Flávio Henrique Teles Vieira

Contribuições ao Cálculo de Banda e de Probabilidade de Perda para Tráfego Multifractal de Redes

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica. Área de concentração: Telecomunicações e Telemática.

Orientador: Lee Luan Ling

Campinas, SP
2006

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

V71c

Vieira, Flávio Henrique Teles Vieira
Contribuições ao cálculo de banda e de probabilidade de perda para tráfego multifractal de redes /Flávio Henrique Teles Vieira. – Campinas, SP: [s.n.], 2006.

Orientador: Lee Luan Ling.
Tese (doutorado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Telecomunicações - Tráfego. 2. Fractais. 3.Previsão. 4. Redes de computação. I. Lee, Luan Ling. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: Contributions to the effective bandwidth and loss probability computing for multifractal network traffic
Palavras-chave em Inglês: Multifractal modeling, Effective bandwidth, Loss probability, Network traffic, Admission Control, Quality of service, Network calculus
Área de concentração: Telecomunicações e Telemática
Titulação: Doutor em Engenharia Elétrica
Banca Examinadora: Dalton Soares Arantes, José Augusto Suruagy Monteiro, Nelson Luis Saldanha da Fonseca, Paulo Cardieri, Shusaburo Motoyama
Data da defesa: 19/12/2006

Flávio Henrique Teles Vieira

Contribuições ao Cálculo de Banda e de Probabilidade de Perda para Tráfego Multifractal de Redes

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica. Área de concentração: Telecomunicações e Telemática.

Banca Examinadora:

Prof. Dr. Dalton Soares Arantes - DECOM/FEEC/UNICAMP
Prof. Dr. José Augusto Suruagy Monteiro - UNIFACS
Prof. Dr. Lee Luan Ling - DECOM/FEEC/UNICAMP
Prof. Dr. Nelson Luis Saldanha da Fonseca - DSC/IC/UNICAMP
Prof. Dr. Paulo Cardieri - DECOM/FEEC/UNICAMP
Prof. Dr. Shusaburo Motoyama - DT/FEEC/UNICAMP

Campinas, SP
2006

Resumo

A modelagem multifractal generaliza os modelos de tráfego existentes na literatura e se mostra apropriada para descrever as características encontradas nos fluxos de tráfego das redes atuais. A presente tese investiga abordagens para alocação de banda, predição de tráfego e estimação de probabilidade de perda de bytes considerando as características multifractais de tráfego. Primeiramente, um Modelo Multifractal baseado em *Wavelets* (MMW) é proposto. Levando em consideração as propriedades deste modelo, são derivados o parâmetro de escala global, a função de autocorrelação e a banda efetiva para processos multifractais. A capacidade de atualização em tempo real do MMW aliada à banda efetiva proposta permite o desenvolvimento de um algoritmo de estimação adaptativa de banda efetiva. Através deste algoritmo é introduzido um esquema de provisão adaptativo de banda efetiva. Estuda-se também a alocação de banda baseada em predição de tráfego. Para este fim, propõe-se um preditor adaptativo *fuzzy* de tráfego, o qual é aplicado em uma nova estratégia de alocação de banda. O preditor *fuzzy* adaptativo proposto utiliza funções de base ortonormais baseadas nas propriedades do MMW. Com relação à probabilidade de perda para tráfego multifractal, deriva-se uma expressão analítica para a estimação da probabilidade de perda de bytes considerando que o tráfego obedece ao MMW. Além disso, uma caracterização mais completa do comportamento de fila é efetuada pela obtenção de limitantes para a probabilidade de perda e para a ocupação média do *buffer* em termos da banda efetiva do MMW. Por fim, é apresentado um esquema de controle de admissão usando o envelope efetivo proposto para o MMW oriundo do cálculo de rede estatístico, que garante que os fluxos admitidos obedeçam simultaneamente aos requisitos de perda e de retardo. As simulações realizadas evidenciam a relevância das propostas apresentadas.

Palavras-chave: Tráfego de Redes, Modelagem Multifractal, Predição, Banda Efetiva, Probabilidade de Perda, Controle de Admissão, Qualidade de Serviço.

Abstract

Multifractal modeling generalizes the existing traffic models and is believed to be appropriate to describe the characteristics of traffic flows of modern communication networks. This thesis investigates some novel approaches for bandwidth allocation, traffic prediction and byte loss probability estimation, by considering the multifractal characteristics of the network traffic. Firstly, a Wavelet based Multifractal Model (WMM) is proposed. Taking into account the properties of this multifractal model, we derive the global scaling parameter, the autocorrelation function and the effective bandwidth for multifractal processes. The real time updating capacity of the WMM in connection with our effective bandwidth proposal allows us to develop an algorithm for adaptive effective bandwidth estimation. Then, through this algorithm, an adaptive bandwidth provisioning scheme is introduced. In this work, we also study a prediction-based bandwidth allocation case. For this end, we develop an adaptive fuzzy predictor, which is incorporated into a novel bandwidth allocation scheme. The proposed adaptive fuzzy predictor makes use of orthonormal basis functions based on the properties of the WMM. Additionally, we derive an analytical expression for the byte loss probability estimation assuming that the traffic obeys the MMW. Besides, a more complete characterization of the queuing behavior is carried out through the estimation of the bounds for the loss probability and mean queue length in buffer in terms of the WMM based effective bandwidth. Finally, an admission control scheme is presented that uses the WMM based effective envelope derived through the statistical network calculus, guaranteeing that the admitted flows simultaneously attend the loss and delay requirements. The computer simulation results confirm the relevance of the presented proposals.

Keywords: Network Traffic, Multifractal Modeling, Traffic Prediction, Effective Bandwidth, Loss Probability, Admission Control, Quality of Service.

"O saber não está na ciência alheia que se absorve, mas, principalmente, nas idéias próprias, que se geram dos conhecimentos absorvidos, mediante a transmutação por que passam no espírito que os assimila"
Rui Barbosa

Agradecimentos

Gostaria de expressar meus agradecimentos

Aos meus pais, pela dedicação, carinho e suporte valiosos e por sempre apostarem em mim.

A todos de minha família, principalmente a minha irmã Renata, por me incentivarem a chegar até aqui.

Ao Prof. Lee Luan Ling, pelo apoio, incentivo, sempre acompanhando de forma prestativa o desenvolvimento deste trabalho.

Ao Prof. Dalton S. Arantes por ter apoiado o meu ingresso à Unicamp.

Ao Prof. Michel Daoud Yacoub pelo companheirismo nas aulas do programa de estágio docência.

Aos membros da banca pelas sugestões apresentadas.

Aos colegas e amigos do laboratório LRPRC e da FEEC Carlos, Christian, Cleisson, Fernando, Firmiano, Gabriel, Gilmar, Glauco, Hélcio, Helder, Juliano, Lígia, Lívio, Miguel, Paulo, Pepe, Talía, Ricardo, pelo auxílio e companheirismo.

A minha namorada Scheila, pelo carinho, apoio e compreensão durante esta jornada.

À CAPES e à Ericsson Telecomunicações SA, pelo apoio financeiro.

A Deus pela oportunidade confiada a mim.

Aos meus pais José Newton Vieira e Maria Helena Teles Vieira

Sumário

Lista de Figuras

Lista de Tabelas

Glossário

AGF- Average Goodness Factor

AQM- Active Queue Management

AR- Auto-Regressivo

ARFA- Agrupamento Regressivo Fuzzy Adaptativo

ATM - Asynchronous Transfer Mode

BB- Bandwidth Broker

CAB- Controle Adaptativo de Banda

CAC- Connection Admission Control

CBR- Constant Bit Rate

CWT - Continuous Wavelet Transform

DEC- Digital Equipment Corporation

Diffserv - Differentiated Services

DWT- Discrete Wavelet Transform

EDF- Earliest-Deadline-First

EEB- Empirical Effective Bandwidth

EQIM- Erro Quadrático Integral Médio

EQM- Erro Quadrático Médio

EQMN- Erro Quadrático Médio Normalizado

fBm - Fractional Brownian Motion

FBO- Funções de Base Ortonormais

FCFS- First-Come First-Served

FCM- Fuzzy C-Means

FCRM- Fuzzy Clustering Regression Model

FEC - Forwarding Equivalent Class

fGn- Fractional Gaussian Noise

GPS- Generalized Processor Sharing

Intserv - Integrated Services

IP - Internet Protocol

LAN- Local Area Network

LBL- Lawrence Berkeley Laboratory

LMS- Least Mean-Square

LRD- Long Range Dependence

LSP - Label Switching Paths

MAN - Metropolitan Area Network

mBm- Multifractal Brownian Motion

MLP- Multilayer Perceptron

MMW - Modelo Multifractal baseado em Wavelet

MPEG- Moving Picture Experts Group

MPLS - Multiprotocol Label Switching

MRA- Multiresolution Analysis

MWM - Multifractal Wavelet Model

NAR- Nonlinear AutoRegressive

NARMA- Nonlinear AutoRegressive Moving Average

NARX- Non-Linear Auto-Regressive with eXogeneuos input

NEB- Norros Effective Bandwidth

NLMA- Non-Linear Moving Average

PEL- Processo Envelope Linear

PEM- Processo Envelope Mínimo

QoS - Quality of Service

RLS- Recursive Least Squares

RSVP- Resource Reservation Protocol

RTRL- Real Time Recurrent Learning

SLA- Service Level Agreements

SON- Service Overlay Network

SP- Static Priority

TCP - Transmission Control Protocol

TEM- Taxa de Envelope Mínima

TSK- Takagi-Sugeno-Kang

UBR- Unspecified Bit Rate

VBR - Variable Bit Rate

VLL- Virtual Leased Lines

VPN - Virtual Private Network

VVGM - Variable Variance Gaussian Model

WAN - Wide Area Network

WRLS- Weighed Recursive Least Squares

WTMM- Wavelet Transform Modulus Maxima

Trabalhos Publicados e Submetidos

1. F. H. T. Vieira, R. P. Lemos e L. L. Lee. "Aplicação de Redes Neurais RBF Treinadas com Algoritmo ROLS e Análise Wavelet na Predição de Tráfego em Redes Ethernet ". VI Congresso Brasileiro de Redes Neurais, São Paulo, 2-05 de Junho, 2003;

2. F. H. T. Vieira, R. P. Lemos e L. L. Lee. "Alocação Dinâmica de Taxa de Transmissão em Redes de Pacotes Utilizando Redes Neurais Recorrentes Treinadas com Algoritmos em Tempo Real". Revista de Engenharia Elétrica do IEEE América Latina, Novembro, 2003;

3. G. R. Bianchi, F. H. T. Vieira e L. L. Ling. "Predictive Dynamic Bandwidth Allocation Based on Multifractal Traffic Characteristics". SAPIR (Service Assurance with Partial and Intermittent Resources)-ICT (International Conference on Telecomunications) 2004, Fortaleza,Ceará, Agosto, 2004;

4. F. H. T. Vieira e L. L. Lee. "Uma Nova Arquitetura Neural Combinada Utilizando Teoria de Ressonância Adaptativa e Aprendizagem RTRL para Predição e Controle de Tráfego de Dados em Tempo Real- Congresso Brasileiro de Automática (CBA), Gramado, RS, 21 a 24 de Setembro, 2004;

5. F. H. T. Vieira, G. R. Bianchi, L. L. Ling e R. P. Lemos. "Estimação de Banda Efetiva Dinâmica em Redes de Computadores Utilizando uma Modelagem Auto-Regressiva Nebulosa". XXI Simpósio Brasileiro de Telecomunicações (SBrT), Belém, Pará, Setembro, 2004;

6. G. R. Bianchi, F. H. T. Vieira e L. L. Ling. "Um Modelo Multifractal Aplicado à Predição de Tráfego de Redes". XXI Simpósio Brasileiro de Telecomunicações, Belém, Pará, Setembro, 2004;

7. F. H. T. Vieira e L. L. Lee. "Fuzzy Modeling and Prediction with Confidence Bound Estimation for Rate Allocation in High-Speed Networks". Simpósio Brasileiro de Redes Neurais- SBRN, São Luís, Maranhão, Brasil, 29 de Setembro a 1 de Outubro, 2004;

8. F. H. T. Vieira e L. L. Lee. "Uma Nova Arquitetura Neural Combinada Utilizando Teoria de Ressonância Adaptativa e Algoritmo de Kalman Estendido para Alocação Dinâmica de Taxa de Transmissão em Redes de Computadores". Simpósio Brasileiro de Redes Neurais- SBRN, São Luís, Maranhão, Brasil, 29 de Setembro a 1 de Outubro, 2004;

9. F. H. T. Vieira, G. R. Bianchi, L. L. Ling e R. P. Lemos. "Fuzzy-AR Modeling for Dynamic Effective Bandwidth Estimation in High-Speed Networks". WSEAS Transactions on Systems, Issue 8. Vol. 3, pp.2680-2685, Outubro, 2004;

10. G. R. Bianchi, F. H. T. Vieira e L. L. Ling. "A Novel Network Traffic Predictor based on Multifractal Characteristic". GLOBECOM'04, Dallas, Texas, USA, 29 de novembro a 3 de dezembro, 2004;

11. F. H. T. Vieira e L. L. Ling "Modelagem Multifractal utilizando Cascata Multiplicativa com Distribuição Generalizada de Multiplicadores", XXII Simpósio Brasileiro de Telecomunicações - SBrT'05, 04-08 de Setembro de 2005, Campinas, SP;

12. C. Jorge, F. H. T. Vieira e L. L. Ling "Escalonamento GPS baseado na Regularidade Local do Tráfego Internet"SBrT'05, 04-08 de Setembro de 2005, Campinas, SP;

13. C. Jorge; F. H. T. Vieira; L. L. Ling. "Predição Adaptativa do Expoente de Hölder para Tráfego Multifractal de Redes". XXVIII Congresso Nacional de Matemática Aplicada e Computacional, São Paulo, 12 a 15 de setembro, 2005;

14. F. H. T. Vieira e L. L. Ling. "Análise de Fila para Tráfego Multifractal utilizando Cálculo de Rede e Parâmetro de Escala Global". Simpósio Brasileiro de Redes de Computadores - SBRC, Curitiba-PR, 29 de Maio a 02 de Junho, 2006;

15. F. H. T. Vieira e L. L. Ling. "Multifractal Traffic Modeling using a Multiplicative Cascade with Generalized Multiplier Distributions"International Conference on Communications-ICC, Istanbul-Turquia, 12 a 15 de Junho, 2006;

16. F. H. T. Vieira e L. L. Ling. "Queueing Analysis for Multifractal Traffic through Network Calculus and Global Scaling Parameter", International Telecommunications Simposium - ITS, Fortaleza, CE, Setembro, 2006;

17. L. M. C. Sousa, F. H. T. Vieira e L. L. Ling. "A Fuzzy Approach for Adaptive Control of MPLS Network Traffic Flows", International Telecommunications Simposium - ITS, Fortaleza, CE, Setembro, 2006;

18. L. M. C. Sousa, F. H. T. Vieira e L. L. Ling. "Controle Fuzzy Adaptativo de Fluxos de Tráfego". Congresso Brasileiro de Automática. Salvador, Bahia, Brasil, 3 a 6 de Outubro, 2006;

19. F. H. T. Vieira e L. L. Ling. "Performance Bounds for a cascade based multifractal traffic model with generalized multiplier distribution". Revista da Sociedade Brasileira de Telecomunicações (SBrT) (submetido);

20. F. H. T. Vieira e L. L. Ling. "Controle de admissão com qualidade de serviço para tráfego multifractal utilizando envelope efetivo". Revista da Sociedade Brasileira de Telecomunicações (SBrT) (submetido);

21. G. R. Bianchi, F. H. T. Vieira e L. L. Ling. "Caracterização e Predição de Tráfego de Redes através do Modelo Tráfego Browniano Fracionário Estendido". Revista da Sociedade Brasileira de Telecomunicações (SBrT) (submetido);

22. C. Jorge, F. H. T. Vieira e L. L. Ling. "Esquema de Escalonamento Baseado Na Regularidade Local de Fluxos de Dados Internet". Revista da Sociedade Brasileira de Telecomunicações (SBrT) (submetido);

23. L. M. C. Sousa, F. H. T. Vieira, L. L. Lee. "Adaptive Fuzzy Modeling for Predictive Control of High-Speed Network Traffic". IWT International Workshop on Telecommunications, Santa Rita do Sapucaí - Minas Gerais, 12 a 17 de fevereiro, 2007.

