Contudo, é importante observar que, devido ao roteamento interno, as chegadas a cada centro não formam, em geral, um processo de Poisson. Portanto, o resultado obtido, usualmente denominado *solução na forma produto,* é surpreendente e deve ser aplicado com cuidado. Em especial, observe que há independência entre o número de usuários em cada centro de serviço para cada t fixado; mas $N(t_1)$ e $N(t_2)$ não são independentes, apesar de terem a mesma distribuição estacionária. A vantagem da *forma produto* é, entre outras, a possibilidade de rapidamente identificar os efeitos da mudança dos parâmetros de um centro sobre o desempenho de toda a rede.

Para utilização como constantes normalizadoras a seguir, definimos

$$b_i^{-1} = \sum_{n=0}^{\infty} \gamma_i^n / [\mu_i(1)\, \mu_i(2) \cdots \mu_i(n)]\,,$$

onde o termo correspondente a $n = 0$ é igual a 1.

Teorema 4.1: Forma produto na rede aberta de Jackson

Sob a condição de $b_i^{-1} < \infty$, $\forall\, i \in \mathcal{J}$, o processo N tem distribuição estacionária dada por:

$$\pi(\overline{n}) = \prod_{i=1}^{J} \pi_i(n_i); \tag{4.1}$$

com $\overline{n} = (n_1, n_2, \cdots, n_J)$ e

$$\pi_i(n_i) = b_i\, \gamma_i^{n_i} / [\mu_i(1)\, \mu_i(2) \cdots \mu_i(n_i)], i \in \mathcal{J}\,.$$

demonstração:

Vamos verificar que a expressão proposta para $\pi(\overline{n})$ satisfaz as equações de balanço global e, portanto, ela é uma distribuição estacionária. As condições para unicidade da distribuição estão associadas à equação de tráfego e são obtidas da teoria dos processos de Markov.

Definimos a seguinte notação auxiliar:

$\delta(k) = min\,(k, 1)\,, \forall\, k \geq 0;$
$\overline{n} + 1_i = (n_1, n_2, \cdots, n_i + 1, \cdots, n_J), \forall\, i \in \mathcal{J};$
$\overline{n} - 1_i = (n_1, n_2, \cdots, n_i - 1, \cdots, n_J), \forall\, i \in \mathcal{J};$
$\overline{n} + 1_j - 1_i = (n_1, n_2, \cdots, n_i - 1, \cdots, n_j + 1, \cdots, n_J), \forall\, i, j \in \mathcal{J}.$

Para todo $\overline{n} \in \{0,1,\cdots\}^{\times J}$, a equação de balanço global é a seguinte:

$$\Big(\sum_{i\in\mathcal{J}}\lambda_i + \sum_{i\in\mathcal{J}}\mu_i(n_i)\Big)\pi(\overline{n}) = \sum_{i\in\mathcal{J}}\mu_i(n_i+1)\,r_{i,J+1}\,\pi(\overline{n}+1_i) + \\ + \sum_{i\in\mathcal{J}}\lambda_i\,\delta(n_i)\,\pi(\overline{n}-1_i) + \sum_{i\in\mathcal{J}}\sum_{j\in\mathcal{J}}\mu_j(n_j+1)\,r_{j,i}\,\pi(\overline{n}+1_j-1_i).$$

Tendo em vista a solução proposta para $\pi(\overline{n})$, valem as seguintes relações:

$$\pi(\overline{n}+1_i) = \frac{\gamma_i}{\mu_i(n_i+1)}\,\pi(\overline{n}),$$
$$\pi(\overline{n}-1_i) = \frac{\mu_i(n_i)}{\gamma_i}\,\pi(\overline{n}),$$
$$\pi(\overline{n}+1_j-1_i) = \frac{\gamma_j\,\mu_i(n_i)}{\gamma_i\,\mu_j(n_j+1)}\,\pi(\overline{n}),$$

as quais substituidas no lado direito da equação de balanço global resulta:

$$\text{Lado direito} = \pi(\overline{n})\Big\{\sum_{i\in\mathcal{J}}\mu_i(n_i+1)\,r_{i,J+1}\Big(\frac{\gamma_i}{\mu_i(n_i+1)}\Big) + \\ + \sum_{i\in\mathcal{J}}\lambda_i\,\delta(n_i)\frac{\mu_i(n_i)}{\gamma_i} + \sum_{i\in\mathcal{J}}\sum_{j\in\mathcal{J}}\mu_j(n_j+1)\,r_{j,i}\,\frac{\gamma_j\,\mu_i(n_i)}{\gamma_i\,\mu_j(n_j+1)}\Big\} = \\ = \pi(\overline{n})\Big\{\sum_{i\in\mathcal{J}} r_{i,J+1}\gamma_i + \sum_{i\in\mathcal{J}}\lambda_i\,\delta(n_i)\frac{\mu_i(n_i)}{\gamma_i} + \sum_{i\in\mathcal{J}}\sum_{j\in\mathcal{J}} r_{j,i}\,\frac{\gamma_j\,\mu_i(n_i)}{\gamma_i}\Big\} = \\ = \pi(\overline{n})\Big\{\sum_{i\in\mathcal{J}}(1-\sum_{j\in\mathcal{J}} r_{i,j})\gamma_i + \sum_{i\in\mathcal{J}}\lambda_i\frac{\mu_i(n_i)}{\gamma_i} + \sum_{i\in\mathcal{J}}\frac{\mu_i(n_i)}{\gamma_i}\sum_{j\in\mathcal{J}} r_{j,i}\,\gamma_j\Big\},$$

onde usamos a igualdade $\delta(n_i)\mu_i(n_i) = \mu_i(n_i)$ e a definição de $r_{i,J+1}$. Finalmente, com a utilização da equação de tráfego, obtemos

$$\text{Lado direito} = \pi(\overline{n})\Big\{\sum_{i\in\mathcal{J}}\lambda_i + \sum_{i\in\mathcal{J}}\lambda_i\frac{\mu_i(n_i)}{\gamma_i} + \sum_{i\in\mathcal{J}}\frac{\mu_i(n_i)}{\gamma_i}(\gamma_i-\lambda_i)\Big\} = \\ = \pi(\overline{n})\Big\{\sum_{i\in\mathcal{J}}\lambda_i + \sum_{i\in\mathcal{J}}\mu_i(n_i)\Big\} = \text{Lado esquerdo.}$$

Concluimos então que as equações de balanço estão satisfeitas e a demonstração está completa.□

Exemplo 4.1:

Para $i \in \mathcal{J}$, seja $\mu_i(n_i) = \mu_i$ para todo $n_i > 0$ e $\mu_i(0) = 0$. Segue pelo teorema 4.1

$$\pi_i(n_i) = (1 - \rho_i)\, \rho_i^{n_i}, \quad n_i = 0, 1, 2, \cdots,$$

onde $\rho_i = \gamma_i / \mu_i < 1$.

A expressão acima corresponde à distribuição estacionária da fila M/M/1 com taxa de entrada γ_i e serviço μ_i. Dessa forma, pela expressão (4.1) o processo $N(t)$ comporta-se como uma coleção de M/M/1 independentes apesar de não ser, pois o processo de entrada em cada fila não é (em geral) Poisson.□

4.2.2 Redes fechadas

Vamos estudar a rede fechada de Jackson seguindo a mesma notação, com algumas adaptações, da rede aberta. Seja M o número total de usuários na rede, isto é, para qualquer t, precisamos sempre ter $\sum_{i=1}^{J} n_i = M$. Como não há entradas nem saídas da rede, a taxa de chegadas (externas) λ_i é igual a zero para todo $i \in \mathcal{J}$. A matriz R perde então a linha e coluna $J + 1$. A equação de tráfego torna-se assim:

$$\gamma_m = \sum_{k=1}^{J} r_{i,k}\, \gamma_k \ , m \in \mathcal{J} \ .$$

Pelas hipóteses já apresentadas na rede aberta, a matriz R é finita e irredutível e a equação acima terá solução única com a imposição de alguma condição extra (por exemplo, fixe o valor de um dos γ's ou de sua soma).

Teorema 4.2: Forma produto na rede fechada de Jackson

O processo N tem distribuição estacionária dada por:

$$\pi(\overline{n}) = b_M \prod_{i=1}^{J} \gamma_i^{n_i} / [\mu_i(1)\, \mu_i(2) \cdots \mu_i(n_i)], \tag{4.2}$$

para $\overline{n} = (n_1, n_2, \cdots, n_J)$, $\sum_{i=1}^{J} n_i = M$ e b_M a constante de normalização.

demonstração:

Podemos seguir, com as devidas adaptações, os passos da demonstração do teorema 4.1. Com a restrição do espaço amostral aos vetores que somem M, a equação de balanço global é a seguinte:

$$\Big(\sum_{i\in\mathcal{J}}\mu_i(n_i)\Big)\pi(\overline{n}) = \quad \sum_{i\in\mathcal{J}}\sum_{j\in\mathcal{J}}\mu_j(n_j+1)\, r_{j,i}\, \pi(\overline{n}+1_j-1_i)\ .$$

A verificação é imediata após substituição da expressão proposta para $\pi(\overline{n})$. A constante b_M é obtida impondo que a soma das probabilidades $\pi(\overline{n})$ seja igual a 1.□

Note que, diferentemente do que aconteceu na rede aberta, a constante de normalização não é fatorável num produto de constantes associadas a cada centro. Assim, não teremos independência entre o número de usuários em cada centro para cada t fixado. Apesar disso, a literatura especializada continua se referindo à expressão 4.2 como sendo uma *forma produto*. O cálculo de b_M fica bastante trabalhoso computacionalmente, uma vez que o número de estados cresce rapidamente com J e M.

Exemplo 4.2: (Gelenbe & Pujolle [1987])

Considere a rede fechada representada na figura 4.2 onde três estações estão modelando um sistema computacional com memória virtual. A estação 1 é o processador central, a 2 representa a memória secundária e a estação 3 os dispositivos de entrada e saída (I/O) acoplados ao sistema computacional.

Admitindo multiprogramação de nível M, vamos supor que cada programa que termina é imediatamente substituído por outro de modo a manter um número constante de programas em andamento na rede.

Programas entrando no sistema são colocados na fila de espera do processador central. A execução do programa continua nessa estação até que um dos três eventos acontece:

i) Termina a execução no processador central, correspondendo ao ramo r_{11}, sendo então imediatamente substituído por outro programa;

ii) Demanda excessiva de computação deslocará o programa para a memória secundária, isto corresponde à rota r_{12};

iii) Entrada ou saída é solicitada, neste caso o programa segue a rota r_{13} dirigindo-se à estação 3.