Capítulo 1

Introdução

O conhecimento da banda requerida pelos fluxos de tráfego de forma a atender a parâmetros de qualidade de serviço (QoS) é de suma importância em redes de comunicações. Busca-se garantir o cumprimento desses parâmetros e ao mesmo tempo atingir alta utilização dos enlaces. Neste sentido, a "banda efetiva" se insere como uma solução adequada para se estimar a banda realmente necessária a ser fornecida aos fluxos das redes. O conceito de banda efetiva para fontes de tráfego é amplamente aceito como um dos parâmetros mais apropriados empregados em controle de admissão e alocação de recursos em redes de dados (Gibbens & Kelly, 1991)(Knightly & Shroff, 1999). Como exemplo, sejam enlaces VBR (*Variable Bit Rate*- Taxa de Bit Variável) que podem transportar fluxos de tráfego multiplexados a uma taxa menor do que a soma total das taxas de pico de todas as conexões envolvidas, assim como tolerar uma pequena perda ou atraso (Beran et al., 1995) (Pancha & Eizarki, 1994). Neste caso, a banda efetiva pode ser vista como um modo de caracterizar as exigências de recurso de conexões com taxas variáveis.

Atualmente, a busca por redes de serviços integrados que utilizam a tecnologia IP (Internet Protocol) tem sido crescente (Ryu, 2003). Porém, ainda hoje a maioria destas redes IP implementam somente serviços de "melhor-esforço", os quais são apropriados para certas aplicações de transmissão de dados, mas não para aplicações em tempo real ou sensíveis ao retardo. A fim de ampliar o suporte a serviços com restrições temporais e com taxas de transmissão variáveis, o conceito de qualidade de serviço (QoS) deve ser incorporado ou implementado nestas redes IP. A primeira arquitetura proposta foi o IntServ (*Integrated Service* - Serviços Integrados) que propõe a realização de reservas de recursos ao longo dos caminhos de transmissão da rede (Braden et al., 1994). Devido à falta de escalabilidade, este modelo de serviço com QoS não pode ser aplicado satisfatoriamente nos pontos da Internet onde o número de fluxos é alto.

Como alternativa, a arquitetura DiffServ (*Differentiated Services*- Serviços Diferenciados) foi desenvolvida para resolver a falta de escalabilidade pela agregação de fluxos de dados (Blake, 1998).

Em vez de se ter tratamento individualizado para cada fluxo de tráfego, os pacotes de dados são designados às classes que provêem parâmetros de QoS distintos. O Diffserv pode ser implementado através da arquitetura MPLS (Multiprotocol Label Switching). Esta arquitetura incorpora Engenharia de Tráfego criando rotas através do estabelecimento de LSPs (Label Switching Paths) e usando FECs (Forwarding Equivalent Class) que divide o tráfego em diferentes fluxos agregados de acordo com o modelo de serviço adotado. Neste sentido, novamente, a banda efetiva pode ser uma ferramenta extremamente útil na estimação dos requisitos destes fluxos de tráfego.

Vários métodos foram propostos na literatura para se determinar a banda efetiva do tráfego de redes baseada em modelos, tais como, fluxos de tráfego Markovianos, tráfego com curta-dependência e tráfego apresentando longa-dependência. Para representar tráfego monofractal com dependência a longo prazo, o movimento Browniano fracionário (fBm) é um modelo auto-similar popular cuja banda efetiva foi derivada em (Norros, 1994). A popularidade do fBm é conseqüência da natureza auto-similar (fractal) do tráfego Internet, demonstrada em (Park & Willinger, 2000). As bandas efetivas de modelos de tráfego, sendo ambos auto-similares e de cauda pesada com distribuições alfa-estáveis, foram estabelecidas em (Harmantzis et al., 2003). Entretanto, pesquisas apontam que o tráfego de redes pode ter propriedades estatísticas e comportamentos em escala mais complexos do que se consideram os modelos auto-similares (Erramilli et al., 1996). Estas constatações experimentais motivaram o surgimento de modelos que descrevem adequadamente estas características do tráfego, os modelos multifractais (Riedi & Véhel, 1997)(Feldmann et al., 1998). Neste contexto, cascatas multiplicativas foram inicialmente sugeridas para modelar processos de tráfego de rede de modo a capturar estas características multifractais encontradas em escalas de tempo pequenas (Park & Willinger, 2000).

Os modelos multifractais englobam várias propriedades de modelos anteriores e descrevem portanto, de forma mais precisa e abrangente, o comportamento do tráfego de redes. Com uma caracterização do tráfego mais detalhada, espera-se obter melhores estimativas para, por exemplo, a banda efetiva e a probabilidade de perda de pacotes para fluxos de tráfego. Este trabalho avalia várias questões relacionadas aos processos multifractais, principalmente aquelas envolvidas no cálculo de banda e de probabilidade de perda. Objetiva-se, de uma forma geral, desenvolver algoritmos e estratégias de alocação de banda e controle de admissão que considerem as características multifractais das séries de tráfego. Para isso, propõe-se uma gama de soluções incluindo um novo modelo multifractal, equações para o cálculo de banda efetiva e de probabilidade de perda, algoritmos de predição, esquema de controle de admissão, entre outras.

Dada a importância dos modelos multifractais, este trabalho propõe um Modelo Multifractal baseado em Wavelet (MMW) que oferece modelagem de tráfego em tempo real (*on-line*) através de poucos parâmetros de entrada. Como será comprovado, este modelo captura com eficiência as características observadas em séries de tráfego reais. O MMW é o modelo de tráfego central desta

tese, de onde são derivadas suas propriedades e parâmetros com a finalidade de se obter soluções para cálculo de banda efetiva, probabilidade de perda, predição e controle de admissão para tráfego multifractal.

Algoritmos de predição de tráfego são ferramentas importantes em controle e gerenciamento de tráfego. A predição dos valores de intensidade de tráfego permite entender melhor sobre a dinâmica do tráfego e este conhecimento pode ser usado na tomada de decisões em questões de controle de fluxo e alocação de banda. Neste trabalho, é feito um estudo sobre algoritmos adaptativos de predição de tráfego. Mais precisamente, incorpora-se algumas propriedades do MMW em um preditor *fuzzy* para que um melhor desempenho de predição seja obtido. Com essa finalidade, desenvolveu-se um algoritmo de treinamento adaptativo para o modelo *fuzzy* proposto que considera a expressão de função de autocorrelação obtida para o MMW. Além disso, o preditor *fuzzy* proposto é incorporado a um esquema adaptativo de alocação de banda baseado em predição.

Em seguida, esta tese inova ao se obter uma expressão analítica da banda efetiva para processos multifractais baseados em cascatas multiplicativas, o que inclui o MMW proposto. De acordo com nosso conhecimento, este é o primeiro trabalho que estabelece uma expressão analítica de banda efetiva para um modelo multifractal. O trabalho de Krishna et al., uma das poucas tentativas de se estabelecer banda efetiva para modelos multifractais, supõe que o tráfego de redes possui algumas estatísticas monofractais (fBm) (Krishna et al., 2001). Enquanto, os autores em (Melo & da Fonseca, 2005) apresentam um algoritmo para o cálculo de banda equivalente com base no modelo multifractal gaussiano mBm (*multifractional Brownian motion*).

Devido ao comportamento extremamente dinâmico dos fluxos de tráfego, a implementação de esquemas de estimação e de alocação adaptativa de banda tem se tornado imperativa. Uma questão importante em gerenciamento de rede consiste em se determinar a banda capaz de atender às exigências de QoS neste ambiente dinâmico. A partir das características de atualização em tempo real do modelo multifractal proposto (MMW), é desenvolvido também um algoritmo adaptativo de estimação de banda efetiva. Vantagens óbvias desta solução incluem que os dados de tráfego não precisam ser armazenados e a banda efetiva pode ser estimada em tempo real. Como conseqüência dos pontos positivos deste algoritmo proposto, acredita-se que este possua grande potencial de uso em aplicações reais, por exemplo, em reconfiguração adaptativa dos LSPs em redes MPLS, dimensionamento adaptativo dos enlaces em redes VPNs (Redes Privadas Virtuais), etc.

Um conceito intimamente ligado à banda efetiva é o de probabilidade de perda de pacotes, uma vez que a primeira é dada em função da segunda. A estimação da probabilidade de perda é freqüentemente considerada como o primeiro passo para se dimensionar o tamanho dos *buffers* nos roteadores a fim de garantir os requisitos de QoS. Dimensionamento dos *buffers* dos roteadores e controle de admissão são exemplos típicos de ações cujos resultados dependem fortemente de uma precisa carac-

terização do comportamento de fila dos dados de tráfego. Neste trabalho, obtém-se uma expressão analítica para a estimação de probabilidade de perda em um enlace com um servidor e *buffer* finito cujo tráfego de entrada é descrito pelo MMW. Ou seja, dados os parâmetros do modelo baseado em *wavelets* proposto (MMW), é possível estimar a probabilidade de perda sem a necessidade de efetuar simulações explícitas que requerem todo conjunto de dados de tráfego. A fim de prover uma análise completa com relação ao comportamento de fila de fluxos de tráfego multifractais, também foram obtidos limitantes para a probabilidade de perda e para o tamanho de fila nos *buffers*, utilizando parâmetros do MMW e alguns resultados do cálculo de rede.

Esta tese finda ao unir os conceitos do cálculo de rede estatístico, do cálculo de banda efetiva e da modelagem multifractal. Neste contexto, se obtém o envelope efetivo para o modelo MMW através de sua banda efetiva. Este envelope efetivo proporciona o estabelecimento de um novo esquema de controle de admissão que considera ambos parâmetros de QoS, probabilidade de perda e de retardo.

A tese está organizada da seguinte maneira:

Capítulo 2: O Capítulo 2 expõe os conceitos relacionados à modelagem multifractal. Inicia-se com a descrição de processos auto-similares e de suas propriedades como a longa-dependência. Em seguida, o conceito de processo multifractal é definido. São exibidas ferramentas da análise multifractal, como por exemplo o espectro multifractal e o expoente de Hölder, que quantificam a característica irregular dos processos multifractais e seus correspondentes métodos de estimação. Ao final do capítulo são apresentados os principais modelos estatísticos multifractais existentes na literatura.

Capítulo 3: Este capítulo propõe o modelo Multifractal baseado em Wavelets (MMW) e seu procedimento de síntese de tráfego. É apresentada uma variedade de testes estatísticos e de desempenho de fila para validar o modelo de tráfego proposto. Neste capítulo, são derivados o parâmetro de escala global e a função de autocorrelação para este modelo multifractal. Além disso, desenvolve-se um algoritmo de estimação adaptativa dos parâmetros de entrada para o MMW.

Capítulo 4: Este capítulo trata da predição de séries temporais de tráfego. Inicialmente, descreve-se a predição linear de séries temporais representada pelos algoritmos de Wiener, LMS e RLS. Quanto à predição não-linear, ênfase é dada à modelagem *fuzzy*, principalmente para o modelo TSK. Alguns resultados de predição de tráfego usando este modelo são apresentados.

Capítulo 5: O Capítulo 5 é dedicado ao desenvolvimento de um preditor *fuzzy* adaptativo no qual são inseridas informações baseadas no MMW e que são estimadas adaptativamente. É verificado seu desempenho de predição em comparação a outros preditores adaptativos. Ao final do capítulo, este preditor, dado o seu excelente desempenho, é inserido em uma proposta de estimação adaptativa de banda.

Capítulo 6: Este capítulo aborda o tema banda efetiva, onde se discute brevemente a Teoria dos

Grandes Desvios, base do conceito de banda efetiva. Em seguida, propõe-se uma expressão analítica para a estimação de banda efetiva baseado nos multiplicadores do MMW. Esta expressão matemática e a característica de atualização em tempo real do MMW são usados para construir um algoritmo adaptativo de estimação de banda efetiva para tráfego multifractal. Este mesmo algoritmo é então incorporado a um esquema adaptativo de provisão de banda. Resultados experimentais das propostas são apresentados por todo capítulo.

Capítulo 7: Neste capítulo, discute-se a caracterização do comportamento de fila do tráfego, principalmente em termos de probabilidade de perda. De forma inovadora, é obtida uma expressão para a probabilidade de perda para tráfego multifractal em um enlace baseada na teoria de muitas fontes ('Many Sources') e que usa os parâmetros e propriedades do MMW. O capítulo também estabelece limitantes para a probabilidade de perda e o tamanho médio da fila para processos descritos pelo MMW através da associação entre cálculo de rede e banda efetiva.

Capítulo 8: Utilizando-se do Cálculo de Rede Estatístico, o Capítulo 8 introduz o envelope efetivo para processos multifractais através da banda efetiva do MMW. Este envelope efetivo permite que seja desenvolvido um esquema de controle de admissão para fluxos de tráfego, onde são garantidos os requisitos probabilísticos de perda e atraso desses fluxos. Um estudo é realizado sobre o número de fluxos que podem ser admitidos em comparação a outras abordagens e para diferentes tipos de escalonadores de tráfego.

Capítulo 9: Finalmente, no Capítulo 9 são apresentadas as conclusões obtidas e possíveis extensões a este trabalho.

1.1 Contribuições do autor

As principais inovações e contribuições dadas pelo presente trabalho são:

1. Novo modelo multifractal de tráfego baseado em análise *wavelet* (MMW);
2. Parâmetro de escala global para o MMW, ou seja, para processos multifractais;
3. Função de autocorrelação do MMW;
4. Algoritmo adaptativo de estimação de parâmetros do MMW a partir de funções multifractais;
5. Modelo preditor *fuzzy* com funções de base ortonormais baseada na função de autocorrelação do MMW;
6. Algoritmo de treinamento adaptativo para o modelo *fuzzy* proposto;
7. Esquema adaptativo de alocação de banda baseada no preditor *fuzzy* proposto;

8. Banda efetiva para o MMW;
9. Estimação adaptativa de banda efetiva para tráfego multifractal;
10. Expressão analítica para a estimação da probabilidade de perda de *bytes* baseada no MMW;
11. Limitantes de desempenho para a probabilidade de perda de pacote e o tamanho de fila;
12. Esquema de controle de admissão utilizando envelope efetivo para processos multifractais.

Capítulo 2

Modelagem Multifractal de Tráfego

2.1 Introdução

O modelo clássico de tráfego de Poisson foi proposto na década de 20 para análise de sistemas de telefonia e depois adaptado para a análise de fila em redes de pacotes (Kleinrock, 1975). Este modelo tem como vantagem o fato de ser analiticamente tratável, permitindo que expressões matemáticas sejam explicitamente derivadas para caracterização do desempenho de fila nos *buffer*s. Mas com o advento das redes com suporte a vários serviços, a natureza do tráfego mudou drasticamente, exibindo características bem diferentes das preditas por modelos de Poisson e de Markov, como por exemplo, comportamento fractal.

As pesquisas sobre tráfego de redes encontraram-se com a teoria dos fractais a partir da publicação do trabalho de Leland, Taqqu, Willinger e Wilson (Leland et al., 1994). Leland et al. constataram experimentalmente que o tráfego coletado na rede Ethernet do *Bellcore Morristown Reaserch and Engineering Center* exibia propriedades fractais tais como auto-similaridade. A característica de dependência de longa duração ou longa-dependência para uma série de tráfego implica em uma estrutura de correlação que decai mais lentamente do que a exponencial. Este decaimento lento da função de autocorrelação, relacionado também à auto-similaridade, foi observado no tráfego gerado por transmissão de vídeo a taxa variável (Beran et al., 1995)(Garret & Willinger, 1994)(Fitzek & Reisslein, 2001), no tráfego em redes de longa distância (WAN-*Wide Area Networks*) e metropolitanas (MAN-Metropolitan Area Network) (Addie et al., 1995)(Paxson & Floyd, 1995), tráfego Internet (Crovella & Bestavros, 1996), dentre outros. Foi constatado que tais propriedades, com destaque para a dependência de longa duração, influenciam fortemente no desempenho e projeto de redes (Park & Willinger, 2000), não sendo adequadamente modeladas por processos estocásticos Markovianos.