Após completar suas tarefas nas estações 2 ou 3, o programa retorna ao processador central (estação 1). A matriz R é dada por:

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

As equações de tráfego correspondem ao sistema $\gamma R = \gamma$. Tomando $\gamma_1 = 1$, os valores γ_2 e γ_3 representam o número médio de visitas às estações 2 e 3, respectivamente, relativas ao número de visitas à estação 1. Teremos então a solução $\gamma_2 = r_{12}$ e $\gamma_3 = r_{13}$ e portanto

$$\pi(\overline{n}) = \pi(n_1, n_2, n_3) = b_M \left(\frac{1}{\mu_1}\right)^{n_1} \left(\frac{r_{12}}{\mu_2}\right)^{n_2} \left(\frac{r_{13}}{\mu_3}\right)^{n_3}.$$

Neste caso, a constante b_M pode ser obtida com mais facilidade. Multiplicando e dividindo por μ_1^M temos

$$b_M = \left(\frac{1}{\mu_1}\right)^M \sum_{n_1+n_2+n_3=M} \left(\frac{\mu_1 r_{12}}{\mu_2}\right)^{n_2} \left(\frac{\mu_1 r_{13}}{\mu_3}\right)^{n_3},$$

o que ainda poderá ser uma computação extensa dependendo do valor de M. Resultados analíticos e de simulação são comparados e apresentados no capítulo 2 de Gelenbe & Pujolle [1987].□

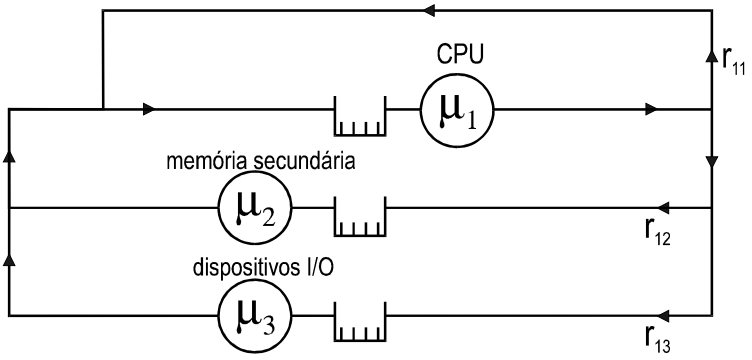

Figura 4.2: Sistema computacional modelado como rede fechada

4.2.3 Fluxo e reversibilidade em redes de Jackson

Nesta sub-seção apresentamos resumidamente alguns resultados sobre fluxo de usuários na rede de Jackson. Há muito material a ser coberto e as demonstrações são bastante elaboradas.

A reversibilidade da rede de Jackson está relacionada ao comportamento da matriz R, conforme estabelece a próxima proposição.

Teorema 4.3: **Reversibilidade na rede de Jackson**

O processo $N(t)$ é reversível se, e somente se, R for uma cadeia de Markov reversível.

demonstração:

Faremos a demonstração no caso de uma rede aberta, para o caso fechado basta repetir a prova com as devidas adaptações.

As equações de balanço detalhado são equivalentes a:

$$\gamma_i \, r_{i,j} = \gamma_j \, r_{j,i} \ , \ i, j \in \mathcal{J} \cup \{J+1\}$$

Definindo $\gamma_i^* = \gamma_i / \sum_{k=1}^{J+1} \gamma_k$, a expressão acima pode ser reescrita como

$$\gamma_i^* \, r_{i,j} = \gamma_j^* \, r_{j,i} \ , \ i, j \in \mathcal{J} \cup \{J+1\},$$

portanto, a verificação dessa equação acontece, se e somente se, a matriz R for reversível, uma vez que γ_i^* faz as vezes da probabilidade estacionária.□

Apesar da rede de Jackson não ser, em geral, reversível, resultados interessantes podem ser obtidos observando o processo reverso. Vamos listar três deles:

i) Em regime estacionário, a cadeia reversa de uma rede aberta de Jackson é também uma rede aberta de Jackson;

ii) Para uma rede de Jackson, aberta e estacionária, o processo de saída de cada estação i é Poisson de taxa $\gamma_i \, r_{i,J+1}$, independentes entre si;

iii) Para qualquer t fixado, os processos de saída da rede antes de t são independentes do estado do sistema $N(t)$.

As demonstrações desses resultados baseiam-se na trajetória do processo original e na definição de reverso e são encontradas em Wolf [1989], Disney & König [1985] e nas referências lá citadas.

Em Melamed [1979], o fluxo em redes abertas de Jackson é estudado em detalhes. Ele demonstra que o tráfego nunca é Poisson exceto nos arcos de saída, isto é, arcos (i, j) tais que $i \rightarrow j$ mas $j \not\rightarrow i$.

O processo de entrada em cada nó é a composição das chegadas externas (Poisson) com as transferências internas (não Poisson). Mesmo as entradas não sendo um processo de Poisson vale o resultado *Pasta*, ou seja, a proporção das entradas (composição), que encontram n usuários num centro, é igual à fração do tempo em que esse mesmo centro tem n usuários.

Para redes fechadas valem resultados similares. Merece destaque o resultado sobre as entradas nos centros de serviço de uma rede fechada. É possível verificar que a proporção das entradas na estação i que encontram k usuários, é igual à fração do tempo em que a estação i tem k usuários, porém numa rede com um freguês a menos no total.

O leitor interessado pode consultar as referências já mencionadas para detalhes e outros resultados.

4.3 Redes de Kelly

Nessa seção estudamos o modelo apresentado em Kelly [1979]. Vamos introduzir nova notação e seguiremos, na medida do possível, a notação original do autor. O modelo permite vários tipos de fregueses que têm sua rota pré-determinada e portanto não escolhem aleatoriamente seu próximo centro de serviço. O próximo exemplo descreve uma aplicação desse modelo.

Exemplo 4.3: (Kelly [1979])

Considere uma pequena linha de produção (ver figura 4.3) constituída de dois tipos de peças. O tipo 1 deve passar pelas máquinas 1, 3 e 4, enquanto o tipo 2 percorre as máquinas 2, 3, 5. Admitimos que a demanda por processamento (chegadas) segue um processo de Poisson para os tipos 1 e 2 e os serviços são exponenciais para todas as máquinas. Os parâmetros das distribuições dependem de cada máquina.

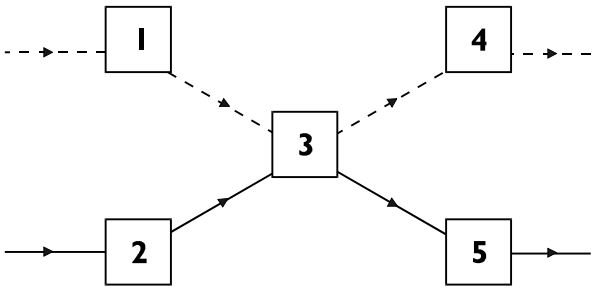

Figura 4.3: Exemplo de uma linha de produção

As peças processadas na máquina 3 não escolhem aleatoriamente seu movimento, as do tipo 1 dirigem-se à máquina 4 e as do tipo 2 à maquina 5. A matriz de roteamento R, da seção anterior, ficaria sem sentido nesse caso.□

Consideramos um sistema com J centros de serviço e com sala de espera ilimitada. O conjunto $\mathcal{J} = \{1, 2, \cdots, J\}$ representará os centros. Existem I tipos diferentes de fregueses caracterizando diferentes rotas através da rede. Isto é, o freguês tipo i segue a rota

$$k(i,1), k(i,2), k(i,3), \cdots, k(i,S(i)),$$

onde, para o freguês tipo i, $k(i,j)$ é o j-ésimo centro visitado e $k(i,S(i))$ é o último. Os fregueses tipo i chegam à rede de acordo com um processo de Poisson com parâmetro λ_i e seu tipo permanece inalterado durante toda sua permanência na rede.

Vamos supor que, para cada nó j, os fregueses estão ordenados. Isto é, o centro j contém os fregueses nas posições $1, 2, 3, \cdots, n_j$, onde n_j é o número presente no centro j. O serviço no centro j é processado da seguinte forma:

i) Cada freguês requer uma quantidade de serviço aleatória com distribuição exponencial de média 1*;*

ii) O total de serviço fornecido é $\phi_j(n_j)$, $(\phi_j(n_j) > 0$ *para* $n_j > 0)$*;*

$iii)$ *A proporção* $\gamma_j(l, n_j)$ *desse esforço é dirigida ao freguês na posição l, isto é, o serviço é processado com taxa* $\phi_j(n_j)\gamma_j(l, n_j)$ *por unidade de tempo; ao deixar o centro, os fregueses das posições* $l + 1$ *a* n_j *movem-se uma posição à frente;*

$iv)$ *Um freguês, chegando ao nó j, ocupa a posição* l *com probabilidade* $\delta_j(l, n_j + 1)$*, fazendo com que os fregueses das posições* l *a* n_j *movam-se uma posição para trás.*

Assumimos independência entre os serviços nos diferentes centros, entre serviços e chegadas e também entre chegadas a cada centro.

Exemplo 4.4

Considere a seguinte opção de parâmetros:

$$\phi_j(n) = \mu_j \, min(c, n)$$

$$\delta_j(l, n) = \begin{cases} 1 & l = n \\ 0 & \text{caso contrário} , \end{cases}$$

$$\gamma_j(l, n) = \begin{cases} \frac{1}{n} & l = 1, 2, \cdots, n; \; n = 1, 2, \cdots, c \\ \frac{1}{c} & l = 1, 2, \cdots, c; \; n = c+1, c+2, \cdots \\ 0 & \text{caso contrário.} \end{cases}$$

Nesse caso, o centro j comporta-se como uma fila com c servidores que atendem em ordem de chegada com parâmetro μ_j.□

Outras escolhas podem gerar diferentes modelos de atendimento. Pode-se modelar, por exemplo, a disciplina LCFS e o serviço em ordem aleatória.

Tendo em vista que o usuário pode visitar mais de uma vez o mesmo centro, vamos definir a classe do freguês como sendo $c_j(l) = (t_j(l), s_j(l))$, onde o par ordenado indica, respectivamente, o tipo do freguês na posição l no centro j e o estágio alcançado na sua rota. Para caracterizar o estado da rede definimos

$$c_j = (c_j(1), c_j(2), \cdots, c_j(n_j)) \text{ , para } j \in \mathcal{J} \text{ ,}$$
$$C = (c_1, c_2, \cdots, c_J).$$

Dessa forma, a variável aleatória $C(t) = C$ indica o estado da rede no instante t. Pode-se verificar que o processo estocástico $\mathbb{C} = \{C(t); t \geq 0\}$ é um processo de Markov com espaço de estados enumerável. As possíveis transições nessa rede são partidas do centro para fora da rede, transferências entre centros (incluindo *feedback*), e chegadas ao centro de fora da rede. Na definição das taxas de transição, precisamos levar em conta que para um mesmo estado de início e de fim pode haver mais de uma possibilidade de ação; dessa forma, efetuamos a somatória de todos os casos possíveis. Supondo que a rede está no estado C, vamos obter as taxas de transição.

Se um usuário na posição l do centro j deixar a rede, diremos que a rede foi para o estado $T_{jl.}C$ e a taxa para essa ocorrência é

$$q(C, l, ., T_{jl.}C) = \phi_j(n_j)\gamma_j(l, n_j) \; .$$

Como já mencionamos, pode acontecer casos onde $T_{jl.}C = T_{jg.}C$ para $l \neq g$. Um exemplo simples dessa situação ocorre quando todos os usuários presentes no centro j são do mesmo tipo. Portanto, partindo do estado C, chegamos ao estado $T_{jl.}C$ de formas diferentes e precisamos somar todas elas de modo a obter a taxa de transição. Temos

$$q(C, T_{jl.}C) = \sum_g q(C, g, ., T_{jg.}C) = \sum_g \phi_j(n_j)\gamma_j(g, n_j),$$

onde, como mencionamos, g é tal que $T_{jl.}C = T_{jg.}C$.

Suponha agora um usuário na posição l do centro j que termina seu serviço e vai ao próximo da sua rota, k, onde ocupa a posição m. Sendo $T_{jlm}C$ o estado resultante vem

$$q(C, l, m, T_{jlm}C) = \phi_j(n_j)\gamma_j(, n_j)\, \delta_k(m, n_k + 1_{\{k \neq j\}}),$$

então

$$q(C, T_{jlm}C) = \sum_g \sum_h q(C, g, h, T_{jgh}C),$$

onde a somatória percorre os g e h tais que $T_{jlm}C = T_{jgh}C$. Note que o indicador $1_{\{k \neq j\}}$ garante que não contemos duas vezes o usuário que faz *feedback*.

Se temos uma chegada externa com o usuário de tipo i ocupando a posição m do primeiro centro da sua rota, a rede ficará no estado $T^{im}C$. A transição correspondente é dada por

$$q(C, ., m, T^{im}C) = \lambda_i \delta_k(m, n_k + 1) \, ,$$

onde k é o primeiro centro da rota do freguês tipo i. A transição entre C e $T^{im}C$ é dada por

$$q(C, T^{im}C) = \sum_h q(C, ., h, T^{ih}C) \, ,$$

e a somatória é para os h's tais que $T^{ih}C = T^{im}C$.

Cabe ainda observar que existem estados em que um ou outro operador não é aplicável e, portanto, as taxas apresentadas acima estão restritas aos casos possíveis. Definimos ainda

$$\alpha_j(i,s) = \begin{cases} \lambda_i & \text{se } r(i,s) = j \\ 0 & \text{caso contrário} \end{cases}$$

$$\text{e} \qquad a_j = \sum_{i \in I} \sum_{s=1}^{S(i)} \alpha_j(i,s)\,.$$

Note que a_j é, em equilíbrio, o número médio (composto) de entradas em cada centro j por unidade de tempo. Utilizaremos a constante de normalização

$$b_j^{-1} = \sum_{n=0}^{\infty} \frac{a_j^n}{\prod_{l=1}^{n} \phi_j(l)}\,, j \in \mathcal{J}.$$

Teorema 4.4: Forma produto na rede de Kelly

Para $b_j^{-1} < \infty$ para todo j, o processo de Markov $\mathbb{C}$ tem distribuição estacionária dada por

$$\pi(C) = \prod_{j=1}^{J} \pi_j(c_j)\,, \tag{4.6}$$

onde para $j \in \mathcal{J}$

$$\pi_j(c_j) = b_j \prod_{l=1}^{n_j} \frac{\alpha_j(t_j(l), s_j(l))}{\phi_j(l)}. \tag{4.7}$$

demonstração:

Pela definição da constante b_j, verificamos imediatamente que a soma das probabilidades na expressão 4.6 é igual a 1. Para auxiliar a demonstração vamos utilizar a proposição c1, apresentada no apêndice, que estabelece uma relação entre o processo direto e seu reverso (taxas representadas por $\hat{q}$). Por essa proposição, para concluir que a expressão (4.7) é a distribuição estacionária dos dois processos precisamos verificar que

$i)\ \pi(C)q(C,D) = \pi(D)\hat{q}(D,C)$ para quaisquer estados C e D;
$ii)\ q(C) = \hat{q}(C)$ para qualquer C;

No processo reverso o usuário de tipo i entra no sistema segundo um processo de Poisson com parâmetro λ_i e percorre a seguinte rota antes de deixar a rede:

$$k(i,S(i)), k(i,S(i)-1), \cdots, k(i,1).$$

A interpretação de γ_j e δ_j são invertidas, isto é, $\gamma_j(l, n_j+1)$ fica sendo a probabilidade de um usuário que chega ao nó j (quando há n_j) ocupar a posição l, enquanto que $\delta_j(l)$ é a proporção de serviço recebida pelo usuário na posição l do centro j. Para o processo reverso temos:

$\hat{q}(T_{jl\cdot}C, ., l, C) = \lambda_i\gamma_j(l, n_j)\ )$, onde $i = t_j(l)$;
$\hat{q}(T_{jlm}C, m, l, C) = \phi_k(n_k + 1_{\{k\neq j\}})\delta_k(m, n_k + 1_{\{k\neq j\}})\gamma_j(l, n_j)\ ,$
onde $k = k(t_j(l), s_j(l)+1)$;
$\hat{q}(T^{im}C, m, ., C) = \phi_k(n_k+1)\delta_k(m, n_k+1)$, onde $k = k(i,1)$.

Podemos obter as taxas de transições do processo reverso somando todas as ações que conduzem ao mesmo estado (de modo análogo ao que fizemos no processo direto).

Substituindo a expressão proposta para $\pi(C)$ obtemos

$$\pi(C)q(C,l,m,T_{jlm}C) = \pi(T_{jlm}C)\hat{q}(T_{jlm}C, m, l, C)\ ,$$

que, após somarmos para todo l e m, produz a seguinte relação entre as taxas de transição dos dois processos

$$\pi(C)q(C,T_{jlm}C) = \pi(T_{jlm}C)\hat{q}(T_{jlm}C, C)\ .$$

Repetindo-se o procedimento com as outras transições, podemos concluir que para todos os estados C e D vale

$$\pi(C)q(C,D) = \pi(D)\hat{q}(D,C)\ .$$

Para concluir a demonstração notamos que

$$q(C) = \hat{q}(C) = \sum_{j=1}^{J}\phi_j(n_j)\sum_{i=1}^{I}\lambda_i\ ,$$

portanto as condições *i)* e *ii)* estão verificadas e o teorema está provado.□

Note que, o processo reverso será uma rede com características similares às descritas para o processo direto. Apresentamos, sem demonstração, dois corolários.

Corolário 4.5:

Em equílibrio, o processo de saída de cada tipo de usuário é um processo de Poisson (taxa λ_i para o tipo i) e são independentes entre si. A variável aleatória $C(t_0)$ é independente das saídas do sistema anteriores a t_0.□

Corolário 4.6:

Em equílibrio, o estado do centro j é independente do resto do estado da rede e é c_j com probabilidade $\pi_j(c_j)$. A probabilidade que o centro j contenha n usuários é dada por

$$P(n_j = n) = b_j \frac{a_j^n}{\prod_{l=1}^{n} \phi_j(l)}, j \in \mathcal{J}.$$

A probabilidade de um usuário na posição l do centro j ser do tipo i no estágio s da sua rota é dada por $\alpha_j(i,s)/a_j$. □

4.4 Redes BCMP

A rede que vamos descrever foi introduzida no artigo Baskett, Chandy, Muntz & Palacios [1975] motivada pela modelagem de sistemas computacionais. O modelo estudado tem 4 tipos diferentes de centros de serviço. Em três deles o tempo de serviço tem distribuição de Cox (ou distribuição com transformada de Laplace-Stieltjes racional) dependendo da classe do usuário. As chegadas e rotas podem também depender do tipo (classe) do usuário. Assumimos que temos I diferentes tipos de usuários e J centros de serviço, representados pelo conjunto $\mathcal{J} = \{1, 2, \cdots, J\}$. No caso de rede aberta, acrescentamos a esse conjunto, os centros fictícios 0 e $J+1$ para representar a entrada e saída do sistema. O movimento dentro da rede segue uma cadeia de Markov com probabilidade de transição dada por:

$$P = \{p_{jr,ks}\} \text{ com } 0 \leq j \leq J, 1 \leq k \leq J+1, 1 \leq r, s \leq I,$$

onde a quantidade $p_{jr,ks}$ indica a probabilidade de um usuário da classe r no centro j deslocar-se para o centro k na classe s, enquanto que $p_{jr,J+1s}$ indica a

saída da rede. Assumimos que P possa ser decomposta em m sub-cadeias ergódicas com estados representados por $E_1, E_2, \cdots, E_m$. Duas classes de usuários podem pertencer à mesma partição se os usuários podem passar de uma classe a outra.

As chegadas podem ser de dois tipos. No primeiro, os usuários chegam de acordo com um processo de Poisson com taxa $\lambda(n)$ com n sendo o número total na rede. Nesse caso uma chegada externa alcança o centro j na classe r com probabilidade $q_{j,r}$. O segundo tipo de chegada consiste em m processos de Poisson independentes, correspondendo às m partições da matriz P. A taxa será $\lambda_\nu(\tilde{n}_\nu)$, com $\nu = 1, 2, \cdots, m$, e $\tilde{n}_\nu$ sendo o número de usuários associados à partição ν. Fixada uma partição qualquer ν, a chegada tem probabilidade $q_{j,r}$ de ser da classe r e ir para o centro j, para (j,r) em E_ν. A rede será fechada com relação à partição ν se $q_{j,r} = 0,\ \forall\ (j,r)$ em E_ν.

Para cada sub-cadeia $E_\nu,\ \nu = 1, 2, \cdots, m$, definimos o conjunto de equações:

$$\sum_{(j,r)\in E_\nu} e_{jr} p_{jr,ks} + q_{ks} = e_{ks}\,, (k,s) \in E_\nu\,,$$

onde e_{ks} é a taxa de entradas (composta) de usuários da classe s no centro k. Assumimos que, no caso de nem todos os q_{ks} serem zero, a equação acima tem solução única.

Antes de apresentar os vários tipos de centros, vamos descrever rapidamente a distribuição de Cox. Ela pode ser representada por um conjunto de servidores exponenciais em série (ver figura 4.4) em que o usuário da classe jr tem uma certa probabilidade $a_{jr.}$ de passar ao próximo estágio. O total de estágios é u_{jr}. Vamos definir, para futura utilização, a quantidade A_{jrl} como sendo a probabilidade do usuário da classe r no centro j passar por l estágios da distribuição, isto é, $A_{jrl} = \prod_{f=1}^{l-1} a_{jrf}$.