Enquanto os modelos de tráfego de curta dependência são significativos por sua simplicidade, eles não capturam a dependência de longa duração presente nos traços reais de tráfego. Há muitos estudos

que revelam a alta variabilidade do tráfego Internet, ou seja, o tráfego possui rajadas em uma gama de escalas de tempo em contraste da suposição de que rajadas de tráfego só existem em escalas curtas de tempo (Leland et al., 1994)(Paxson & Floyd, 1995). Foi mostrado que estas incidências de rajadas multiescalas têm um impacto significativo no desempenho das redes (Leland et al., 1994)(Paxson & Floyd, 1995)(Erramilli et al., 1996). Os resultados publicados em (Erramilli et al., 1996) mostram que o conhecimento das características do tráfego em múltiplas escalas ajuda a melhorar a eficiência dos mecanismos de controle de tráfego.

Diferentes modelos de tráfego foram propostos na literatura com o objetivo de representar a característica auto-similar constatada no tráfego de redes. Identificado como o modelo que de maneira mais simples incorpora matematicamente as características auto-similares observadas no tráfego, o modelo tráfego Browniano fracionário (*fractional Brownian traffic*, fBt) tornou-se amplamente utilizado. Entretanto, observou-se que enquanto em escalas de tempo da ordem de centenas de milissegundos e maiores, o comportamento do tráfego era bem representado por modelos auto-similares, em escalas de tempo da ordem de centenas de milissegundos e menores, as características de tais modelos afastavam-se das apresentadas pelo tráfego real.

Outras propriedades do tráfego de redes foram também descobertas apontando para um comportamento do tráfego mais complexo do que se supunham os modelos monofractais. Os resultados em (Erramilli et al., 1996) indicam que o desempenho de fila depende mais da variabilidade do tráfego em certas escalas de tempo do que do valor do parâmetro de Hurst H, que mede o grau de auto-similaridade de um processo. Realmente, verificou-se que diferentes processos com longa dependência possuindo o mesmo valor do parâmetro de Hurst podem gerar comportamentos de fila sensivelmente diferentes (Grossglauser & Bolot, 1999). A origem dessas propriedades, consideradas presentes em pequenas escalas de tempo, é atribuída à ação dos protocolos predominantes nas redes em questão, e dos mecanismos fim-a-fim de controle de congestionamento existentes na Internet atual, que determinam o comportamento do fluxo de informações entre diferentes camadas na hierarquia de protocolos TCP/IP (Feldmann et al., 1998). Tais constatações motivaram a busca por modelos de tráfego mais abrangentes, que possibilitassem uma descrição mais completa do tráfego de redes, como por exemplo, modelos multifractais baseados em cascatas multiplicativas.

Investigações envolvendo tráfego WAN TCP/IP (Riedi & Véhel, 1997)(Feldmann et al., 1998) constataram que essas diferentes propriedades e comportamentos do tráfego eram convenientemente descritos utilizando-se a análise multifractal. Para muitos processos de tráfego de rede, seus gráficos de energia em escala dos coeficientes *wavelet* ou os de variância-tempo normalmente não têm comportamento linear. Muitos destes processos apresentam uma combinação de comportamentos fractais, com parâmetro de Hurst variado em diferentes pequenas escalas de tempo, ou seja, são multifractais (Li & Mills, 1999). De fato, o desempenho de fila depende intensamente das irregularidades do

tráfego em escalas de tempo pequenas devido à dinâmica complexa das redes de dados (Grossglauser & Bolot, 1999). Feldmann, et.al discutem que esta combinação de comportamentos em escala é melhor representada por um processo multifractal (Feldmann et al., 1998). Em (Erramilli et al., 2000), Erramilli et al. confirmaram esta hipótese, e indicaram que este comportamento pode ter um impacto significativo no desempenho de filas. Em (Véhel & Riedi, 1997), Lévy Véhel e Riedi mostraram que uma versão multifractal do movimento Browniano fracionário (fBm) pode refletir melhor as propriedades do tráfego de rede. Em (Riedi et al., 1999), os autores apresentaram o modelo MWM (Multifractal Wavelet Model) mostrando que este modelo prediz melhor o comportamento de traços de tráfego TCP (Paxson & Floyd, 1995) e de Ethernet (Leland et al., 1994) quando comparado com o movimento Browniano fracionário (fBm). Em (Gao & Rubin, 1999a), Gao e Rubin mostraram que ambos processos de tempo entre chegada de pacotes e tamanho de pacotes, podem ser considerados multifractais. Em (Gao & Rubin, 2000), os mesmos autores examinam as propriedades de fluxos de tráfego multifractais em superposição. Em (Krishna et al., 2001), Krishna et al. propuseram um modelo baseado em cascata multiplicativa em que se assumia as distribuições dos multiplicadores da cascata como sendo gaussianas (VVGM-Variable Variance Gaussian Model) para modelar processos de tempo entre chegada de pacotes.

No presente capítulo, são apresentadas as definições formais de processos monofractais e multifractais, seus parâmetros descritores, propriedades do tráfego e métodos da análise multifractal como estimação da função de partição, do espectro multifractal e da regularidade local. O capítulo termina com uma breve descrição de três modelos multifractais de destaque: o movimento Browniano multifracionário (mBm), cascatas multiplicativas e o MWM (Multifractal Wavelet Model).

2.2 Processos Fractais

Algumas características do tráfego real, como correlação de longo prazo e rajadas em diferentes escalas de tempo, podem ser modeladas usando fractais. Mandelbrot descreve fractal como uma entidade caracterizada por irregularidades que governam sua forma e complexidade, possuindo uma estrutura com detalhes em todos os níveis de resolução (Mandelbrot, 1977). A definição de fractais não é trivial e depende de outra definição, a da dimensão de conjunto adotada. Qualquer conjunto matemático possui um número característico a ele associado chamado de dimensão topológica. Intuitivamente a dimensão topológica de um conjunto é igual ao número de parâmetros necessários para descrever todos os pontos deste conjunto, por exemplo, tanto em coordenadas polares quanto retangulares, são necessários dois parâmetros para descrever os pontos de um circulo, portanto trata-se de um conjunto bi-dimensional. Não há uma definição rápida e precisa de fractal, mas sim um conjunto de propriedades características dos fractais. Os fractais exibem propriedades tais como

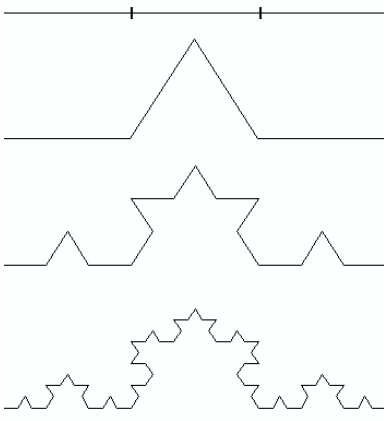


Fig. 2.1: Três iterações da subdivisão de Von Koch. A curva de Von Koch é um fractal obtido no limite de um número infinito de subdivisões.

auto-similaridade, espectro segundo uma lei de potência, estrutura irregular e dimensão de Hausdorff maior do que sua dimensão topológica.

O conjunto triádico de Cantor e a curva de Von Koch são exemplos típicos de conjuntos fractais (Falconer, 1990). A curva de Von Koch possui um comprimento infinito em um quadrado finito em $\mathbb{R}^2$ (Figura 2.1). Causado por sua intrínseca irregularidade, o tamanho de objetos fractais não é mensurável em termos da geometria euclidiana clássica. Ao contrário, seu tamanho pode ser caracterizado definindo-se uma dimensão não-inteira (dimensão fractal) (Peitgen et al., 1994). A medida usual de comprimento é portanto mal adaptada para caracterizar as propriedades topológicas de tais curvas fractais. Este fato motivou Hausdorff, em 1919, a introduzir uma nova definição de dimensão, baseada no tamanho das variações dos conjuntos quando medidos em diferentes escalas, a qual definiremos a seguir.

Seja U um subconjunto não-vazio do espaço euclidiano n-dimensional $\mathbb{R}^n$, sendo o diâmetro de U dado por $|U| = \sup\{|x-y| : x, y \in U\}$. Dado $\{ U_i \}$, uma coleção finita de conjuntos com diâmetro de valor máximo δ, que cobre um subconjunto *F* em $\mathbb{R}^n$, tal que $F \subset \bigcup_{i=1}^{\infty} U_i$ com $0 < |U_i| \leq \delta$ para todo i. Neste caso, $\{U_i\}$ é denominado de coleção de conjuntos de cobertura δ de *F*. Assumindo que F seja um subconjunto de $\mathbb{R}^n$ e s seja um número não-negativo, a definição de medida de Hausdorff s-dimensional é dada como:

Definição 2.2.1 *Seja* $\mid U_i \mid$ *o diâmetro do conjunto* U_i*. Define-se* $H^s_\delta(F)$ *para algum* $\delta > 0$ *como*

$$H^s_\delta(F) = \inf \left\{ \sum_{i=1}^{\infty} \mid U_i \mid^{\mathrm{s}} : \ \{U_i\} \quad \textit{é uma coleção de conjuntos de cobertura } \delta \text{ de } F \right\} \qquad (2.1)$$

A medida de Hausdorff s*-dimensional de F é definida como o limite de* $H^s_\delta(F)$ *quando* δ *tende a zero.*

Ou seja,

$$H^s(F) = \lim_{\delta \to 0} H^s_\delta(F). \tag{2.2}$$

Pode-se demonstrar que o limite que define a medida de Hausdorff existe para qualquer subconjunto *F* em $\mathbb{R}^n$, sendo seu valor usualmente igual a 0 ou ∞ (Mandelbrot, 1977).

Definição 2.2.2 *O valor crítico de s para o qual a medida de Hausdorff s-dimensional $H^s(F)$ muda instantaneamente de ∞ para 0 é definido como dimensão de Hausdorff de F.*

Como já foi mencionado, um fractal é caracterizado por possuir dimensão de Hausdorff maior do que sua dimensão topológica. Por isso, esta dimensão é importante e como será mostrado, ela se insere também na formulação do espectro multifractal.

Ao se constatar a presença de propriedades fractais no tráfego de redes, os modelos de tráfego apresentando tais propriedades, como a auto-similaridade, tiveram grande impacto na área de comunicações. Impacto este, devido também ao fato de que as propriedades fractais do tráfego influenciam fortemente o desempenho das redes e até então não eram descritas pelos modelos existentes (Leland et al., 1994).

2.2.1 Processos Auto-similares

Muitos pesquisadores observaram que o modelo de Poisson, que era usado na análise de filas, nem sempre era capaz de modelar adequadamente o tráfego de rede (Paxson & Floyd, 1995). Na escolha de um modelo de tráfego apropriado, vários fatores são decisivos, tais como: generalidade, facilidade de implementação, precisão, enfim, características próximas às das fontes reais.

Tráfego com rajadas em várias escalas temporais pode ser descrito usando o conceito de auto-similaridade. A auto-similaridade é a propriedade associada aos fractais, objetos cuja aparência não muda apesar da mudança de escala conforme mostrado pela Figura 2.2. O parâmetro de Hurst mede o grau de auto-similaridade do processo e representa basicamente uma medida do decaimento da função de autocorrelação do processo.

Modelos de tráfego com longa dependência se baseiam principalmente em processos auto-similares. O termo auto-similaridade se refere normalmente a processos assintoticamente auto-similares de segunda ordem ou monofractais (Park & Willinger, 2000). No contexto de processos a tempo discreto, a auto-similaridade é definida em termos de seus processos agregados. Neste sentido, seja $Y(t)$ um processo auto-similar com incrementos estacionários $X(t)$ no tempo discreto $t \in \mathbb{Z}^+$. A partir do processo de incrementos a tempo discreto, $X(t), t \in \mathbb{Z}^+$, pode-se obter o processo agregado $X^m(t)$ de $X(t)$, definido como a média amostral do processo $X(t)$ em blocos não sobrepostos de tamanho

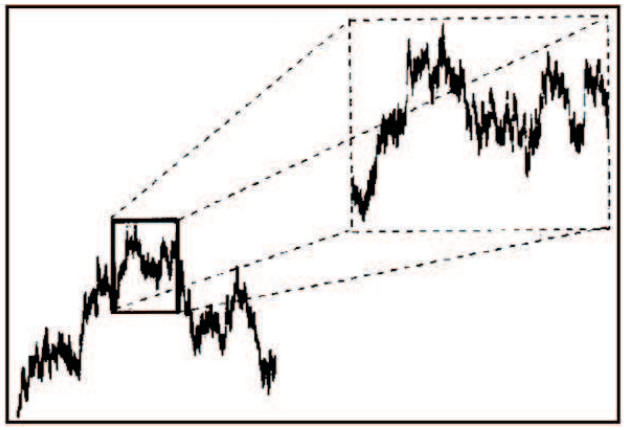


Fig. 2.2: Auto-similaridade estatística. Uma parte dilatada da série não pode ser estatisticamente distinguida de toda série.

m, ou seja,

$$X^m(t) = \frac{1}{m} \sum_{i=m(t-1)+1}^{mt} X(i), \qquad t = 1, 2, 3, \ldots \tag{2.3}$$

Dado que o processo $X(t)$ é estacionário e $E[Y(t)] = 0$, então para todo número inteiro m tem-se que

$$X^m(t) \stackrel{d}{=} m^{H-1} X. \tag{2.4}$$

onde $\stackrel{d}{=}$ denota igualdade em distribuição.

No caso de séries temporais auto-similares, quando vistas em diferentes escalas temporais sua estrutura relacional permanece inalterada. Como resultado, essas séries, definidas a seguir, possuem rajadas em uma gama de escalas temporais.

Definição 2.2.3 *Seja um processo estocástico $X(t)$ no tempo discreto $t \in \mathbb{Z}$. Este processo é dito auto-similar com parâmetro de Hurst $H \in (0; 1)$ se, para todo $m > 0$ e $t \geq 0$, os processo $X(t)$ e $m^{1-H} X^m(t)$ são identicamente distribuídos, ou seja,*

$$X(t) \stackrel{d}{=} m^{1-H} X^m(t). \tag{2.5}$$

Se a equação (2.5) é válida apenas para grandes valores de m, i.e., escalas de agregação maiores, se diz que o processo $X(t)$ é assintoticamente auto-similar. Em muitas aplicações, apenas algumas estatísticas do tráfego são consideradas. Normalmente, o tráfego é caracterizado em termos de estatísticas de segunda ordem, de modo que é suficiente considerar que a auto-similaridade se manifeste nestas estatísticas. Isto permite relaxar a definição anterior e introduzir a seguinte definição:

Definição 2.2.4 *Seja $R_X(k) = E[X(t)X(t+k)]$ a função de autocorrelação do processo estacionário discreto $X(t), t \in \mathbb{Z}^+$, com parâmetro de Hurst (1/2 < H < 1). Diz-se que:*

1) $X(t)$ é exatamente auto-similar de $2^{\underline{a}}$ ordem se

$$R_X(k) = \frac{1}{2}((k+1)^{2H} - 2k^{2H} + (k-1)^{2h}), m \geq 1; \tag{2.6}$$

2) $X(t)$ é assintoticamente auto-similar de $2^{\underline{a}}$ ordem se

$$\lim_{m\to\infty} R_{X^m}(k) = R_X(k) = \frac{1}{2}((k+1)^{2H} - 2k^{2H} + (k-1)^{2h}). \tag{2.7}$$

Um conceito relacionado à auto-similaridade é o de distribuição de cauda pesada (*heavy tail*) (Park & Willinger, 2000) . A distribuição de cauda pesada da duração ou tamanho das sessões ou conexões que originam tráfego agregado é apontada como causa da característica auto-similar observada (Park & Willinger, 2000) (Crovella & Bestavros, 1996). Rajadas ocasionais presentes nos processos de tráfego de rede como o tempo entre chegada de pacotes, geram este comportamento de 'cauda pesada', onde a densidade de probabilidade desses processos decai lentamente. Mais precisamente, um processo X possui distribuição de cauda pesada se $P[X > \mathrm{x}] \approx x^{-a}, x \to \infty$ onde $0 < a < 2$ e a é denominado de índice de cauda. Este processo possui variância infinita para $0 < a < 2$, e se $a < 1$, a média de X também será infinita. Um exemplo de distribuição de cauda pesada é a distribuição de Pareto, dada por

$$F(x) = P[X \leq x] = 1 - \left(\frac{k}{x}\right)^a. \tag{2.8}$$

O parâmetro a que representa o grau de cauda pesada de um processo pode ser estimado pelo método de Hill (Hill, 1975) ou analisando o comportamento em escala dos dados do processo (Crovella & Taqqu, 1999) (Samorodnistiky & Taqqu, 1994). Com relação à origem da auto-similaridade, o trabalho de K. Park et al. mostra que a transferência de arquivos cujos tamanhos são descritos por distribuições de cauda pesada é suficiente para gerar auto-similaridade no tráfego (Park & Willinger, 2000). O índice de cauda a da distribuição de cauda pesada do tamanho dos arquivos pode diretamente determinar o grau de auto-similaridade no tráfego. Park verificou uma relação linear entre o parâmetro de Hurst H e o expoente a da distribuição do tamanho dos arquivos (Park & Willinger, 2000).

A auto-similaridade introduz dificuldades na otimização do desempenho da rede e na garantia de qualidade de serviço, podendo causar impactos significativos como aumento dos retardos e da taxa de perda de pacotes (Park & Willinger, 2000)(Tuan & Park, 1998). Quanto mais o tráfego apresenta características de cauda pesada mais são degradados parâmetros como perda de pacotes e atraso. Como há uma relação entre perda de pacotes e atraso, se diminui um, aumenta o outro, se torna

mais difícil de se garantir os requisitos de QoS para tráfego com rajadas. Por outro lado, pode haver períodos de baixo tráfego, quando se tem subutilização dos recursos da rede.

Introduzido por Mandelbrot, o movimento Browniano fracionário (fBm) é capaz de descrever o comportamento auto-similar do tráfego, sendo de grande interesse também em outras áreas, tais como, hidrologia, processamento de sinais e matemática financeira (Mandelbrot & Ness, 1968). O movimento Browniano fracionário é definido a partir do movimento Browniano, conforme as seguinte definições :

Definição 2.2.5 *O movimento Browniano é um processo aleatório* $\{B(t), t \in [0, \infty)\}$ *tal que:*

1) Para cada $t > 0$ *e* $u > 0$*, os incrementos* $B(t+u) - B(t)$ *possuem distribuição normal com média zero e variância* u*.*

2) O processo $B(t)$ *possui incrementos estacionários e independentes;*

3) $B(t)$ *é uma função contínua no tempo e* $B(0) = 0$*.*

Definição 2.2.6 - *Seja* ***B****(t) o movimento Browniano puro. O movimento Browniano fracionário (fBm-* fractional Brownian motion*) de expoente H* $\in$ *(0;1) é definido como:*

$$Z(t) = \frac{1}{\Gamma(H+1/2)} \left\{ \int_{-\infty}^{0} \left[(t-s)^{H-1/2} - (-s)^{H-1/2} \right] dB(s) + \int_{0}^{t} (t-s)^{H-1/2} \, dB(s) \right\} \quad (2.9)$$

com $Z(0) = 0$ *e* Γ *é a função gamma.*

O movimento Browniano fracionário possui as seguintes propriedades:

1. $Z(t)$ é Gaussiano;
2. $Z(t)$ é um processo contínuo;
3. $Z(t)$ possui incrementos estacionários;
4. $E[Z(t)] = 0$;
5. $E[Z(t)^2] = \sigma^2 t^{2H}$.

A partir das propriedades acima, a autocorrelação do movimento Browniano fracionário (fBm) $Z(t)$ é dada por

$$R_Z(s,t) = \frac{\sigma^2}{2} \left(|s|^{2H} + |t|^{2H} - |s-t|^{2H} \right), \quad (2.10)$$

ou seja, trata-se de um processo não-estacionário. Se $\sigma^2 = 1$, $Z(t)$ é chamado de movimento Browniano fracionário padrão. Para $H = 0,5$ e $\sigma^2 = 1$, $Z(t)$ é simplesmente o movimento Browniano padrão, com incrementos independentes e estacionários.

Dado um processo fBm $Z(t)$, pode-se definir um processo de incrementos em tempo discreto $X = \{X_n\}_{n=0}^{\infty}$ onde

$$X_n = Z(n+1) - Z(n). \tag{2.11}$$

O processo $X = \{X_n\}_{n=0}^{\infty}$ definido anteriormente é denominado ruído Gaussiano fracionário (*fractional Gaussian noise*, fGn). A definição formal do fGn é dada a seguir.

Definição 2.2.7 - *O processo $X = \{X_n\}_{n=0}^{\infty}$, é denominado ruído Gaussiano fracionário (fractional Guassian noise, fGn) com parâmetro H, se X é o processo de incrementos do fBm com parâmetro H. O processo X é estacionário, Gaussiano, com média $E[X_n] = 0$, variância $E[X_n^2] = \sigma^2 > 0$, e sua função de autocorrelação é dada por*

$$R_X(k) = \frac{\sigma^2}{2}\left(|k+1|^{2H} + |k|^{2H} - |k-1|^{2H}\right). \tag{2.12}$$

Para $H = 1/2$, o fGn reduz-se ao ruído Gaussiano branco, portanto, X torna-se completamente descorrelacionado. Caso $\sigma^2 = 1$, o processo X é denominado ruído Gaussiano fracionário padrão.

Por suas características, no contexto da modelagem de tráfego, o fBm é utilizado para descrever o volume acumulado de tráfego até um dado instante. Já o seu processo de incrementos, o fGn, é utilizado para descrever o volume de tráfego que atravessa um determinado ponto da rede em intervalos regulares de tempo.

2.2.2 Dependência de Longa Duração

A longa dependência ou dependência de longa duração de um processo é usualmente definida em termos de sua função de autocorrelação ou sua densidade espectral de potência. A longa dependência está relacionada a processos estacionários em sentido amplo com variância finita, enquanto processos auto-similares são geralmente não-estacionários. Entretanto, pode-se encontrar uma definição de longa-dependência para processos com variância infinita em (Brockwell & R.Davies, 1991). Assim como, uma classe de processos não estacionários com longa dependência também é definida em (Roughan & Veitch, 1999). Convém ressaltar que no caso de processos assintoticamente auto-similares de $2^{\underline{a}}$ ordem restringindo-se o parâmetro de Hurst ao intervalo $(1/2 < H < 1)$, auto-similaridade implica em longa dependência e vice-versa. Assim, diz-se que processos auto-similares com $H > 1/2$ apresentam dependência de longa duração; isto significa que a função de autocorrelação do processo apresenta decaimento hiperbólico e portanto não é somável (i.e $\sum_{k=-\infty}^{\infty} R(k) \to \infty$).

Para $H = 1/2$, nota-se que a função de autocorrelação de um processo auto-similar de segunda ordem é equivalente ao ruído branco, portanto somável (i.e $\sum_{k=-\infty}^{\infty} R(k) < \infty$). Neste caso, diz-se que o processo apresenta dependência de curta duração. Pode-se demonstrar que processos cuja distribuição é de cauda pesada também exibem longa dependência (Cappe et al., 2002). A seguinte definição expressa a longa dependência em termos da função de autocorrelação e da densidade espectral do processo em questão.

Definição 2.2.8 *Um processo estacionário $X(t)$ com média zero e variância finita apresenta longa dependência se sua função de autocorrelação $r(k) = E[X(t+k)X(t)]$ ou sua densidade espectral de potência $S_X(f)$ (transformada de Fourier da função de autocorrelação) satisfaz (Veitch & Abry, 1999):*

$$r(k) \sim c_r k^{-\beta}, \qquad k \to \infty \tag{2.13}$$

$$S_X(f) \sim c_f |f|^{-\gamma}, \qquad f \to 0 \tag{2.14}$$

onde $0 < \gamma < 1$, $0 < \beta < 1$, $\beta = 1 - \gamma$, e c_f e c_r são constantes não-nulas.

Note que o conceito de auto-similaridade realmente está ligado ao de longa dependência uma vez que o parâmetro de Hurst H se relaciona com o expoente γ por $H = 1 - \gamma/2$.

A dependência de longa duração é relevante para a inferência estatística, fazendo com que a estimação de parâmetros estatísticos tais como média, desvio padrão e intervalo de confiança seja mais difícil (Beran, 1995). Entretanto, a queda lenta das autocorrelações pode agir em benefício da qualidade da predição de valores futuros, uma vez que quanto maior a dependência entre a observação futura e as passadas, melhor será a predição das amostras futuras, desde que esta dependência seja explorada apropriadamente.

2.2.3 Múltiplos Comportamentos em Escala

Análises do tráfego TCP/IP revelam que este possui diferentes comportamentos em escala (*multiple scaling*) (Riedi & Véhel, 1997). Para entender esse fenômeno, vamos recorrer à Análise *Wavelet*. Seja $X(t)$ um processo auto-similar com parâmetro de Hurst $H \in (0.5, 1)$, então a esperança matemática da energia E_j do processo na escala j em uma banda 2^{-j} e em torno da freqüência $2^{-j}\omega$ pode ser expressa como (ver Apêndice A):

$$E[E_j] = E\left(\frac{1}{n_j}\sum_k |d_x(j,k)|^2\right) = c|2^{-j}\omega|^{1-2H}, \tag{2.15}$$

onde c é um fator que não depende de j e n_j denota o número de coeficientes *wavelets* $d_x(j,k)$ na escala j.

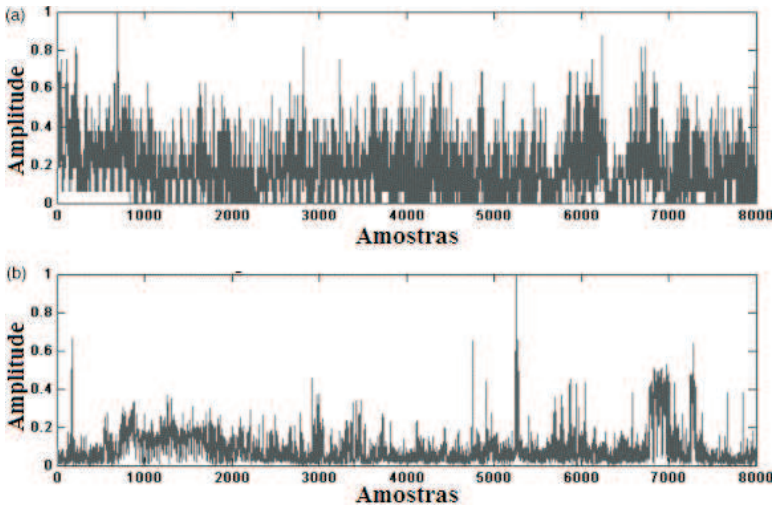


Fig. 2.3: Tráfego normalizado (a) Ethernet (monofractal); (b) Internet (multifractal).

A estimativa de E_j revela o comportamento do momento de segunda ordem do processo $X(t)$ em cada escala j, servindo como um estimador não-paramétrico da variância dos coeficientes *wavelet*. De fato, como o processo correspondente aos coeficientes *wavelet* de $X(t)$ possui curta-dependência (dependência de curto prazo), implica que E_j é um estimador aproximadamente ótimo do comportamento de segunda ordem de $X(t)$ na escala j (Veitch & Abry, 1999). Na seção 2.4 estendemos essa definição de energia dos coeficientes *wavelet* com o uso da função de partição *wavelet*. Por enquanto, saibamos que para um processo exatamente auto-similar a energia dos coeficientes *wavelet* em várias escalas tende a uma reta em um escala logarítmica (comportamento em escala-*scaling*). Isso pode ser constatado para o processo monofractal fBm. Entretanto, para tráfego real esses gráficos de energia em escala muitas vezes desviam da linearidade, possuindo distintas regiões com comportamento aproximadamente linear, o que torna evidente que há múltiplos domínios de escala presente nos dados. Neste contexto, se inserem os modelos multifractais, os quais permitem caracterizar esse comportamento.

Os modelos monofractais por possuírem um único valor para o expoente de Hölder (neste caso, igual ao parâmetro de Hurst), expoente este que indica a regularidade local do processo, foram considerados em alguns trabalhos inapropriados para modelar tráfego em pequenas escalas (Feldmann et al., 1998) (Riedi & Véhel, 1997). A Figura 2.3 mostra a diferença entre dois traços de tráfego, sendo um monofractal e o outro multifractal. Neste último caso, observa-se que o tráfego parece ter diferentes tipos de rajadas.

2.3 Processos Multifractais

A complexidade geométrica de um fractal pode ser descrita, pelo menos de uma forma global, por suas dimensões (Riedi & Véhel, 1997) (Riedi, 1997). Mas se levarmos em consideração que as propriedades em escala do tráfego são causadas por uma dinâmica caótica e por processos aleatórios, é natural que existam vários comportamentos em escala diferentes (*multiple scaling*). De fato, diferentes comportamentos em escala freqüentemente são encontrados em diferentes instante de tempo, gerando o fenômeno denominado multifractal. Na maioria das situações práticas, o grau de auto-similaridade do tráfego é variante com o tempo e a suposição de Gaussianidade para o processo de tráfego reduz o horizonte de aplicação dos modelos monofractais como o fBm (Krishna et al., 2003). Estas propriedades do tráfego de redes e de outras séries temporais influenciaram o surgimento de modelos multifractais. O conceito de multifractal foi introduzido por Mandelbrot no contexto de turbulência nos anos 70. Desde então, a teoria multifractal é usada em vários campos como processamento de imagem, geofísica, etc.

Um processo multifractal é caracterizado por um conjunto de dimensões fractais e a estes fractais se associa um espectro multifractal. Assim, para se obter informações mais detalhadas destes fractais, usa-se ferramentas como a análise multifractal. Esta análise é capaz de descrever o comportamento local de medidas, distribuições e funções de forma geométrica e estatística. A análise multifractal compara o comportamento em escala de momentos estatísticos dos processos para estimar suas regularidades locais (Feldmann et al., 1998)(Riedi et al., 1999). Através da análise multifractal algumas propriedades encontradas em processos multifractais podem ser verificadas. O tráfego de redes ao ser considerado multifractal significa que possui uma estrutura de forte dependência inerente, com incidência de rajadas em várias escalas (Riedi et al., 1999) (Park & Willinger, 2000). Estas características podem degradar o desempenho de rede em comparação a fluxos de tráfego modelados como sendo gaussianos ou de curta-dependência. Processos multifractais são definidos por leis de escala e momentos estatísticos dos processos de incrementos em intervalos de tempo finitos conforme definição a seguir.

Definição 2.3.1 *Um processo estocástico $X(t)$ é multifractal se satisfaz a equação:*

$$E(|X(t)|^q) = c(q)t^{\tau(q)+1} \tag{2.16}$$

onde $t \in T$ e $q \in$ Q, T e Q são intervalos na reta real, $\tau(q)$ e $c(q)$ são funções com domínio Q. Normalmente, assume-se que T e Q têm comprimentos positivos, e que $0 \in T$, $[0,1] \subseteq Q$.

A definição 2.3.1 descreve o comportamento multifractal em termos de momentos estatísticos onde $\tau(q)$ e $c(q)$ são conhecidos como a função de escala e o fator de momento de um processo mul-

tifractal, respectivamente. Se $\tau(q)$ é linear em q, o processo é chamado monofractal; caso contrário, é multifractal.