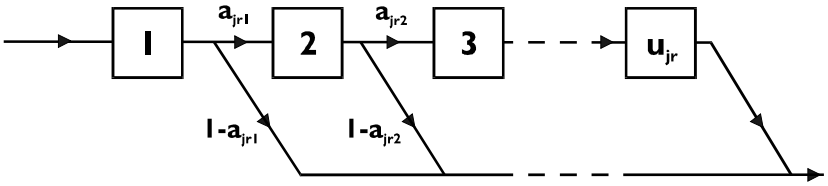

Figura 4.4: Distribuição de Cox

A família das distribuições de Cox é idêntica à família das distribuições que têm transformada de Laplace-Stieltjes racional. Erlang e hiperexponencial são membros dessa família e a soma de distribuições de Cox também tem distribuição de Cox. Uma outra propriedade importante assinala que qualquer distribuição pode ser aproximada por uma distribuição de Cox. Sabemos que uma função qualquer pode ser aproximada por uma função escada e esta, por sua vez, pode ser aproximada por uma soma de Erlangs-n (n grande).

Os centros de serviço podem ser de 4 tipos:

Tipo 1: A disciplina de serviço é FCFS, existe apenas um servidor que atende em tempo exponencial com parâmetro dependendo do centro e do número de usuários presentes, ou seja, para um centro j o parâmetro é $\mu_j(n_j)$.

Tipo 2: A disciplina é tempo compartilhado (processor sharing) e assim cada usuário recebe $1/k$ segundos de serviço se existem k no centro. O tempo de serviço tem distribuição de Cox que pode ser diferente para cada classe de usuário.

Tipo 3: O número de servidores é suficiente para sempre haver um livre quando um usuário chega (não há espera para serviço nesse centro). O tempo de serviço tem distribuição de Cox que pode ser diferente para cada classe de usuário.

Tipo 4: A disciplina de serviço é LCFS com interrupção. Um único servidor interrompe o atendimento para começar a atender o usuário que acabou de chegar. O usuário com serviço interompido fica na frente da fila esperando o fim do serviço em andamento para retomar a ser servido. Note que isto pode ocasionar uma cascata de serviços incompletos; o que deve ser enfatizado é que um usuário só completa seu serviço quando todos que chegaram depois dele completarem. Como nos centros tipo 2 e 3 o tempo de serviço tem distribuição de Cox que pode ser diferente para cada classe de usuário.

Definimos o estado da rede pelo vetor $\overline{x} = (x_1, x_2, \cdots, x_J)$ onde a interpretação de cada componente depende do tipo do centro. Para $j \in \mathcal{J}$,

$i)$ *se o centro j é do tipo 1:* $x_j = (x_{j1}, x_{j2}, \cdots, x_{jn_j})$, *onde x_{jl} é a classe do usuário na posição l do centro j e n_j é o total de usuários presentes em j;*

$ii)$ *se o centro j é do tipo 2 ou 3:* $x_j = (v_{j1}, v_{j2}, \cdots, v_{jI})$, *onde* $v_{jr} = (m_{jr1}, m_{jr2}, \cdots, m_{jru_{jr}})$ *e m_{jrl} representa o número de usuários do centro j que são da classe r*

e no l-ésimo estágio de seu serviço. O parâmetro u_{jr} é o número de estágios para usuários da classe r em serviço no centro j;

iii) se o centro j é do tipo 4: $x_j = ((r_1, m_1), (r_2, m_2), \cdots, (r_{n_j}, m_{n_j}))$ onde (r_n, m_n) é, respectivamente, a classe e o estágio do n-ésimo usuário na ordem LCFS.

Vamos definir as funções auxiliares $d(\overline{x})$ e $f_j(x_j)$ da seguinte forma:

$$d(\overline{x}) = \begin{cases} \prod_{l=0}^{n-1} \lambda(l)\,, & \text{para chegadas dependendo do total } n\,; \\ \prod_{\nu=1}^{m} \prod_{l=0}^{\tilde{n}_\nu - 1} \lambda_\nu(l)\,, & \text{para chegadas dependendo das partições;} \\ 1 & \text{para redes fechadas.} \end{cases}$$

Para $f_j(x_j)$ precisamos distinguir entre os diversos tipos de centros:

$$f_j(x_j) = \begin{cases} \left(1/\mu_j(n_j)\right)^{n_j} \prod_{l=1}^{n_j} e_{jx_{jl}}\,, & \text{se } j \text{ é do tipo 1;} \\ n_j! \prod_{r=1}^{I} \prod_{l=1}^{u_{jr}} \frac{\left(e_{jr} A_{jrl}\, \mu_{jrl}^{-1}\right)^{m_{jrl}}}{m_{jrl}!} & \text{se } j \text{ é do tipo 2;} \\ \prod_{r=1}^{I} \prod_{l=1}^{u_{jr}} \frac{\left(e_{jr} A_{jrl}\, \mu_{jrl}^{-1}\right)^{m_{jrl}}}{m_{jrl}!} & \text{se } j \text{ é do tipo 3;} \\ \prod_{l=1}^{n_j} e_{jr_l} A_{jr_l m_l}\, \mu_{jr_l m_l}^{-1} & \text{se } j \text{ é do tipo 4.} \end{cases}$$

Para cada t fixado, a variável aleatória $\overline{x}(t) = \overline{x}$ indica o estado da rede nesse instante. Apesar da notação complexa que pode dificultar a imediata constatação, pode-se verificar que o processo estocástico $\mathcal{X} = \{\overline{x}(t); t \geq 0\}$ é um processo de Markov. Note que a distribuição de Cox, com cada um dos estágios tendo tempo exponencial, preserva a característica Markoviana.

Teorema 4.7: Forma produto na rede BCMP

Para uma rede aberta, fechada ou mista, em que cada centro de serviço é do tipo 1, 2, 3 ou 4, as probabilidades estacionárias são dadas por:

$$\pi(x_1, x_2, \cdots, x_J) = b\, d(\overline{x}) f_1(x_1) f_2(x_2) \cdots, f_J(x_J)\,,$$

onde b é uma constante de normalização.

demonstração:

Não vamos fazer a demonstração, apenas indicamos que ela consiste em verificar que as *equações de balanço local* estão satisfeitas. Essas equações igualam, para cada centro, o fluxo (probabilidade $\times$ taxa) entre entradas e saídas de usuários de uma determinada classe e estágio de serviço. As equações são apresentadas a seguir, onde separamos os casos de fim de serviço daqueles correspondendo à mudança de estágio.

$\pi(\overline{x}) \times [$ taxa de saída do estágio l de usuários da classe r no centro j $] =$

$\sum_{\overline{x}'} \pi(\overline{x}') \times [$ taxa de entrada no estágio l de usuários da classe r no centro j $]$

e $\pi(\overline{x}) \times [$ taxa de saída do centro j para usuários da classe r $] =$

$\sum_{\overline{x}'} \pi(\overline{x}') \times [$ taxa de entrada no centro j para usuários da classe r].

Somando essas equações para todas as classes e estágios da distribuição de Cox, obtemos a equação de balanço global. Dessa forma, a solução da equação de balanço local será também solução na equação de balanço global (o contrário não vale).□

Comentários: A expressão da probabilidade estacionária, mesmo no caso do serviço ter a distribuição de Cox, envolve somente as médias do tempo de serviço (por estágios ou total). Tendo em vista a *forma produto*, a rede pode ser estudada separadamente por estações e isto facilita cálculos de desempenho. Como nos outros modelos de redes, o cálculo da constante de normalização fica bastante complexo com o aumento de estágios, centros de serviço e classes. Limitações no número de usuários, por classe ou globalmente, podem ser introduzidas afetando apenas o cálculo da constante de normalização.

Exemplo 4.5: (Gelenbe & Pujolle [1987])

Na transmisssão de mensagens é usual sua divisão em pequenos pedaços, referidos como *pacotes* na literatura especializada. Uma rede de comutação de pacotes dirige os usuários (pacotes) de um ponto inicial a um ponto final através de uma série de dispositivos (em geral computadores). Esses dispositivos comutam os pacotes dentre as várias linhas de transmissão e têm capacidade de armazená-los, se necessário. Os pacotes têm comprimento aleatório e as linhas de transmissão uma velocidade constante. Na modelagem por redes de filas, as linhas

de transmissão fazem o papel dos centros de serviço com os dispositivos de comutação fazendo o papel de sala de espera.

Na figura 4.5 está representada uma rede de comutação de pacotes com 7 pontos, representados pelas letras A' até G', indicando a origem e o destino dos pacotes. Vamos admitir que as chegadas de mensagens seguem um processo de Poisson distribuindo-se ao acaso entre os 7 terminais de entrada. Os dispositivos de comutação e armazenagem são microcomputadores representados pelas letras A até G. As linhas de transmissão entre esses microcomputadores pode ser modelada por dois centros de serviço, um para cada direção. Teremos assim um total de 16 centros de serviço, representados na figura por $c_1, c_2, \cdots, c_{16}$. Por exemplo, c_7 corresponde à transmissão de pacotes de C para E. A ligação entre os 7 terminais e os microcomputadores não é modelada por centros de serviço pois a velocidade da linha é muito rápida, indicando atendimento instantâneo em tempo zero para ambas as direções.

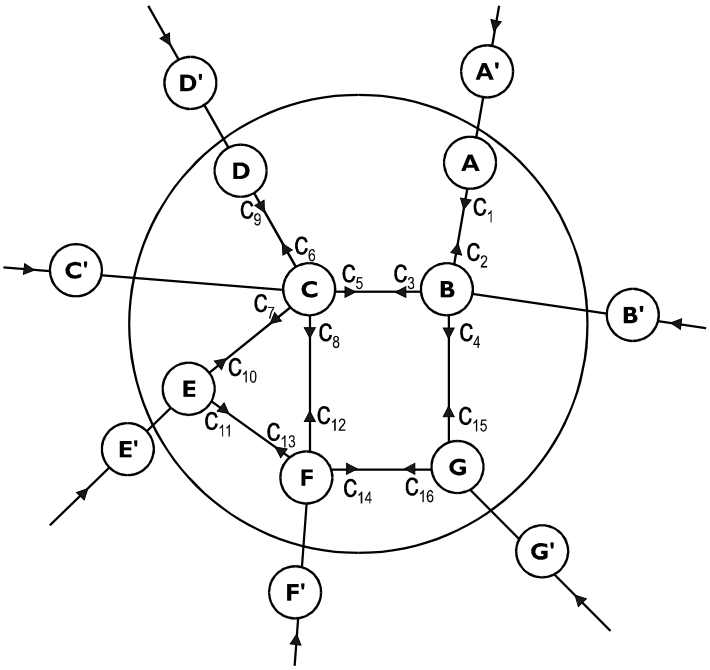

Figura 4.5: Exemplo de uma rede de comutação de pacotes

O roteamento dos pacotes dentro da rede é apresentado na tabela 4.1. Casos não mencionados na tabela, têm rotas escolhidas pelo menor caminho, entendido aqui como menor número de dispositivos de comutação intermediários.