Para processos auto-similares com parâmetro de Hurst H, pode-se mostrar que $\tau(q) = qH - 1$ e $c(q) = E(|X(1)|^q$. Processos que obedecem a propriedade (2.16) são invariantes em escala (Muzy et al., 2000). São exemplos de processos multifractais: o movimento Browniano multifracionário (mBm-Multifractal Brownian motion) (Peltier & Véhel, 1995), o processo de Lévy (Embrechts & Maejima, 2000) e cascatas multiplicativas (Riedi, 2003).

2.3.1 Caracterização da Regularidade Local

O parâmetro de Hurst é uma propriedade global de um processo estocástico e quantifica como todo o processo varia com a mudança de escala. Uma caracterização do comportamento em escala local também pode ser feita utilizando conceitos de singularidade local de uma função em um ponto. Define-se ponto singular como um ponto em que uma equação, curva, superfície etc, possua transições ou torna-se degenerada. Pontos singulares freqüentemente carregam informações essenciais em um sinal. Em vários tipos de sinais, desde sinais de eletrocardiograma até sinais de fala, a informação de interesse está contida nas singularidades presentes (Daoudi & Véhel, 1995) (West et al., 2004). Particularmente para sinais de tráfego de redes de computadores, o grau da sua regularidade está intimamente ligado ao grau das rajadas de dados (*burstiness*) (Krishna et al., 2003) (Seuret & Gilbert, 2000).

Sinais fractais possuem singularidades não-isoladas, ou seja, apresentam comportamento singular em quase todos os pontos. Para caracterizar estruturas singulares, é necessário precisamente quantificar a regularidade de um sinal. Existem várias maneiras de se medir a regularidade de um sinal. Um método de natureza geométrica muito utilizado é o de encontrar a dimensão fractal do gráfico de uma função. Falando de forma superficial, é preciso determinar como esse gráfico preenche o espaço em pequenas escalas. A precisão deste método é muito sensível à dimensão fractal utilizada, sendo as mais freqüentes a de Hausdorff, a da caixa e a de Tricot (Falconer, 1990). Outra forma de quantificar a regularidade do sinal, a qual será considerada neste trabalho, consiste nos métodos baseados nos expoentes de Hölder. Estes são os métodos mais utilizados para a caracterização da regularidade local de um sinal e utilizam o expoente de Hölder local ou, em especial, o expoente de Hölder pontual. O expoente de Hölder pontual é a medida mais utilizada para quantificar a regularidade pontual de sinais. Ele indica o grau de 'rajadas' presentes no tráfego. O tipo de expoente escolhido depende da aplicação a ser feita. Antes de definirmos estes dois tipos de expoentes, inicialmente precisamos definir alguns conceitos.

Definição 2.3.2 *Seja $\Omega \subset \mathbb{R}^n$, $k > 0$ e a função $f : \Omega \to \mathbb{R}$. $C^k(\Omega)$ define o conjunto de funções f*

que são k vezes diferenciáveis, com derivadas contínuas.

Definição 2.3.3 *(Espaço métrico): Um espaço métrico é um par ordenado (X, ρ) em que X é um conjunto não-vazio com uma função $\rho : X \times X \to [0, \infty)$, satisfazendo:*

1. *(não-degeneração) $\rho(x, y) = 0 \Leftrightarrow x = y$*
2. *(simetria) $\rho(x, y) = \rho(y, x)$ para todo $x, y \in X$*
3. *(desigualdade triangular) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$, para todo $x, y, z \in X$*

A função ρ é denominada de uma métrica em *X* e o número real não-negativo *r(x,y)* é chamado de distância de *x* a *y*.

Definição 2.3.4 *Seja (X,ρ) um espaço métrico, r um número real estritamente positivo e $x, x_0 \in X$. O conjunto*

$$B(x_0, r) = \{x : \rho(x_0, x) < r\} \tag{2.17}$$

é denominado de bola aberta com centro em x_0 e raio r.

Definição 2.3.5 *Seja α um número real estritamente positivo, $x_0 \in \mathbb{R}$, e uma função $f : \mathbb{R} \to \mathbb{R}$. Caso $0 < \alpha < 1$, podemos dizer que $f \in C^\alpha(B(x_0, r))$ se existir uma constante K tal que, para todo x, y em $B(x_0, r)$,*

$$|f(x) - f(y)| \leq K|x - y|^\alpha \tag{2.18}$$

De um modo mais geral, considerando $m < \alpha < m + 1$ ($m \in N$), podemos dizer que $f \in C^\alpha(B(x_0, r))$ caso exista uma constante K tal que, para todo x,y em $B(x_0, r)$,

$$|\partial^m f(x) - \partial^m f(y)| \leq K|x - y|^{\alpha - m} \tag{2.19}$$

em que ∂^m é um operador diferencial de ordem m.

Definição 2.3.6 *(Expoente de Hölder local): Seja f uma função tal como descrita na definição 2.3.5 e,*

$$\alpha_l(x_0, B(x_0, r)) = \sup\{\alpha : f \in C^\alpha(B(x_0, r))\}. \tag{2.20}$$

Note que $\alpha_l(x_0, B(x_0, r))$ é não crescente em função de r. O expoente de Hölder local α_l de f em x_0 é definido como

$$\alpha_l(x_0) = \lim_{r \to 0} \alpha_l(x_0, B(x_0, r)). \tag{2.21}$$

Este tipo de expoente de Hölder é estável sob a ação de operadores diferenciais ou integradores. No entanto, sua principal desvantagem é que a função de Hölder local, $x \rightarrow \alpha_l(f,x)$, é uma função semicontínua inferior, ou seja,

$$\forall x_0 \in \mathbb{R}, \forall \varepsilon, \exists \eta :$$
$$y \in B(x_0,\eta) \Rightarrow \alpha_l(f,y) > \alpha_l(f,x_0) - \varepsilon \tag{2.22}$$

Um subconjunto X de $\mathbb{R}$ é denso em $\mathbb{R}$ se todo intervalo aberto de $\mathbb{R}$ contém algum elemento de X. Segundo (Hobson, 1958), duas funções semicontínuas inferiores, coincidentes em um conjunto denso, são iguais. Assim, por exemplo, não é possível observar um ponto isolado "regular" em um ambiente "irregular" (sinal repleto de singularidades) por meio do expoente de Hölder local (Daoudi et al., 1998). Uma solução alternativa e mais versátil é se considerar o expoente de Hölder pontual.

Definição 2.3.7 *(Expoente de Hölder pontual): Seja α um número real estritamente positivo, K uma constante e $x_0 \in \mathbb{R}$. A função $f : \mathbb{R} \rightarrow \mathbb{R}$ é $C^\alpha(x_0)$ se existe um polinômio P_n, de grau $n < \alpha$, tal que*

$$|f(x) - P_n(x - x_0)| \leq K|x - x_0|^\alpha. \tag{2.23}$$

O expoente de Hölder pontual α_p da função f em x_0 é definido como

$$\alpha_p(x_0) = \sup\{\alpha > 0 | f \in C^\alpha(x_0)\}. \tag{2.24}$$

Esta caracterização da regularidade de uma função é amplamente utilizada em Análise Multifractal. Por exemplo, em (Seuret & Gilbert, 2000), mostra-se que o expoente de Hölder pontual pode quantificar o grau da variação instantânea de um sinal de tráfego de redes. Mais precisamente, este expoente pode indicar o grau das rajadas de dados presentes neste sinal. Note que $\alpha > 1$ corresponde a instantes com pequenas variações do tráfego, enquanto $\alpha < 1$, indica regiões com alto nível de variações ou rajadas.

Do ponto de vista matemático, funções de Hölder pontuais (ou seja, funções do tipo $x \rightarrow \alpha_p(f,x)$), são dadas pelo limite inferior de uma seqüência de funções contínuas (Daoudi et al., 1998). Isto permite a aplicação destas funções em uma grande variedade de situações.

2.3.2 Espectro Multifractal

A informação local das singularidades de um sinal é dada pelo expoente de Hölder em cada ponto do espectro, enquanto a informação global é capturada pela caracterização da distribuição geométrica

ou estatística dos expoentes de Hölder, denominado espectro "multifractal". O espectro multifractal, representado por $f(\alpha)$, é uma representação conveniente para a distribuição dos expoentes de Hölder em um processo.

Considerando as equações (2.23) e (2.24), verifica-se que para uma realização fixa de um determinado processo *Z(t)*, suas variações infinitesimais nas proximidades de *t* são descritas por

$$|Z(t+\Delta t) - Z(t)| \sim C_t(\Delta t)^{\alpha(t)}, \tag{2.25}$$

onde C_t é chamado de pré-fator (Mandelbrot et al., 1997). Percebe-se através da equação (2.25) que $\alpha(t)$ pode ser visto como um fator de escalonamento local em *t*.

Mostra-se a partir da equação (2.25), que o expoente de Hölder de uma realização de um processo contínuo em um instante *t* é dado por

$$\alpha(t) = \sup\left\{\alpha :\ \alpha = \frac{\ln|Z(t+\Delta t) - Z(t)|}{\ln \Delta t}\ quando\ \Delta t \to 0\right\}. \tag{2.26}$$

A partir da equação (2.26), define-se um estimador para o expoente de Hölder de um processo $Z(t)$.

Definição 2.3.8 *- Seja um processo Z(t) com suporte no intervalo* $[0,T]$ *. Subdivida iterativamente o intervalo [0,T] em* b^k *partes de mesmo tamanho, onde k identifica o estágio na seqüência de subdivisões. Calculando-se o valor* $\left|Z(t_i + b^{-k}T) - Z(t_i)\right|$ *para cada* b^k *subdivisões, o expoente de Hölder aproximado (coarse Hölder exponent) é definido por*

$$\alpha_k(t_i) \equiv \frac{\ln\left|Z(t_i + b^{-k}T) - Z(t_i)\right|}{\ln b^{-k}}. \tag{2.27}$$

A equação (2.27) pode conduzir a um método para estimar a probabilidade de que um ponto aleatoriamente escolhido no intervalo $[0,T]$ tenha um dado expoente de Hölder. Para isso, é necessário dividir a faixa de α ´s em pequenos intervalos não sobrepostos, $(\bar{\alpha}_j, \bar{\alpha}_j + \Delta\alpha]$ tal que $N_k(\bar{\alpha}_j)$ seja o número de expoentes de Hölder aproximados $\alpha_k(t_i)$ contidos em cada intervalo $(\bar{\alpha}_j, \bar{\alpha}_j + \Delta\alpha]$. Quando $k \to \infty$, a razão $N_k(\alpha)/b^k$ converge para a probabilidade que um ponto *t* aleatoriamente escolhido possua expoente de Hölder igual a α (Mandelbrot et al., 1997).

Embora nos processos multifractais exista um valor de expoente de Hölder α_0 mais freqüente, outros valores de expoentes de Hölder também ocorrem. Tais expoentes de Hölder com valores diferentes de α_0 são bastante importantes, uma vez que a maior parte das variações em uma função multifractal encontra-se em instantes com expoente de Hölder diferentes de α_0. Tal característica permite discriminar multifractais de monofractais, dando origem a definição 2.3.9.

Definição 2.3.9 *Seja $N_k(\bar{\alpha}_j)$ o número de expoentes de Hölder aproximados iguais a α que ocorrem ao subdividir o processo $Z(t)$ em b^k partes de mesmo tamanho. Então o espectro multifractal, representado por $f(\alpha)$, é definido por:*

$$f(\alpha) \equiv \lim \left\{ \frac{\ln N_k(\alpha)}{\ln b^k} \right\} \; para \; k \to \infty \tag{2.28}$$

Caso o limite anterior exista e $f(\alpha)$ seja definido e positivo em um suporte maior que um único ponto, então diz-se que $Z(t)$ é multifractal. Processos cujo espectro $f(\alpha)$ é definido para apenas um ponto, apresentando um único expoente de Hölder, são classificados como monofractais. Para processos multifractais, o espectro apresenta uma forma parabólica côncava onde $f(\alpha) \leq \alpha(t)$, para todo $\alpha(t)$ e $f(\alpha) \leq f(\alpha_0)$ para todo $\alpha(t)$, onde $f(\alpha_0)$ é o valor máximo de $f(\alpha)$ (Riedi, 1997).

2.4 Estimação da Característica Multifractal

As características multifractais podem ser constatadas utilizando as ferramentas da análise multifractal, as quais serão abordadas nesta seção. A primeira abordagem é baseada na estimação da função de partição do processo, a segunda está ligada a estimação do espectro multifractal e a terceira está relacionada a estimação da regularidade do processo, ou seja, do expoente de Hölder.

2.4.1 Estimação da Função de Partição

A partir da definição 2.3.1 percebe-se que a descrição de um processo multifractal envolve tanto o conhecimento da função $c(q)$ quanto da função $\tau(q)$. A seguir, apresenta-se um método simples para testar a característica multiescala do processo, assim como estimar as funções $c(q)$ e $\tau(q)$.

Considere os dados $(Z_i)_{i=1}^N$ com suporte no intervalo $[0, T]$, em uma escala $\delta = {}^{T}/_{N}$. Define-se a soma partição como

$$S_m^Z(q) = \sum_{k=1}^{N/m} \left(\bar{Z}_k^{(m)} \right)^q, \tag{2.29}$$

onde

$$\bar{Z}_k^{(m)} = \sum_{l=1}^{m} Z_{(k-1)m+l} \tag{2.30}$$

é o processo original observado em uma escala de agregação $\delta = {}^{T.m}/_{N}$.

Para um valor fixo de q_i, variam-se os valores de *m* em uma faixa apropriada, obtendo-se um conjunto de pontos no plano $\log m$ x $\log S_m^Z(q_i)$. O coeficiente angular da reta obtida através da aproximação por mínimos quadrados é denominado $\tau(q_i)$, sendo $c(q_i)$ igual ao ponto de intersecção

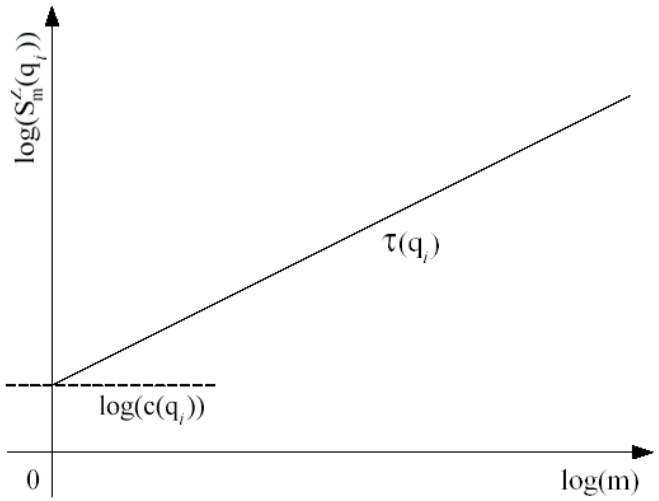


Fig. 2.4: Estimação das funções $\tau(q)$ e $c(q)$.