A trajetória e topologia da rede definem 42 classes de usuários. Os pacotes são assumidos terem comprimento exponencial, o que implica que os tempos de serviço também são exponenciais, já que a velocidade das linhas é constante. O tempo médio de serviço é 0.25 segundos para todos os centros, exceto c_1, c_2, c_3 e c_5 em que o tempo médio é 0.33. A taxa de chegada λ é igual para todas as classes. Cálculos foram feitos para dois valores de λ: 0.143 e 0.214.

Tabela 4.1: Roteamento

Origem	Destino	Via
A'	F'	C
B'	F'	C
C'	G'	F
D'	G'	F
G'	C', D'	B
F'	A', B'	C

Utilizando a rede de BCMP podemos aplicar o teorema 4.7 para analisar o comportamento da rede. O número médio de pacotes presentes nos diversos centros (linhas de transmissão) é apresentado na tabela 4.2.

Tabela 4.2: Ocupação na rede de comutação de pacotes

centro	$\lambda = 0.143$	$\lambda = 0.214$
1	0.428	0.828
2	0.428	0.818
3	1	3
4	0.154	0.250
5, 6	0.667	1.5
7	0.464	0.66
8, 9	0.667	1.5
10	0.364	0.66
11	0.154	0.25
12	0.364	0.66
13	0.154	0.25
14,15	0.364	0.66
16	0.154	0.25

Pode-se assim verificar que a linha de transmissão correspondente ao centro de serviço 3 apresenta a maior taxa de ocupação. Essa indicação pode ajudar no estudo de eventuais gargalos que estariam produzindo atrasos exagerados na transmissão.□

Para outros exemplos de aplicação da rede BCMP, o leitor interessado pode consultar, além de Gelenbe & Pujolle [1987], as referências Allen [1990] e Kleinrock [1976].

4.5 Redes de estações quase-reversíveis

Vamos discutir nesta seção a propriedade de quase-reversibilidade em filas (isoladas) e a extensão do conceito para rede de filas. Faremos algumas suposições genéricas sobre o funcionamento da fila. Assumimos que todos os fregueses que entram, saem em algum momento do sistema e não mais de um evento (chegada ou saída) ocorre num instante t fixado. Cada freguês tem uma classe escolhida de um conjunto enumerável $\mathfrak{C}$ e não muda sua classe enquanto estiver no sistema.

Assumimos ainda que um processo de Markov a parâmetro contínuo $x(t)$ descreve o estado do sistema no instante t. Este processo tem informação suficiente para determinar quantos fregueses de cada classe existem no sistema.

Uma fila é chamada *quase-reversível* se seu estado $x(t)$ é um processo de Markov estacionário e o estado em t_0, $x(t_0)$, é independente dos tempos de chegada dos fregueses da classe $c \in \mathfrak{C}$, subsequentes à t_0 e também independente dos tempos de partida dos fregueses da classe $c \in \mathfrak{C}$, anteriores à t_0.

Alguns exemplos de filas que são quase-reversíveis são a M/M/1, a M/G/1/PS (*processor sharing*), a M/G/1/LCFS com interrupção e a M/G/∞. A verificação desses resultados pode ser encontrada em Walrand [1988].

Apresentamos a seguir, sem demonstração, algumas propriedades de uma fila quase-reversível.

Teorema 4.8

Se uma fila é quase-reversível então:
$i)$ Os tempos de chegada dos fregueses da classe $c \in \mathfrak{C}$ formam processos de Poisson independentes;
$ii)$ Os tempos de partida dos fregueses da classe $c \in \mathfrak{C}$ formam processos de Poisson independentes.

demonstração:

Ver Kelly [1979].□

Como consequência do teorema acima, temos que, numa fila quase-reversível, valem as equações de balanço local e a taxa de chegada de fregueses da classe c depende apenas de c, mas não depende do estado da fila. Lembrando que reversibilidade requeria a verificação das equações de balanço detalhado, podemos estabelecer as diferenças entre os dois conceitos. Quase-reversibilidade requer uma condição mais forte do que reversibilidade sobre as taxas de chegadas e mais fraca sobre o fluxo de probabilidade.

Vamos nos ocupar agora da construção de uma rede aberta de estações quase-reversíveis. A idéia é partir de estações que isoladamente são quase-reversíveis para estabelecer uma regra de transição (e movimento) para toda a rede.

Com a mesma notação definida na seção 4.3, seja (i, s) a classe do freguês de tipo i no estágio s da sua rota e $k(i, s)$ o centro visitado pelo freguês tipo i no s-ésimo estágio da sua rota. Representamos por $\pi_j(x_j)$ a distribuição de equilíbrio de uma fila j, que é quase-reversível e tem chegadas de fregueses da classe (i, s) formando um processo de Poisson de taxa $\alpha_j(i, s)$. Sejam $q_j(x_j, y_j)$ as taxas de transição nessa fila; denotamos por $\mathcal{S}_j(i, s, x_j)$ o conjunto de estados que, em relação ao estado x_j, contém um freguês a mais da classe (i, s) e o mesmo número de fregueses das outras classes.

Construimos um processo de Markov $X(t) = (x_1(t), x_2(t), \cdots, x_J(t))$ com as seguintes taxas de transição:

$i)$ $q_h(x_h, y_h)$: para entradas no sistema de freguês do tipo i ao centro $h = k(i, 1)$, causando mudança do estado x_h para $y_h \in \mathcal{S}_h(i, s, x_h)$;

$ii)$ $q_j(y_j, x_j)$: para saídas do sistema de freguês do tipo i que causa ao centro $j = k(i, S(i))$ a transição do estado $y_j \in \mathcal{S}_j(i, s, x_j)$ para x_j;

$iii)$ $q_j(y_j, x_j)\frac{q_h(x_h,y_h)}{\alpha_h(i,s+1)}$: para saídas de freguês da classe (i, s), $s < S(i)$, da fila $j = k(i, s)$ entrando na fila $h = k(i, s + 1)$ como freguês classe $(i, s + 1)$; a fila j troca seu estado de $y_j \in \mathcal{S}_j(i, s, x_j)$ para x_j e a fila h troca de x_h para o estado $y_h \in \mathcal{S}_h(i, s + 1, x_h)$;

$iv)$ $q_j(y_j, x_j)$: para transições internas à fila j (sem saída nem entrada de fregueses).

A rede definida por esses J centros e caracterizada pelo processo de Markov $X(t)$ é chamada de *rede aberta de estações quase-reversíveis*. Note que, em geral, as estações podem não obedecer as condições para serem quase-

reversíveis. O nome está, assim, associado à construção da rede, que foi efetuada a partir de estações que, isoladamente, são quase-reversíveis.

Pode-se verificar que a distribuição de equilíbrio é dada por

$$\pi(x_1, x_2, \cdots, x_J) = \pi(x_1)\pi(x_2)\cdots\pi(x_J),$$

valendo assim a *forma produto*.

Pode-se verificar que as redes de Jackson, BMCP e Kelly são exemplos de redes que são quase-reversíveis.

O próximo teorema é apresentado sem demonstração e sumariza os principais resultados.

Teorema 4.9

A rede aberta de estações quase-reversíveis tem as seguintes propriedades:
$a)$ Os estados das estações individuais são independentes.
$b)$ Para uma estação individual e freguês de uma dada classe, a distribuição de equilíbrio e a distribuição encontrada por um freguês chegando, são idênticas; ambas são ainda iguais à distribuição que teria a estação, quando estivesse isolada das demais, com chegadas de fregueses de cada classe formando processos de Poisson.
$c)$ Sob tempo reverso, o sistema se torna uma outra rede aberta de estações quase-reversíveis.
$d)$ O próprio sistema é quase-reversível e, portanto, saídas do sistema para fregueses de cada tipo formam processos de Poisson independentes; o estado do sistema no instante t_0 é independente das saídas do sistema anteriores a t_0.

demonstração:

Ver Kelly [1979].□

4.6 Exercícios

1) Considere duas filas M/M/1 em série, isto é, as saídas da primeira são as entradas da segunda. Verifique que esse sistema pode ser considerado uma rede de Jackson e identifique os diversos parâmetros do modelo.

2) Para o sistema apresentado no exercício 1, determine o número médio de usuários e o tempo médio de espera no sistema, se $\lambda_1 = 1$ e $\mu_1 = \mu_2 = 2$.

3) Verifique se o sistema apresentado no exercício 1 é reversível.

4) (Gelenbe & Pujolle [1987]) Uma rede aberta de Jackson tem dois centros de serviço com um único servidor, que atende com taxa μ_i, $i = 1, 2$. Só há entradas pelo centro 1 e a matriz de roteamento é dada por (incluimos o centro 3 para representar o exterior da rede)

$$R = \begin{bmatrix} 0 & p_1 & 1-p_1 \\ p_2 & 0 & 1-p_2 \\ 1 & 0 & 0 \end{bmatrix}.$$

Escreva as equações de balanço global, local e detalhado. Verifique se a rede é reversível.

5) Considere uma rede fechada e cíclica com três estações (o usuário percorre sucessivamente as estações 1, 2 e 3). Para um total de 3 usuários no sistema e serviço exponencial de parâmetros $\mu_1 = \mu_2 = \mu_3 = 1$, determine a distribuição estacionária.

6) Numa rede fechada com três estações e 2 usuários temos a seguinte matriz de roteamento:

$$R = \begin{bmatrix} 0 & 0.2 & 0.8 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

O serviço em cada centro é exponencial com parâmetro 1. Determine a distribuição estacionária do sistema. Qual o número médio de usuários na estação 1?

7) (Kelly [1979]) Para um centro de serviço j na rede de Kelly contendo c servidores mostre que:
a) O centro terá disciplina de serviço LCFS sem interrupção, se escolhermos

$$\delta_j(l, n_j) = \begin{cases} 1 & l = n_j\,;\; n_j = 1, 2, \cdots, c \\ 1 & l = c+1;\; n_j = c+1, c+2, \cdots. \\ 0 & \text{caso contrário} \end{cases}$$

b) O serviço será em ordem aleatória se

$$\delta_j(l, n_j) = \frac{1}{n_j - c}\;,\; l = c+1, c+2, \cdots, n_j;\; n_j = c+1, c+2, \cdots.$$

8) (Kelly [1979]) Verifique que, na rede apresentada no exemplo 4.3, o processo definido pelo vetor $(n_1, n_2, n_3, n_4, n_5)$ não é um processo de Markov (falta acrescentar o tipo de usuário na definição do espaço de estados).

9) (Kelly [1979]) Suponha que $\phi_j(n_j) = \phi_j$ para $n_j > 0$ e $j = 1, 2, \cdots, J$ (dessa forma, cada fila tem um único servidor). Para um centro j qualquer:
a) Calcule o número médio de usuários.
b) Determine a distribuição de probabilidade do número de usuários do tipo i e no estágio s da sua rota.

10) (Wolf [1989]) Considere uma rede com dois centros e dois tipos de fregueses. O tipo 1 percorre as estações 1, 2 e 1, enquanto o tipo 2 tem rota 2, 1 e 2; as chegadas externas para os dois tipos são Poisson com parâmetros λ_1 e λ_2, respectivamente. O serviço é exponencial com parâmetro μ_i para a estação i e a disciplina de atendimento é FCFS. Mostre que:
a) $P(N_i = n) = (1 - \rho_i)\rho_i^n$, para $i = 1, 2$, com N_i sendo o número de usuários no centro i, $\rho_i = a_i/\mu_i$ e $n \geq 0$.
b) Dado $N_1 > 0$, a probabilidade de que, um usuário tipo 1 no estágio 1 da sua rota esteja na estação 1, vale λ_1/a_1.

11) (Baskett, Chandy, Muntz & Palacios [1975]) Considere a seguinte rede BCMP fechada, com dois centros e duas classes de usuários. O centro 1 é do tipo 3 e o centro 2, do tipo 2. O número de usuários nas classes 1 e 2 são M_1 e M_2, respectivamente. Os serviços são todos exponenciais com parâmetro μ_{ir} com i, sendo o centro e r a classe do usuário. Note que, nesse caso, a distribuição de Cox tem só um estágio. Para o movimento na rede, temos

$$p_{1,2;2,2} = p_{2,2;1,2} = p_{2,1;1,1} = 1 \ , \ p_{1,1;1,1} + p_{1,1;2,1} = 1,$$

com todos os outros elementos da matriz P sendo zero. Defina n_{ir} como o número de usuários da classe r no centro i. Escreva as equações de balanço global e local, no caso $n_{ir} > 0, \ i, r = 1, 2$.

Apêndice A

Integral de Riemann-Stieltjes e Transformada de Laplace-Stieltjes

O material apresentado aqui é um sumário dos principais resultados sobre a integral de Riemann-Stieltjes e a transformada de Laplace-Stieltjes. Maiores detalhes podem ser encontrados em Rudin [1976], Kleinrock [1975], Cooper [1981] e Gross & Harris [1985].

A integral de Riemann, familiar ao leitor, não contém a generalidade necessária para tratar com funções de distribuição que não sejam absolutamente contínuas. A integral apresentada aqui permite uma abordagem unitária para as variáveis aleatórias, sejam discretas ou contínuas, o que é conveniente para resolver vários problemas em teoria das filas.

Antes de apresentar a integral de Riemann-Stieltjes precisamos de algumas preliminares. Sejam $\alpha(t)$ uma função monótona não decrescente em t no intervalo $[a, b]$ e $\beta(t)$ uma função qualquer definida no mesmo intervalo. Em geral, na área de probabilidade e estatística, a função $\alpha(t)$ será uma função de distribuição. Para uma sequência de pontos $a = t_0 < t_1 < t_2 < \cdots < t_n = b$ que particiona o intervalo $[a, b]$, consideremos a norma $\|\Delta\|$, definida da seguinte forma:

$$\|\Delta\| = \underset{1 \leq i \leq n}{max} (t_i - t_{i-1}),$$

isto é, a norma indica o comprimento do maior sub-intervalo.

Definição a1: Integral de Riemann-Stieltjes

A integral de Riemann-Stieltjes (RS) da função $\beta(t)$ com respeito à função $\alpha(t)$ no intervalo $[a, b]$ é definida por:

$$\int_a^b \beta(t)\, d\alpha(t) = \lim_{\|\Delta\| \to 0} \sum_{i=1}^{n} \beta(\xi_i) \Big[\alpha(t_i) - \alpha(t_{i-1})\Big],$$

com $t_{i-1} \leq \xi_i \leq t_i$, $i = 1, 2, \cdots, n$.□

O lado esquerdo da expressão acima é a notação usualmente utilizada para representar a integral RS. É importante observar que o limite inferior, a, está incluído na integral, de modo que um eventual valor positivo de $\beta(a)\alpha(a)$ é computado. Uma notação mais precisa seria substituir a por a^-, o que não faremos, para não sobrecarregar a notação (essa é também a opção da maioria dos autores). Uma condição suficiente para a existência do limite é que $\beta(t)$ seja contínua em $[a, b]$ e, em geral, esse são os casos de nosso interesse.

Se $\alpha(t) = t$, então a integral RS torna-se uma simples integral de Riemann. Se $\alpha(t)$ é a função distribuição $F(t)$ de uma variável aleatória não negativa X, com densidade $f(t)$, a integral RS da função $\beta(t)$ em relação à F torna-se

$$\int_0^\infty \beta(t)\, dF(t) = \int_0^\infty \beta(t)\, f(t) dt = E(\beta(X)),$$

com essa última igualdade valendo pela definição de valor esperado.

Várias das propriedades da integral de Riemann também são válidas nas integrais RS. Em particular,

$$\int_a^b d\alpha(t) = \alpha(b) - \alpha(a);$$

$$\int_a^b [k_1\beta_1(t) + k_2\beta_2(t)]\, d\alpha(t) = \int_a^b k_1\beta_1(t)\, d\alpha(t) + \int_a^b k_2\beta_2(t)\, d\alpha(t),$$

com k_1 e k_2 constantes arbitrárias.

Se $\beta(t)$ é contínua e $\alpha(t)$ é a função de distribuição de uma variável aleatória discreta com saltos nos pontos $c_1 < c_2 < \cdots < c_n$ pertencentes à $[a, b]$, então

$$\int_a^b \beta(t)\, d\alpha(t) = \sum_{i=1}^{n} \beta(c_i)\, [\alpha(c_i^+) - \alpha(c_i^-)]\ .$$

Apresentamos a seguir a fórmula de integração por partes para as integrais RS. Para $\beta(t)$ monótona não decrescente, temos,

$$\int_a^b \beta(t)\, d\alpha(t) + \int_a^b \alpha(t)\, d\beta(t) = \beta(b)\, \alpha(b) - \beta(a)\, \alpha(a).$$

Utilizando a integral RS, vamos definir uma transformação que é de grande auxílio na teoria das probabilidades e, em especial, em teoria das filas.

Definição a2: Transformada de Laplace-Stieltjes

Seja X uma variável aleatória não negativa com função de distribuição F. A transformada de Laplace-Stieltjes (LS) de F é dada por:

$$E(e^{-sX}) = \tilde{F}(s) = \int_0^{\infty} e^{-st}\, dF(t),\ Re(s) \geq 0\,.\square$$

Note que s, na definição acima, pode ser um número complexo e o extremo inferior da integral é 0^-, como já mencionado anteriormente.

A transformada LS facilita operações que desejamos fazer com as funções originais. Além disso, existe correspondência um-a-um entre a função original e a transformada. Essas são características desejáveis em qualquer transformada e se verificam, por exemplo, em funções geradoras de probabilidade e de momentos. Vamos discutir em seguida algumas propriedades da transformada LS.

Sejam F_1 e F_2 duas funções de distribuição e consideremos a operação de convolução entre elas, definida por

$$G(t) = F_1 * F_2(t) = \int_0^t F_1(t-x)\, dF_2(x) = \int_0^t F_2(t-x)\, dF_1(x)\,,$$

onde $*$ indica a operação de convolução. Observe que, pela definição acima, a convolução é uma operação comutativa que resulta em uma outra função de distribuição, $G(t)$. É possível verificar que a transformada LS de $G(t)$ será o produto das transformadas de F_1 e F_2, isto é,

$$\tilde{G}(s) = \tilde{F}_1(s)\tilde{F}_2(s).$$

Os momentos da distribuição $F(t)$ podem ser calculados a partir de $\tilde{F}(s)$, através de sucessivas diferenciações,

$$E(X^n) = \int_0^{\infty} t^n\, dF(t) = (-1)^n \Big(\frac{d^n}{ds^n}\, \tilde{F}(s) \Big) \Big|_{s=0}.$$

Em geral, após realizar as operações desejadas, precisamos retornar do domínio das transformadas para o domínio das funções de distribuição. Nesses casos, é útil estabelecer a relação entre a transformada de Laplace-Stieltjes e a de Laplace, pois esta última tem tabelas de inversão facilmente encontráveis em

textos de probabilidade aplicada. Inicialmente, relembramos a definição de transformada de Laplace,

$$L(F(t)) = \int_0^{\infty} e^{-st}\, F(t)dt\,.$$

Aplicando a integração por partes, temos

$$\int_0^b e^{-st}\, dF(t) - s\int_0^b e^{-st}\, F(t)dt = e^{-sb}\, F(b) - e^{-s0}\, F(0^-).$$

Assumindo $F(0^-) = 0$ e tomando limite para $b \to \infty$, vem

$$L(F(t)) = \frac{1}{s}\, \tilde{F}(s)\ .$$

Dessa forma, para efetuar a inversão de uma transformada LS pode-se usar as tabelas da transformada de Laplace.

Apêndice B

Principais Processos Estocásticos Utilizados em Filas

1. Processos de Renovação

Considere um experimento aleatório qualquer e sejam $W_1, W_2, \cdots$ as variáveis aleatórias representando os intervalos de tempo entre sucessivas ocorrências. Os tempos

$$T_0 = 0 \text{ e } T_{n+1} = T_n + W_{n+1}, \text{ para } n \geq 0,$$

definem os instantes das sucessivas ocorrências.

Definição b1: Processo de Renovação

Se $W_1, W_2, \cdots$ são independentes e identicamente distribuídos diremos que a sequência $\mathcal{T} = \{T_n; n \geq 0\}$ é um processo de renovação.□

Vários autores preferem definir o processo de renovação através do processo de contagem do número de ocorrências até um instante t. Pode-se demonstrar que as duas definições são equivalentes. O processo de contagem é definido por

$$N(t) = \sum_{n=1}^{\infty} I_{[0,t]}(T_n) \,,$$

com

$$I_{[0,t]}(T_n) = \begin{cases} 1\,, & 0 \leq T_n \leq t; \\ 0\,, & \text{caso contrário.} \end{cases}$$

A distribuição de $N(t)$ pode ser obtida da distribuição dos W_n da seguinte forma:

$$P(N(t) = k) = P(T_{k-1} \leq t) - P(T_k \leq t) = F^{k-1}(t) - F^k(t)\,,$$

com $F(t)$ sendo a função distribuição de W_n e $F^k(t)$ representanto a k-ésima convolução de F consigo mesma. Note que, o uso de transformadas de Laplace-Stieltjes é de grande auxílio, pois a operação de convolução se torna produto de transformadas.