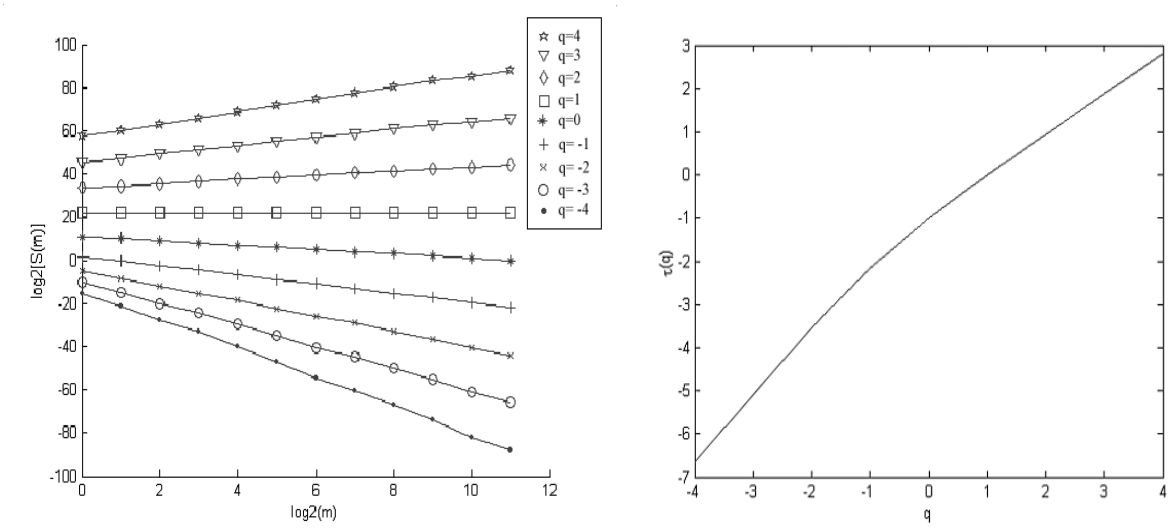
entre o eixo cartesiano y e a reta em questão. A Figura 2.4 ilustra melhor a forma de obtenção dos valores de $\tau(q_i)$ e $c(q_i)$, relacionados pela equação

$$\log S_m^Z(q_i) \cong \tau(q_i).\log m + \log c(q_i). \tag{2.31}$$

Uma vez conhecidos $\tau(q_i)$ e $c(q_i)$ para diferentes valores de q_i, obtém-se então as funções $\tau(q)$ e $c(q)$. A Figura 2.5(a) mostra a curva da função soma-partição obtida para a série representativa do número de *bytes* por quadro de vídeo codificado MPEG-4 do filme *Silêncio dos Inocentes* (Fitzek & Reisslein, 2000). Para a obtenção da função soma-partição foram considerados valores de $q_i \in$ [-4;4] e $m \in$ [1;2048]. A Figura 2.5(b) exibe a curva correspondente à função $\tau(q)$ obtida a partir dos diferentes valores de $\tau(q_i)$ para a soma partição da seqüência de vídeo considerada.

Pode-se observar que a função $\tau(q)$ não é linear para processos multifractais. Tipicamente o valor da derivada da função $\tau(q)$ varia muito pouco, normalmente estando no intervalo $[1/2; 2]$, levando o gráfico de $\tau(q)$ parecer quase linear. Sendo assim, uma análise feita apenas através de inspeção visual da função $\tau(q)$ pode levar a falhas. Portanto, a análise feita através do espectro multifractal é geralmente mais informativa.

Uma outra função importante na análise multifractal é a função de partição baseada em *wavelets*. Ao invés de se calcular os expoentes de regularidade locais, estatísticas do comportamento local de um processo podem ser obtidas por sua função de partição baseada em *wavelet* $S_j(q)$. Considere $Y_{j,k}$ denotando o volume de tráfego (total de *bytes*) observado no instante de tempo k em uma escala



(a) (b)

Fig. 2.5: Soma partição da série representativa do tamanho em *bytes* dos quadros de vídeo codificado MPEG-4 do filme *Silence of the Lambs*. (b) Função partição $\tau(q)$ obtida para a seqüência de vídeo codificado MPEG-4 considerada.

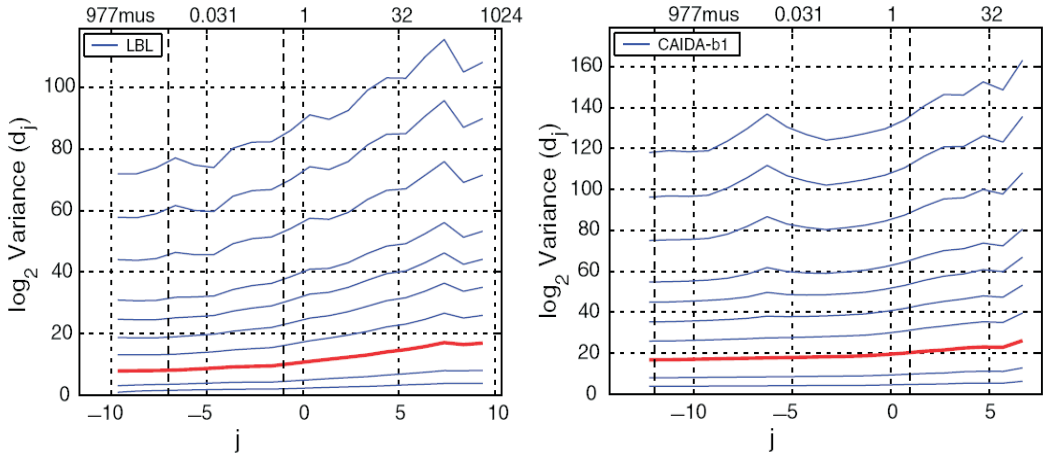
temporal j. A função de partição baseada em *wavelet* é definida como:

$$S_j(q) = E|W_{j,k}|^q \tag{2.32}$$

onde $W_{j,k}$ são os coeficientes *wavelet* de Haar do processo na escala de tempo j e no instante de tempo k, dados por:

$$W_{j,k} = 2^{j/2}(Y_{j+1,2k} - Y_{j+1,2k+1}) \tag{2.33}$$

Esta função de partição calcula os momentos de ordem q dos coeficientes *wavelet* em função da escala e consegue capturar o comportamento local de um processo. A explicação para esta última afirmação é que um expoente de Hölder local de baixo valor gera um coeficiente *wavelet* de valor alto, sendo sua presença evidenciada ainda mais, ao se tomar a potência q desses valores. De forma análoga, quando uma processo é localmente suave, seus coeficientes *wavelet* são pequenos. A Figura 2.6 ilustra o cálculo da função de partição baseada em *wavelet* para diferentes valores de q e distintas séries de tráfego. A partir destes gráficos pode se analisar o comportamento em escala para diferentes valores de q, e concluir se há presença da propriedade de multi-escala (*multiscaling*), considerada evidência para o comportamento multifractal. Ou seja, se para um determinado valor de q encontra-se um reta no gráfico apresentado para uma gama de escalas, este fenômeno é denominado de caracte-



(a) (b)

Fig. 2.6: Diagrama Log-escala de ordem q para as séries de tráfego LBL-TCP-3 (esquerda) e CAIDA-b1 (direita). Debaixo para cima as ordens são $q = 0.5, 1, 2, 3, 4, 5, 6, 8, 10, 12$

rística multiescala. A função de partição se comporta assintoticamente $(j \to \infty)$ da seguinte forma:

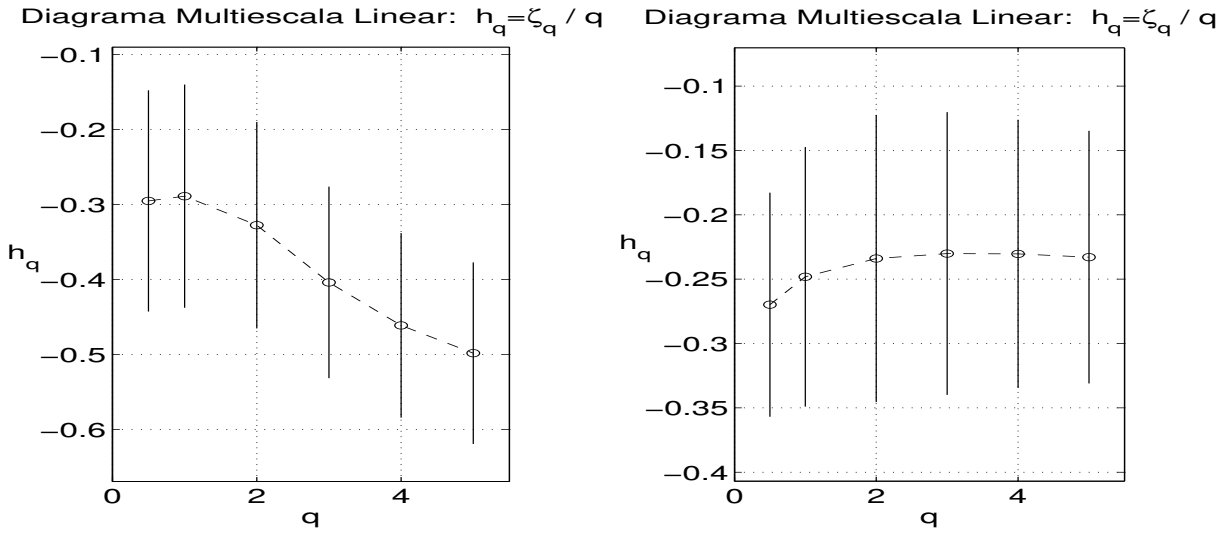
$$\log_2 S_j(q) \sim q.constante + j\alpha_q \tag{2.34}$$

Assim, a inclinação de $\log_2 S_j(q)$ em relação a j provê uma estimativa de α_q. Para verificar se um processo é multifractal, duas medidas são avaliadas (Veitch et al., 2005): $\zeta_q = \alpha_q - q/2$ e $h_q = \zeta_q/q$. Para processos monofractais como o fBm (*fractional Brownian motion*) e seu processo de incrementos o fGn (*fractional Gaussian noise*) (Park & Willinger, 2000), temos $h_q = H$ (Parâmetro de Hurst). Portanto, um gráfico constante de h_q versus q caracteriza um processo monofractal, enquanto o mesmo não ocorre para processos multifractais. Isso pode ser visto pela Figura 2.7, onde é apresentado o gráfico h_q versus q para a série multifractal de tráfego 10-7-S-1 (LRPRC, 2002) na escala de 100ms e para um processo monofractal fGn.

2.4.2 Estimação do Espectro Multifractal

A função $f(\alpha)$ tal como apresentada na definição 2.3.9 é chamada de espectro de granularidade grosseira (*coarse graining spectrum*) ou também espectro de grandes desvios (*large deviation spectrum*). Além do espectro multifractal de grandes desvios, o espectro de Hausdorff e espectro de Legendre também merecem destaque (Falconer, 1990).

Basicamente, qualquer um dos três espectros provê informações sobre quais singularidades ocorrem em um sinal, e quais são as singularidades que predominam. O espectro é uma curva unidimen-



(a) Traço de tráfego multifractal 10-7-S-1

(b) Processo monofractal fGn

Fig. 2.7: Diagrama Multiescala Linear

sional, normalmente com um perfil côncavo, onde a abscissa representa os expoentes de Hölder que efetivamente existem no sinal, e a ordenada está relacionada com a quantidade de pontos onde uma dada singularidade é encontrada. Por exemplo, se um espectro $f(\ \alpha\ )$ possui apenas um máximo em $\alpha = a$, com $f(a) = 1$, então ao selecionarmos aleatoriamente um ponto deste sinal, este possuirá quase certamente expoente de Hölder com valor *a*. Por outro lado, se $\alpha = b$ tal que $f(b) \cong 0$, então existe um conjunto muito esparso de pontos para os quais teremos expoente de Hölder igual a *b*, ou seja, será pequena a probabilidade de ocorrência deste valor de expoente de Hölder no sinal. No caso onde $\alpha = c$ tal que $f(c) = -\infty$, não ocorrerão no sinal expoentes de Hölder com valores iguais a c.

A breve descrição de espectro anteriormente apresentada na seção 2.3.2 é apenas uma visão superficial do conceito de espectro. Na realidade, existem diferenças essenciais entre os três espectros, conforme as definições apresentadas a seguir.

Das mais variadas dimensões fractais existentes, a dimensão de Hausdorff é provavelmente a mais importante, que tem como vantagem ser definida para qualquer conjunto (Falconer, 1990). Normalmente, o conjunto dos pontos com um mesmo grau de singularidade α constituem um conjunto fractal, cuja descrição geométrica é precisamente descrita através de sua dimensão de Hausdorff (Falconer, 1990). O espectro de Hausdorff provê uma informação geométrica pertinente à dimensão fractal dos conjuntos de pontos em um sinal que possuem um dado expoente de Hölder.

Definição 2.4.1 - *Seja K_α o conjunto dos pontos de um sinal Z(t) que apresentem regularidade de*

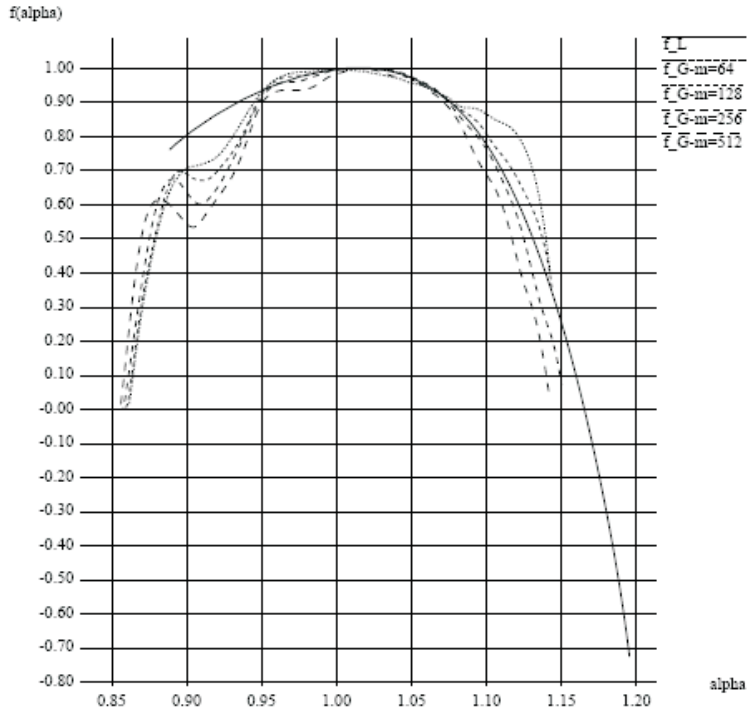


Fig. 2.8: Típico espectro de Legendre e espectro de grandes desvios em diferentes resoluções

Hölder igual a α*. Ou seja:*

$$K_\alpha := \left\{ x \in R^d : \alpha(x) := \alpha \right\}. \tag{2.35}$$

O espectro de Hausdorff de Z(t) é dado por

$$f_H(\alpha) := \dim(K_\alpha), \tag{2.36}$$

onde $\dim(K_\alpha)$ *é a dimensão Hausdorff do conjunto* K_α*.*

Do ponto de vista matemático, este é o espectro multifractal mais preciso, sendo também o mais difícil de ser estimado (Falconer, 1990).

Dado pela definição 2.3.9, o espectro de grandes desvios provê informações estatísticas relacionadas à probabilidade de encontrar no sinal, um ponto com um dado expoente de Hölder. Mais precisamente, o espectro de grandes desvios permite medir como esta probabilidade se comporta quando submetido a mudanças de resolução. Embora o espectro de grandes desvios não seja exatamente a densidade correspondente aos α´s, mas sim uma dupla normalização logarítmica desta densidade, a estimação do espectro de grandes desvios exige a aplicação de ferramentas de estimação de densidade de probabilidade. Neste caso, para a estimação da densidade de probabilidade normal-

mente são empregadas ferramentas clássicas como o método de kernel duplo (Devroye, 1989) (ver Apêndice C).

O espectro de Legendre é uma aproximação côncava do espectro de grandes desvios. Este espectro é de grande interesse pois normalmente permite estimações robustas, embora para alguns sinais específicos (Riedi & Véhel, 1997), omita algumas informações possíveis de serem obtidas através do espectro de grandes desvios. A robustez e a simplicidade de estimação do espectro de Legendre o tornam o mais atrativo espectro multifractal. A Figura 2.8 mostra um típico espectro multifractal de Legendre assim como espectros de grandes desvios em diferentes resoluções de agregação *m´*s.

Em decorrência de sua robustez e por ser o mais atrativo espectro do ponto de vista numérico, o espectro multifractal de Legendre foi o espectro utilizado nesta tese. A já mencionada simplicidade de estimação do espectro de Legendre decorre de sua definição.

Definição 2.4.2 *Seja $\tau(q)$ a função partição de um sinal $Z(t)$. O espectro de Legendre de $Z(t)$ é dado por*

$$f_L(\alpha) := \tau^*(\alpha), \tag{2.37}$$

onde $\tau^*(\alpha)$ é a transformada de Legendre da função partição $\tau(q)$, dada por $\tau^*(\alpha) = \inf_q(q\alpha - \tau(q))$.

O método de estimação do espectro de Legendre de um determinado sinal é dado pela transformada de Legendre de sua função partição $\tau(q)$, obtida por sua vez, através do método apresentado anteriormente na seção 2.4.1.