O "mais famoso" processo de renovação é o processo de Poisson. Nesse caso, o intervalo entre ocorrências tem distribuição exponencial e o processo de contagem segue a distribuição de Poisson. Isto é,

$$F(t) = 1 - e^{-\lambda t}\,, t \geq 0\,, \lambda > 0\,,$$

$$\text{e } P(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}\,,\ k = 0, 1, 2, \cdots.$$

2. Cadeias e Processos de Markov

Definição b2: Cadeias de Markov

A sequência de variáveis aleatórias $X_1, X_2, \cdots$, assumindo valores num conjunto E, forma uma cadeia de Markov (em tempo discreto) se para todo $n = 1, 2, \cdots$ temos,

$$P(X_n = j|\, X_1 = i_1, X_2 = i_2, \cdots, X_{n-1} = i) = P(X_n = j|X_{n-1} = i),$$

para toda escolha de valores $j, i_1, i_2, \cdots, i\ \in E$.□

A igualdade acima estabelece que a dependência da evolução do processo fica reduzida à última ocorrência. Ou, em outras palavras, o futuro depende apenas do presente, mas não do passado. Um processo estocástico com essa propriedade é dito ter a *propriedade de Markov*.

O conjunto E é chamado de *espaço de estados da cadeia.* Vamos assumir (o que usualmente é feito) que a cadeia é homogênea e as probabilidades não dependem do índice n. Assim, definimos a matriz de transição da cadeia, em uma etapa, por

$$P = [p_{ij}], i, j \in E\,, \text{ com } p_{ij} = P(X_n = j|X_{n-1} = i), \forall n \geq 1.$$

A cadeia de Markov será irredutível, se todo estado da cadeia pode ser alcançado a partir de qualquer outro, em uma ou mais etapas. Para classificar os estados da cadeia, vamos definir a probabilidade de retorno a um estado. Sejam,

$$f_j^{(n)} = P(\text{ partindo de } j,\ 1^{\underline{o}} \text{ retorno a } j \text{ ocorre na etapa } n),$$

$$f_j = \sum_{n=1}^{\infty} f_j^{(n)} = P(\text{partindo de } j, \text{ retornar a } j \text{ em alguma etapa}).$$

O estado j será dito recorrente se $f_j = 1$ e transitório se $f_j < 1$. Se os retornos a um estado j se dão apenas em um número de etapas que é múltiplo de δ, então j é dito ser periódico de período δ. Se $\delta = 1$, o estado é aperíodico. Para os estados recorrentes podemos calcular o tempo médio de recorrência pela expressão

$$\mu_j = \sum_{n=1}^{\infty} n f_j^{(n)} .$$

Se μ_j é finito, j é dito recorrente não nulo, caso μ_j seja infinito o estado será denominado recorrente nulo.

A partir de uma distribuição inicial $\pi^{(0)} = (\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}, \cdots)$, o sistema evolui fazendo transições de acordo com a matriz P. Na etapa n, a chance da cadeia estar em cada estado será dada por $\pi^{(n)}$, com elementos $\pi_j^{(n)}$, $j \in E$. Definimos a distribuição estacionária da cadeia como sendo o vetor de probabilidades π que, uma vez escolhido como distribuição inicial, implicará que teremos $\pi^{(n)}$ igual à distribuição estacionária, para qualquer etapa n. Podemos então dizer que ao começar com a distribuição estacionária ficamos com ela para sempre.

Apresentamos a seguir, sem demonstração, duas importantes proposições:

Proposição b1:

Os estados de uma cadeia de Markov irredutível são todos do mesmo tipo. Se forem periódicos, todos têm o mesmo período.□

Proposição b2:

Em uma cadeia de Markov homogênea, irredutível e aperiódica, a distribuição de probabilidade limite,

$$\eta_j = \lim_{n \to \infty} \pi_j^{(n)}$$

existe para todo $j \in E$ e é independente da distribuição de probabilidade inicial.

Temos ainda que uma das seguintes situaçõesse verifica:

a) Todos os estados são transitórios ou todos são recorrentes nulos. Nesse caso não existe distribuição estacionária e $\eta_j = 0$ para todo $j \in E$.

b) Todos os estados são recorrentes não nulos e $\eta_j > 0$ para todo $j \in E$. Nessa situação os η_j's coincidem com a distribuição estacionária e $\eta_j = 1/\mu_j$. O sistema de equações

$$\sum_i x_i = 1\,;$$

$$x_j = \sum_{i \in E} x_i\, p_{ij}\ ,\ j \in E,$$

determinará unicamente as duas distribuições (estacionária e limite).□

Definição b3: Processo de Markov

Um processo estocástico $\{X(t); t \geq 0\}$, com espaço de estados enumerável E, finito ou infinito, é um processo de Markov (ou cadeia de Markov a parâmetro contínuo) se, para todo $t, s \geq 0$ e estado $j \in E$,

$$P(X(t+s) = j |\, X(u); u \leq s) = P(X(t+s) = j |\, X(s))\ .\square$$

Como fizemos no caso discreto, vamos assumir que o processo é homogêneo, isto é, podemos escrever a probabilidade de transição

$$P(X(t+s) = j |\, X(s) = i) = p_{ij}(t)\ ,$$

não dependendo de s, mas apenas dos estados i , j e da duração do intervalo t.

Pode-se demonstrar que, num processo de Markov, o tempo de permanência em um estado antes da transição tem distribuição exponencial. O parâmetro dessa exponencial pode depender do estado presente, mas não do que será visitado a seguir.

Usualmente, a definição do processo se dá através das *taxas de transição infinitesimal*, dadas por

$$q_{ii} = \lim_{\Delta t \to 0} \frac{p_{ii}(\Delta t) - 1}{\Delta t}$$

$$q_{ij} = \lim_{\Delta t \to 0} \frac{p_{ij}(\Delta t)}{\Delta t}\ ,\ i \neq j\,.$$

A matriz $\Lambda = [q_{ij}]$, com $i, j \in E$, é denominada de *gerador infinitesimal*. Note que, q_{ii} é negativo com valor igual ao oposto da soma dos outros elementos da linha i.

Um processo de Markov é *irredutível* se, para todo $i, j \in E$, $p_{ij}(t)$ é positivo para algum t. Em outras palavras, o processo é irredutível se existe probabilidade positiva de alcançar qualquer estado, partindo de qualquer estado, em algum tempo.

Se o processo é irredutível, então a distribuição limite existe e é independente da distribuição inicial, isto é, $\lim_{t \to \infty} p_{ij}(t)$ existe para todo $j \in E$.

Para calcular a distribuição estacionária Π, resolve-se a equação vetorial

$$\Pi\Lambda = 0\,,$$
$$\Pi e = 1\,,$$

com e sendo um vetor coluna de $1's$. Se esse sistema tiver solução única, então a distribuição limite coincide com a distribuição estacionária. Em geral, essa é a situação de maior interesse prático.

3. Processos de Nascimento e Morte

Vamos descrever um caso especial dos processos de Markov, denominado processo de nascimento e morte. O nome tem origem nos estudos de população animal realizados em experimentos biológicos. Vamos considerar aqui apenas o caso contínuo, por ser o de maior aplicação em Teoria das Filas.

Considere um processo de Markov com transições apenas entre estados adjacentes e com as seguintes taxas,

$$q_{ij} = \begin{cases} \lambda_i, & \text{se } j = i+1; \\ \mu_i, & \text{se } j = i-1,\ i > 0; \\ -(\lambda_i + \mu_i), & \text{se } j = i,\ i > 0; \\ -\lambda_0, & \text{se } j = i = 0. \end{cases}$$

Note que as taxas λ_i correspondem a saltos "para cima" enquanto μ_i a saltos "para baixo". Usualmente, dizemos que λ_i é a taxa de nascimento e μ_i a taxa de morte do processo.

O gerador infinitesimal é dado por

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

4. Processos de Renovação Markovianos

O processo que vamos discutir agora inclui, como caso particular, todos os processos apresentados nas seções anteriores. O processo de renovação Markoviano combina a teoria da renovação com a teoria das cadeias de Markov. Os tempos de permanência em cada estado têm distribuição geral, dependendo do estado presente e do próximo a ser visitado e a sequência de estados visitados forma uma cadeia de Markov.

Para cada $n = 0, 1, \cdots$, considere o par de variáveis aleatórias (X_n, T_n) com X_n tomando valores num conjunto enumerável E e T_n em $[0, \infty)$ sendo $0 = T_0 \leq T_1 \leq \cdots$.

Definição b4: Processo de Renovação Markoviano

O processo estocástico $(X, T) = \{(X_n, T_n); n \geq 0\}$ é um processo de renovação Markoviano, com espaço de estados E, se

$$P(X_n = j, T_n - T_{n-1} \leq t \,|(X_k, T_k), k \leq n-1) = P(X_n = j, T_n - T_{n-1} \leq t \,|X_{n-1}),$$

para todo $j \in E$, $n \geq 1$ e $t \geq 0$.□

Pela definição acima observamos que a transição futura do processo só depende do presente, e não de toda história do processo. Como usual, vamos assumir que o processo é homogêneo e assim definir o núcleo de transição por $Q(t) = [Q_{ij}(t)], i, j \in E,$ com

$$Q_{ij}(t) = P(X_n = j, T_n - T_{n-1} \leq t \,|X_{n-1} = i)\,, \forall n > 0.$$

Um processo de renovação Markoviano torna-se um processo de Markov, se o tempo de permanência em cada estado é exponencial com parâmetro não dependendo do próximo estado a ser visitado. Ele será uma cadeia de Markov, se o tempo entre transições for constante e igual a 1. Temos ainda que o processo de renovação Markoviano torna-se um processo de renovação, no caso do espaço de estados conter apenas 1 elemento.

Se considerarmos apenas a sequência de estados visitados pelo processo, ou seja, se estamos interessados em $\{X_n; n \geq 0\}$, obtemos uma cadeia de Markov, que é denominada a *cadeia imersa* no processo (X, T). A matriz de transição, em uma etapa, dessa cadeia é $P = [p_{ij}],\ i, j \in E$, com

$$p_{ij} = \lim_{t \to \infty} Q_{ij}(t)\,.$$

Sob algumas circunstâncias, quando isolamos a componente T do processo de renovação Markoviano, o processo resultante será um processo de renovação. Esse é o conceito de equivalência, definido a seguir.

Definição b5: Equivalência

O processo de renovação Markoviano (X, T) será equivalente ao processo de renovação $\mathcal{T}$ com distribuição $F,$ se, para todo $n \in \mathbb{N}$ e $t_1, t_2 \cdots t_n \in \mathbb{R}_+$, tivermos

$$\Pi Q(t_1)Q(t_2)\cdots Q(t_n)e = F(t_1)F(t_2)\cdots F(t_n)\,,$$

com Π sendo a distribuição estacionária da cadeia imersa.□

A equivalência requer que, qualquer distribuição conjunta de n intervalos entre realizações sucessivas do processo (X,T), sejam independentes. Tendo em vista a dificuldade de aplicação prática da definição acima, apresentamos a seguir um critério para equivalência.

Proposição b3:

O processo de renovação Markoviano (X,T) é equivalente ao processo de renovação $\mathcal{T}$, se uma das seguintes condições se verificam

$$i)\ \Pi Q(t) = F(t)\Pi$$
$$ii)\ Q(t)e = e\,F(t).\ \square$$

Essas condições são utilizadas frequentemente na verificação de características dos processos de saída ou de partida em filas. Elas também estão relacionadas com a propriedade de *quase-reversibilidade*, apresentada no capítulo 4. Consulte Disney & Kiessler [1987] para maiores detalhes.

Apêndice C

Processos Reversos e Reversibilidade

A idéia de processo reverso constitui-se numa ferramenta importante para o estudo de redes de filas. Apresentaremos as principais definições e resultados e indicamos ao leitor que consulte Disney & Kiessler [1987], Kelly [1979] e as referências lá mencionadas, para maiores detalhes.

Definição c1: Processo Reverso

Seja $\mathcal{X} = \{X(t);\ -\infty < t < +\infty\}$ um processo estocástico qualquer. Para um número real τ, $\hat{\mathcal{X}} = \{X(\tau - t);\ -\infty < t < +\infty\}$ é definido como o processo reverso de $\mathcal{X}$.□

Podemos interpretar o processo reverso como sendo a evolução do processo no sentido contrário do tempo original.

Definição c2: Processo reversível

Um processo será dito reversível se ele, e seu reverso, têm a mesma distribuição probabilística. Nesse caso, o processo direto é estatisticamente indistinguível do seu reverso.□

A partir dessas definições gerais, vamos apresentar resultados para as cadeias e processos de Markov. Começamos por indicar como obter os processos reversos.

Seja $\mathcal{X} = \{X_n; n \in \mathbb{N}\}$ uma cadeia de Markov estacionária com espaço de estados E, matriz de transição P com elementos p_{ij} e distribuição estacionária dada por Π. A cadeia reversa $\hat{\mathcal{X}}$ é a cadeia de Markov com o mesmo espaço de estados E e com matriz de transição $\widehat{P}$, cujos elementos são

$$\widehat{p}_{ij} = \frac{\pi_j}{\pi_i} p_{ji}\ ,\ \forall i, j \in E.$$

Considere agora um processo de Markov $\mathcal{X} = \{X(t);\ t \in \mathbb{R}\}$, estacionário, com gerador Λ e distribuição estacionária Π. O reverso de $\mathcal{X}$ é o processo de Markov $\hat{\mathcal{X}}$ que tem gerador $\hat{\Lambda}$, cujos elementos são

$$\hat{\Lambda}_{ij} = \frac{\pi_j}{\pi_i}\ \Lambda_{ji}\ ,\ \forall i, j \in E.$$

Note que, por construção, o processo reverso têm a mesma distribuição estacionária do processo direto. Isto concorda com a interpretação da distribuição estacionária como sendo a proporção de permanência nos estados do processo. Em várias situações em que a verificação das equações de equilíbrio é complicada, mas o processo reverso e suas taxas são fáceis de serem "chutados", podemos ter um grande auxílio no cálculo da distribuição estacionária. O resultado a seguir indica como podemos tirar proveito dessa situação. A versão que apresentaremos é para os processos de Markov. A demonstração e o resultado análogo para cadeias podem ser encontrados em Kelly [1979].

Proposiçao c1:

Seja $\mathcal{X}$ um processo de Markov estacionário com taxas $\Lambda_{ij}, i, j \in E$. Se podemos obter uma coleção de números $\hat{\Lambda}_{ij}, i, j \in E,$ tais que

$$\sum_{j \in E} \Lambda_{ij} = \sum_{j \in E} \hat{\Lambda}_{ij}\ ,$$

e uma coleção de números positivos π_j, $j \in E$, que somem $1,$ tais que

$$\pi_i\ \Lambda_{ij} =\ \pi_j\ \hat{\Lambda}_{ji}\ ,\ \forall i, j \in E\ ,$$

então $\hat{\Lambda}_{ji}\ ,\ i, j \in E$, são as taxas de transição do processo reverso $\hat{\mathcal{X}}$ e π_j, $j \in E$, é a distribuição estacionária dos dois processos (direto e reverso). □

Apresentamos a seguir, sem demonstração, alguns critérios para reversibilidade em cadeias e processos de Markov.

Proposição c2:

Uma cadeia ou processo de Markov estacionário é reversível se e só se estão satisfeitas as equações de balanço detalhado. Isto é,

$i)\ \pi_i\ p_{ij} =\ \pi_j\ p_{ji}\ ,\ \forall i, j \in E$, para as cadeias;
$ii)\ \pi_i\ \Lambda_{ij} =\ \pi_j\ \Lambda_{ji}\ ,\ \forall i, j \in E$, para os processos. □

Proposição c3: Critério de Kolmogorov

Uma cadeia de Markov estacionária é reversível se e, só se, suas probabilidades de transição satisfazem

$$p_{j_1 j_2}\, p_{j_2 j_3} \cdots p_{j_{n-1} j_n}\, p_{j_n j_1} = p_{j_1 j_n}\, p_{j_n j_{n-1}} \cdots p_{j_3 j_2}\, p_{j_2 j_1},$$

para qualquer sequência finita de estados $j_1, j_2, \cdots, j_n \in E$. □

O critério de Kolmogorov estabelece que, para qualquer trajetória que comece e termine no mesmo estado, a probabilidade da sua realização é a mesma não importanto a direção do movimento. O critério é muito útil também para uma rápida verificação de não reversibilidade. Isto em geral é feito escolhendo-se trajetórias críticas do processo. A versão para processos de Markov é análoga.

Bibliografia

ALLEN, A. O. (1990). *Probability, Statistics and Queueing Theory with Computer Science Applications.*2a. ed. New York: Academic Press.

BASKETT, F., CHANDY, M., MUNTZ, R., PALACIOS, J. (1975). Open, closed, and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.* 22:248-260.

BUNDAY, B. D. (1986). *Basic Queueing Theory*. London: E. Arnold Publishers.

BURKE, P. J. (1956). The output of a queueing system. *Oper. Res.* 4:699-704.

CLARKE, A. B., DISNEY, R. L. (1985). *Probability and Random Processes: A first course with applicattions.*2a. ed. New York: John Wiley (Wiley Series in Probability and Mathematical Statistics).

COOPER, R. B. (1981). *Introduction to Queueing Theory*. 2a. ed. New York:North Holland.

ÇINLAR, E. (1975). *Introduction to Stochastic Processes.* Englewood Cliffs, New Jersey: Prentice-Hall.

DAIGLE, J. (1991). *Queueing Theory for Computer Communications.* New York: Addison-Wesley.

DISNEY, R. L., KIESSLER, P. C. (1987). *Traffic Processes in queueing Networks: A Markov Renewal Approach.* Baltimore: John Hopkins University Press.

DISNEY, R. L., KÖNIG, D. (1985). Queueing Networks: A survey of their random processes. *SIAM Review* 27, 335-403.

DISNEY, R. L., FARREL, R.L., DE MORAIS, P. R. (1973). A characterization of M/G/1 queues with renewal departures. *Management Science* 20,1222-1228.

GELEMBE, E., PUJOLLE, G. (1987). *Introduction to Queueing Networks.* Paris: Wiley.

GORDON, W. J., NEWELL, G. F. (1967). Closed queueing networks with exponential servers. *Oper. Res.* 15:254-265.

GROSS, D. & HARRIS, C. M. (1985). *Fundamentals of Queueing Theory.* 2a.ed. New York: John Wiley, (Wiley Series in Probability and Mathematical Statistics).

HUNTER, J. J. (1985). Filtering of Markov renewal queues IV: Flow-processes infeedback queues. *Adv. in Applied Probab.* 17:386-407.

JACKSON, J.R. (1957). Networks of waiting Lines. *Oper. Res.* 5:518-521.

JACKSON, J. R. (1963). Jobshop-like queueing systems. *Management Sci.* 10:131-142.

KELLY, F. P. (1979). *Reversibility and Stochastic Networks.* New York: John Wiley & Sons Ltda.

KENDALL, D. G. (1953). Stochastic processes occurring in the theory of queuesand their analysis by the method of imbedded Markov chains. *Ann. Math. Statist.* 24:338-354.

KLEINROCK, L. (1975). *Queueing Systems vol. 1: Theory.* New York: Wiley Interscience.

KLEINROCK, L. (1976). *Queueing Systems vol. 2: Computer applications.* NewYork: Wiley Interscience.

LITTLE, J. D. C. (1961). A proof for the queueing formula L=λW. *Oper. Res.* 9:383-387.

MELAMED, B. (1979). On Poisson traffic processes in discrete state Markovian systems with applications to queueing theory. *Advance in Applied Probability* 11:218-239.

MURDOCH, J. (1978). *Queueing Theory: Worked examples and Problems.* London: The MacMillan Press Ltda.

NOVAES, A. G. N. (1975). *Pesquisa Operacional e Transportes: Modelos Probabilisticos.* São Paulo: EDUSP e McGraw Hill.

RUDIN, W. (1976). *Principles of Mathematical Analysis.* 3a. ed. New York: McGraw-Hill.

WALRAND, J. (1988). *Introduction to Queueing Networks.*Englewood Cliffs, New Jersey: Prentice Hall.

WOLFF, R. W. (1988). *Stochastic Modelling and the Theory of Queues.* Englewoods Cliffs, New Jersey: Prentice Hall.

Índice Remissivo

Marcos Nascimento Magalhães

Professor do Departamento de Estatística do Instituto de Matemática e Estatística da Universidade de São Paulo. Licenciado e Mestre em Estatística pelo IME- USP. Doutor em Engenharia Industrial e Pesquisa Operacional pela Virginia Polytechnic Institute and State University, Virginia, EUA. Sua área de pesquisa é Teoria das Filas e Processos Estocásticos Aplicados.