2.4.3 Estimação da Regularidade Local

O conhecimento do grau de rajadas presentes no tráfego é um importante fator em estratégias de controle que pode ser aproveitado pela rede. Esta seção trata da estimação da regularidade local de um sinal tendo o expoente de Hölder pontual como indicador dessa regularidade. A estimação deste expoente é realizada com base no decaimento do valor absoluto dos coeficientes *wavelet* do sinal analisado. Para isso, é feita uma breve apresentação da transformada *wavelet* a seguir.

Estimação pelo Algoritmo *WTMM*

A transformada *wavelet* é uma ferramenta poderosa para caracterização da regularidade local de um sinal (Daubechies, 1992). Apresentamos aqui algumas definições oriundas do estudo desta transformada a fim de estabelecermos os passos necessários para a estimação do expoente de Hölder. Conceitualmente, a transformada *wavelet* é um produto-convolução do sinal analisado com a *wavelet* mãe ψ (Ver Apêndice A). Neste processo, a *wavelet*-mãe deve ser ajustada a uma determinada escala

j e transladada até um ponto $2^j k$ de um sinal $f(x)$, com $j, k \in \mathbb{Z}$. Desta forma, o coeficiente *wavelet* $d_{j,k}$ é dado por (Daubechies, 1992):

$$d_{j,k} = 2^{-j} \int_{-\infty}^{\infty} f(x)\psi\left(2^{-j}x - k\right) dx \tag{2.38}$$

Um dos métodos mais usados para detecção de singularidades em sinais e que envolve o conhecimento dos coeficientes de detalhes $d_{j,k}$ de uma série temporal é o método dos Máximos em Módulos da Transformada Wavelet (*Wavelet Transform Modulus Maxima* - WTMM) (Struzik, 2000). O algoritmo WTMM faz uso das seguintes definições de máximo em módulo e linha de máximos (Mallat & Hwang, 1992):

Definição 2.4.3 *Sejam* $d_{j,k}{}'s$ *os coeficientes* wavelet *de um processo expressos pela equação (2.38), tem-se que:*

1. *É denominado de* ***máximo em módulo****, qualquer ponto* (j_0, k_0) *tal que* $|d_{j_0,k}| < |d_{j_0,k_0}|$ *, na qual k pertence tanto à vizinhança à esquerda ou à direita de* k_0*.*
2. *É chamado de* ***linha de máximos****, qualquer curva conectada no espaço-escala* (k, j) *formada por pontos que são máximos em módulo.*

Na realidade, a linha de máximos contém os pontos mais significativos dentre aqueles que estão imersos em um cone invertido, definido por

$$|2^j k - x_0| \leq K 2^j, \tag{2.39}$$

onde K é uma constante. Dentro deste cone, a linha de máximos converge para a singularidade presente no ponto $x_0 = 2^j k_0$ de um sinal, obedecendo a seguinte relação (Seuret & Gilbert, 2000):

$$|d_{j,k}| \leq A 2^{j\alpha}, \tag{2.40}$$

ou, equivalentemente:

$$\log |d_{j,k}| \leq \log A + \alpha \log(2^j), \tag{2.41}$$

nas quais *A* é uma constante.

A desigualdade (2.41) mostra que o expoente de Hölder pontual α em x_0 é o maior coeficiente angular das retas que estão acima de $|d_{j,k}|$ na escala logarítmica. A Figura 2.9 apresenta uma representação do cone referente a uma das linhas de máximos, e de alguns pontos que são máximos em módulo. Temos então representado pela equação (2.41) a relação entre o expoente de Hölder e os coeficientes de detalhes. Tomando isto como base e levando em consideração o tipo de singularidades

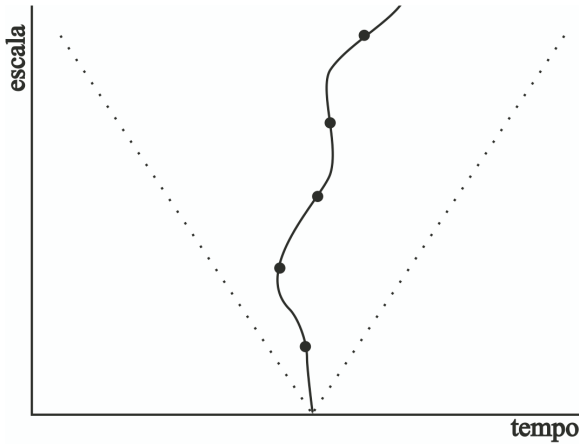


Fig. 2.9: Cone referente a uma linha de máximos

presentes em um sinal, Seuret et. al. propuseram um algoritmo de estimação dos expoentes de Hölder pontuais (Seuret & Gilbert, 2000). Este algoritmo, apresentado a seguir, permite o cálculo do valor da regularidade pontual nos instantes de tempo desejados.

Assim, seja um sinal amostrado, contendo 2^n amostras. Seja também $d_{j,k}$ seus coeficientes *wavelets* estimados usando a transformada *wavelet* não-dizimada (ou seja, redundante) (Daubechies, 1992). A estimação do expoente de Hölder pontual para cada amostra k_0 é feita da seguinte maneira (Seuret & Gilbert, 2000):

Algoritmo 2.4.1

Passo 1) Construa, em uma mesma figura, para cada $0 < j \leq n$ *, a seguinte curva paramétrica (com parâmetro* $k \leq 2^n$*):*

$$x_j(k) = \log_2(2^j + |2^j k - x_0|) \tag{2.42}$$

$$y_j(k) = \log_2(|d_{j,k}|) \tag{2.43}$$

Passo 2) Encontre todas as retas D: y = ax + C que satisfaçam as seguintes restrições:

1.A reta D está acima de todos os pontos (x_j*(k),*y_j*(k)), ou seja:*

$$\forall j, \forall k, \quad y_j(k) \geq \alpha x_j(k) + C \tag{2.44}$$

2.A reta D toca uma das curvas paramétricas; ou seja, existe uma seqüência de pares (j_m*,* k_m*) tal que:*

$$\lim_{m \to \infty} y_{j_m}(k_m) - (\alpha x_{j_m}(k_m) + C) = 0 \tag{2.45}$$

Passo 3) Considere α_{max} *o maior coeficiente angular encontrado entre todas as retas D que satisfaçam as restrições (2.44) e (2.45). O coeficiente* α_{max} *é o* ***expoente de Hölder pontual*** *do sinal*

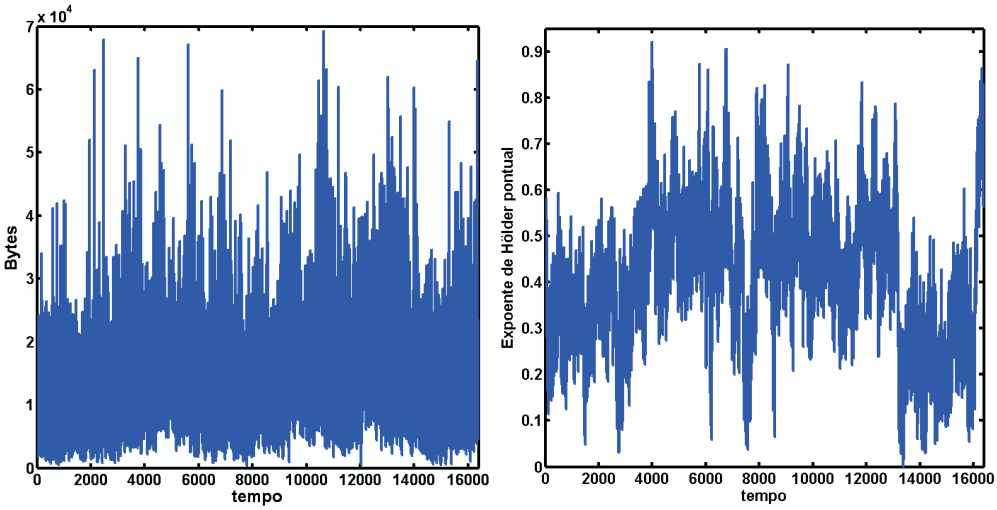
para a amostra k_0.

Este é o método de estimação do expoente de Hölder pontual utilizado neste trabalho. Em (Jorge et al., 2005a) (Jorge et al., 2005b), apresentamos resultados sobre o desempenho deste algoritmo em janelas de tempo ou seja, o empregando na estimação adaptativa do expoente de Hölder utilizando as amostras correspondentes a essas janelas de tempo. Nos referidos trabalhos, também aplicamos ferramentas de predição adaptativa para estimar valores futuros da série formada pelos expoentes de Hölder com o intuito de antecipar a caracterização do comportamento do tráfego e assim inserir essa informação em um escalonador GPS (Generalized Processor Sharing). A Figura 2.10 refere-se ao traço de tráfego lbl-pkt-5 (na escala de tempo de 100 ms), onde é mostrado seus expoentes de Hölder pontuais obtidos com o algoritmo acima descrito, assim como seu espectro multifractal de Legendre.

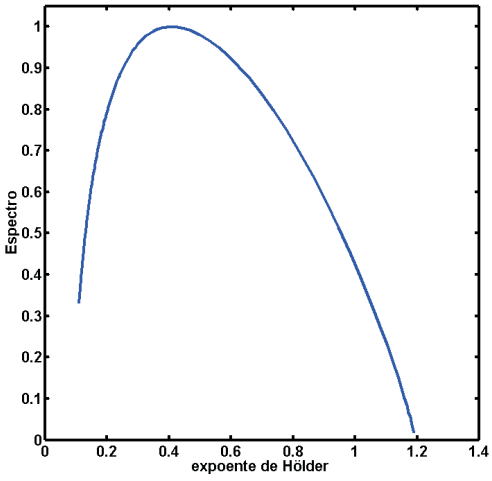
2.5 Modelagem Multifractal de Tráfego

Em alguns trabalhos (Riedi & Véhel, 1997)(Feldmann et al., 1998), mostrou-se que o comportamento em escala do tráfego de redes WAN pode ser dividido em duas principais regiões de resolução: em grandes escalas de tempo (da ordem de centenas de milissegundos e maior) onde o comportamento em escala é caracterizado pelo fenômeno da auto-similaridade, enquanto, em pequenas escalas de tempo (da ordem de centenas de milissegundos e menor), o tráfego WAN é mais bem descrito através da análise multifractal. A abordagem desta análise tem sido recentemente questionada por alguns trabalhos, principalmente para tráfego de *backbone* Internet (Zhang et al., 2003) e em termos do grau de 'multifractalidade' presente nos traços de redes (Rolls et al., 2005). Entretanto, é importante notar que a análise multifractal generaliza e refina de uma forma natural o comportamento auto-similar observado no tráfego de redes. Processos auto-similares de segunda ordem ou, mais genericamente, monofractais, apresentam regularidade e comportamento em escala constantes no tempo, e normalmente dependem apenas de um único parâmetro, o parâmetro de Hurst H. Por outro lado, processos multifractais permitem que tais características variem no tempo, portanto, possibilitando maior flexibilidade em descrever fenômenos irregulares localizados no tempo.

Modelos estatísticos derivados de processos multifractais são capazes de representar de forma mais completa e precisa o real comportamento do tráfego de redes. Neste sentido, foram propostos modelos baseados na análise multifractal, dentre os quais podem ser destacados: modelo *wavelet* multifractal (*Multifractal Wavelet Model*, MWM) (Riedi et al., 1999), cascatas multiplicativas (Krishna et al., 2003) e movimento Browniano multifracionário (*multifractional Brownian motion*, mBm) (Peltier & Véhel, 1995), os quais são apresentados nas próximas seções. Com relação aos resultados de desempenho de fila que levam em conta as características de processos multifractais, Riedi et



(a) (b)



(c)

Fig. 2.10: Acima à esquerda: amostras de tráfego da série lbl-pkt-5 na escala de tempo de 100 ms. Acima à direita: expoentes de Hölder pontuais referentes às amostras citadas. Abaixo: espectro multifractal

al. (Riedi et al., 1999) propuseram um modelo baseado em *wavelets* e desenvolveram uma análise de fila multiescala. Gao e Rubin (Gao & Rubin, 1999b) simularam filas alimentadas com processos multifractais. Em (Dang et al., 2003), um modelo multifractal é usado para derivar uma aproximação analítica para a probabilidade de perda.

2.5.1 Movimento Browniano Multifracionário

O fBm é um processo auto-similar capaz de descrever convenientemente sinais irregulares que ocorrem em diversas situações, e sua regularidade pontual é constante e igual ao parâmetro de Hurst em todos os pontos. Entretanto, alguns conjuntos de dados reais possuem regularidade pontual variável, e para tais casos, um único escalar *H* pode não prover uma descrição adequada da regularidade do conjunto. Para superar as limitações existentes no fBm, Peltier e Véhel introduziram o movimento Browniano multifracionário (*multifractional Brownian motion*, mBm) (Peltier & Véhel, 1995). Diferente do fBm, no mBm é permitido que o expoente de Hölder varie no tempo, sendo descrito por uma função *H*(*t*). Tal característica é muito útil quando se faz necessário modelar processos em que a regularidade varia no tempo, tal como o tráfego Internet.

Definição 2.5.1 *- Sejam (X, d_X) e (Y, d_Y) dois espaços métricos. Uma função f : X → Y é denominada uma função Hölder com expoente $\beta > 0$, se para cada x,y ∈ X tal que $d_X(x, y) < 1$ tenha-se que*

$$d_Y(f(x), f(y)) \leq c.d_X(x, y)^{\beta} \tag{2.46}$$

para alguma constante c > 0.

Definição 2.5.2 *- Seja H: [0, ∞) → [a,b] ⊂ (0,1) uma função Hölder com expoente $\beta > 0$. Para t ≥ 0, a seguinte função aleatória W é chamada movimento Browniano multifractal com função H(t)*

$$W_{H(t)}(t) = \int_{-\infty}^{0} \left[(t-s)^{H(t)-1/2} - (-s)^{H(t)-1/2} \right] dB(s) + \int_{0}^{t} (t-s)^{H(t)-1/2} \, dB(s), \tag{2.47}$$

onde B é o movimento Browniano.

Definição 2.5.3 *- O processo W(t), t ≥ 0, é chamado de movimento Browniano multifracionário padrão se for verificada a seguinte propriedade:*

$$var\left(\frac{W(t+h) - W(t)}{h^{H(t)}} \right) \underset{h \to 0}{\to} 1, \tag{2.48}$$

onde $H(t)$ é a função Hölder do processo W(t).

O movimento Browniano multifracionário (mBm) perde algumas das propriedades do fBm, e embora seja um processo gaussiano, seu processo de incrementos em geral não é estacionário. Sendo uma generalização do fBm, quando $H(t) = H$ para todo t, o mBm se torna então simplesmente o fBm com expoente H. Outra importante propriedade do mBm diz respeito à regularidade do processo descrita pelo expoente de Hölder. Para o mBm, em cada ponto $t_0 \geq 0$, o expoente de Hölder $\alpha(t_0)$ é dado pelo valor da função Hölder do processo naquele ponto, ou seja, $\alpha(t_0) = H(t_0)$.

O modelo mBm é adequado para caracterização de séries reais gaussianas assim como para aplicação na predição de valores futuros de intensidade de tráfego para essas séries. Demonstramos estes fatos nos seguintes artigos (Bianchi et al., 2004b) (Bianchi et al., 2004c) (Bianchi et al., 2004a).

2.5.2 Cascatas Multiplicativas

Cascatas multiplicativas foram inicialmente propostas por Kolmogorov para modelagem de turbulência (Kolmogorov, 1962). Atualmente, o modelo de cascatas multiplicativas tem encontrado aplicações em diversas áreas que necessitam de modelar fenômenos não-lineares e que apresentam estrutura multiplicativa, tais como modelagem de tráfego (Riedi et al., 1999), fenômenos geofísicos (Gupta & Waymire, 1993), evolução do DNA (Bickel & West, 1998), etc. Os modelos multifractais baseados em cascata permitem o uso da análise *wavelet* na estimação de seus parâmetros e apresentam também distribuição de cauda pesada, mais precisamente, lognormal com todos os momentos finitos, em concordância com a distribuição dos dados de tráfego medido (Feldmann et al., 1998).

A cascata binomial é o método mais simples de se obter um processo multifractal, consistindo de um procedimento iterativo no intervalo compacto [0,1]. Sejam m_0 e m_1 (multiplicadores da cascata) dois números positivos cuja soma é 1. No estágio $k = 0$ da cascata, obtemos a medida inicial μ_0 do processo com valor aleatório entre [0,1]. Conforme pode ser visto pela Figura 2.11, no estágio $k = 1$, a medida μ_1 distribui massa, sendo, m_0 no subintervalo [0,1/2] e massa igual a m_1 em [1/2, 1]. Em $k = 2$, o intervalo [0,1/2] é subdividido em [0,1/4] e [1/4,1/2] e o mesmo acontece com intervalo [1/2,1], obtendo-se (Mandelbrot et al., 1997):

$$\mu_2[0, 1/4] = m_0 m_0 \quad \mu_2[1/4, 1/2] = m_0 m_1,$$
$$\mu_2[1/2, 3/4] = m_1 m_0 \quad \mu_2[3/4, 1] = m_1 m_1.$$

Com a repetição desse processo podemos gerar a seqüência de medidas μ_k, que converge então para o processo multifractal μ. Considere o intervalo diádico $[t, t + 2^{-k}]$ onde t pode ser escrito na forma binária $t = 0.\eta_1 \ldots . \eta_k = \sum_{i=1}^{k} \eta_i 2^{-i}$ e $\eta_i \in \{0, 1\}$. Esta notação é útil na determinação da medida μ de um intervalo arbitrário no estágio k. Sejam φ_0 e φ_1 as freqüências relativas de 0's e 1's, respectivamente, no desenvolvimento da cascata. A medida μ no intervalo diádico $t, t + 2^{-k}$ é dada

por

$$\mu[t, t + 2^{-k}] = \mu[\Delta_k] = m_0^{k\varphi_0} m_1^{k\varphi_1}. \tag{2.49}$$

Este processo de divisão preserva em cada estágio a massa dos intervalos diádicos, por isso é chamado de cascata conservativa ou microcanônica. Existe uma classe de cascatas denominadas de canônicas, que estendem esta condição e apenas requerem que a medida seja conservada na média (Mandelbrot et al., 1997). Se em cada estágio da cascata os intervalos são divididos em $b > 2$ intervalos de tamanhos iguais, este processo é chamado de cascata multinomial. Em particular, se para $b = 2$ o multiplicador m_0 tem um valor fixo, então a cascata multiplicativa é do tipo binomial determinística com função de escala: $\tau_0(q) = \tau(q) + 1 = -\log_2(m_0^q + m_1^q)$ (Mandelbrot et al., 1997). A Figura 2.12 ilustra o processo obtido com uma cascata binomial com 10 estágios de divisão.

Seja I_k o intervalo diádico $[i2^{-k}, (i + 1)2^{-k}]$, onde $i = 0, \ldots, 2^k - 1$ e k é o estágio da cascata. O expoente de Hölder no intervalo I_k de uma cascata determinística binomial resultante é dado por

$$\alpha(I_k) = \frac{\log \mu(I_k)}{\log(2^{-k})} = \frac{\log[m_0^{k\varphi_0} m_1^{k\varphi_1}]}{-k \log 2}. \tag{2.50}$$

As cascatas multiplicativas podem ser caracterizadas por seus espectros multifractais, o que nos fornece a dimensão fractal do conjunto de intervalos cujo expoente de Hölder é α. Fazendo $\alpha_{min} = -\log_2 m_0$; $\alpha_{max} = -\log_2 m_1$ e $\alpha_0 = (\frac{\alpha_{min} + \alpha_{max}}{2})$, pode-se demonstrar que o espectro multifractal da cascata determinística binomial é dado por

$$f(\alpha) = 1 - \frac{2}{\ln 2} \left(\frac{\alpha - \alpha_0}{\alpha_{max} - \alpha_{min}} \right). \tag{2.51}$$

Da expressão acima, algumas propriedades do espectro multifractal podem ser verificadas. O máximo de $f(\alpha)$ ocorre para $\alpha = \alpha_0$, assim como o comportamento de $f(\alpha)$ nas proximidades de

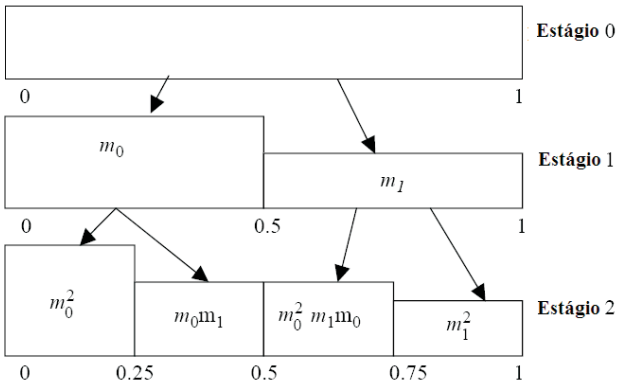


Fig. 2.11: Processo de construção de cascata binomial

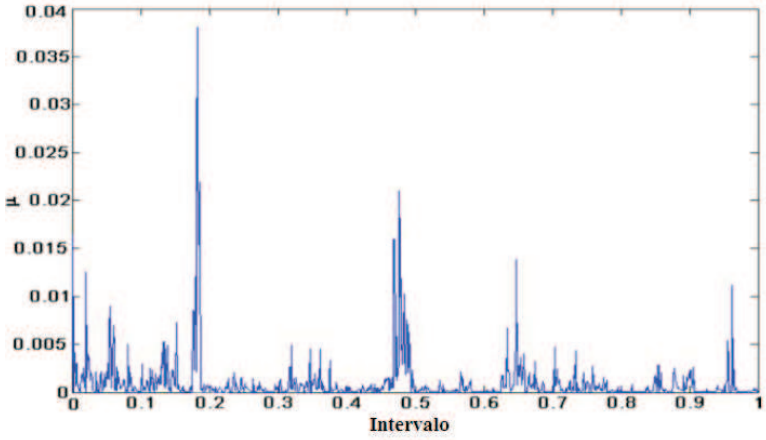


Fig. 2.12: Uma cascata binomial conservativa após 10 iterações

$\alpha = \alpha_0$ é quadrático. A Figura 2.13 apresenta o espectro multifractal e os expoentes de Hölder para uma cascata binomial com 14 estágios, onde pode-se verificar o que foi mencionado sobre o seu espectro e ainda observar o comportamento aproximadamente 'periódico' dos expoentes de Hölder.

Até então, analisamos a cascata multiplicativa com multiplicadores fixos m_0 e m_1. Ao se permitir que os multiplicadores da cascata sejam variáveis aleatórias independentes em [0,1] com densidade de probabilidade $f_R(x)$, obtém-se uma estrutura mais geral do que a determinística em que os multiplicadores são valores fixos. Dessa forma, o processo multifractal obtido $\{\mu(\Delta t_k)\}_{k=1}^{2^N}$ terá no estágio *i* da cascata e no intervalo diádico de comprimento $\Delta t_k = 2^{-k}$, que começa em $t = 0.\eta_1 \ldots .\eta_k = \sum_{i=1}^{k} \eta_i 2^{-i}$, a medida μ:

$$\mu(\Delta t_k) = R(\eta_1).R(\eta_1, \eta_2), \ldots, R(\eta_1, \ldots, \eta_k), \tag{2.52}$$

onde $R(\eta_1, \ldots, \eta_i)$ é o multiplicador no estágio *i* da cascata. Uma vez que os multiplicadores $R(\eta_1, \ldots, \eta_i)$ são i.i.d, pode-se demonstrar que a medida μ satisfaz a relação de escala (Mandelbrot et al., 1997):

$$E(\mu(\Delta t_k)^q) = (E(R)^q)^k = (\Delta t_k)^{\tau(q)+1} = \Delta t_k^{-\log_2 E(R^q)} \tag{2.53}$$

que define um processo multifractal com função de escala $\tau(q) = -\log_2 E(R^q) - 1$. Comparando a equação (2.53) com a Definição 2.3.1, pode-se notar que a cascata binomial satisfaz a definição de processo multifractal. Supondo que os multiplicadores da cascata em cada estágio sejam i.i.d., que a variância dos multiplicadores é a mesma em todos os estágios e o momento de segunda ordem W_2^N para uma cascata com N estágios seja dado por

$$E[(R(\eta_1).R(\eta_1, \eta_2), \ldots, R(\eta_1, \ldots, \eta_k))^2] = W_2^N, \tag{2.54}$$

então, pode-se inferir a média e a variância da cascata multiplicativa como

$$E[(\mu(\Delta t_k)] = E[R(\eta_1).R(\eta_1,\eta_2),\ldots,R(\eta_1,\ldots,\eta_k)] = \left(\frac{1}{2}\right)^N; \tag{2.55}$$

$$var[(\mu(\Delta t_k)] = E[(R(\eta_1).R(\eta_1,\eta_2),\ldots,R(\eta_1,\ldots,\eta_k))^2] - \left(\frac{1}{2}\right)^{2N} = W_2^N - \left(\frac{1}{2}\right)^{2N}. \tag{2.56}$$

Sob as condições citadas, o momento M_q^N de ordem q para a série agregada pode ser escrito da seguinte forma:

$$M_q^N = \frac{1}{m}\sum_{k=1}^{m}(\mu(\Delta t_k) = \frac{1}{m}\sum_{k=1}^{m}(R(\eta_1).R(\eta_1,\eta_2),\ldots,R(\eta_1,\ldots,\eta_k))^q] = W_q^N \tag{2.57}$$

Para processos com longa dependência a relação da variância com o tempo pode ser expressa por $Var(V^m) \sim m^{2H-2}$ onde $1/2 < H < 1$, $m = 2^k$ e $V^m = \frac{1}{m}\sum_{k=1}^{m}(\mu(\Delta t_k))$. Para as cascatas multiplicativas obtém-se

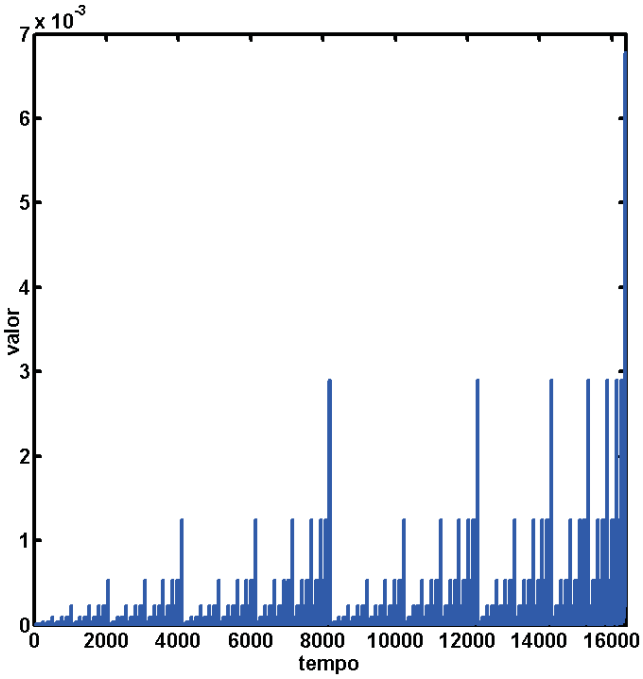
$$V^m = W_2^N(4W_2^N)^{-k} - 2^{-2N}. \tag{2.58}$$

Nota-se que a variação de $\log V^m$ versus $\log m$ é linear, o que provê uma estimativa do parâmetro de Hurst igual a $1 - \frac{\log_2(4W_2^N)}{2}$. Portanto, tráfego com dependência de longo prazo pode ser modelado usando cascatas multiplicativas.

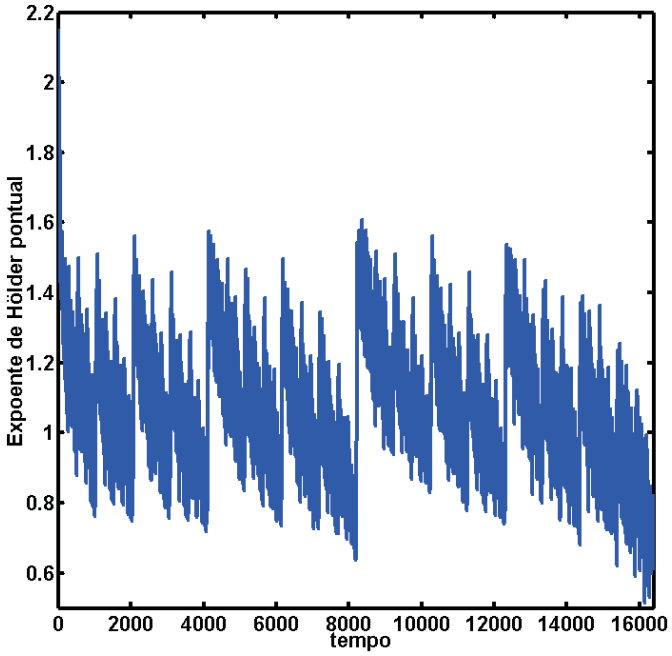
As cascatas multiplicativas são processos não-estacionários e foram inicialmente introduzidas no contexto de redes para modelar processos de taxa de tráfego em escalas pequenas de tempo, menores do que algumas centenas de milissegundos (Feldmann et al., 1998). A presença desse comportamento multifractal foi atribuída aos mecanismos de rede operando nestas escalas de tempo. Infelizmente o processo de construção da cascata requer um número elevado de parâmetros, tipicamente da ordem de $O(\log_2 \breve{N})$, onde $\breve{N}$ é o tamanho da série de tráfego. Este problema será abordado no Capítulo 3, onde as propriedades das *wavelets* são exploradas para se obter um modelo mais flexível que possa ser atualizado à medida que dados são disponibilizados. Enquanto isso, apresenta-se na próxima seção o algoritmo para estimação dos multiplicadores proposto por Feldman et. al (Feldmann et al., 1998) e uma extensão sugerida com a qual se pode obter funções analíticas para as densidades de probabilidade dos multiplicadores.

2.5.3 Estimação da Densidade de Probabilidade dos Multiplicadores

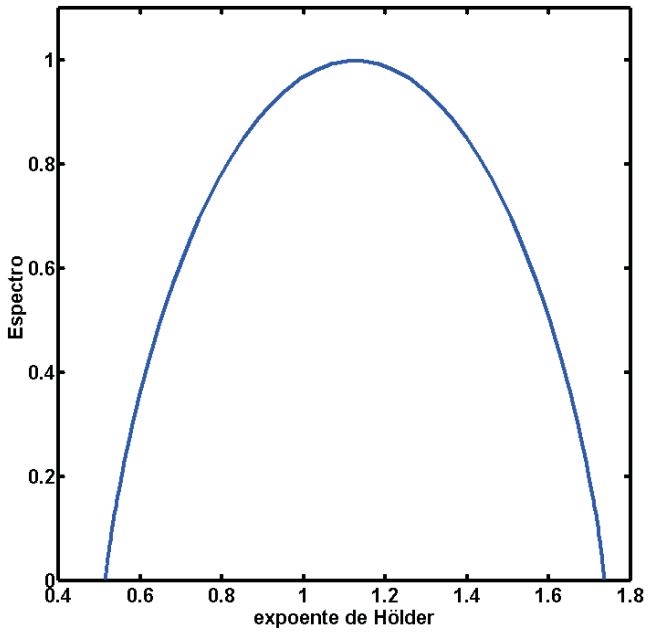
Sejam os dados de tráfego no estágio *N* da cascata, X_i^N. A série de tráfego no estágio $(N-1)$ da cascata pode ser obtida agregando valores consecutivos do estágio posterior *N* em blocos não-sobrepostos de tamanho 2. De forma análoga, dada a série na escala $(N-j)$, $X_i^{N-j}(i = 1,\ldots,2^{N-j})$,



(a)



(b)



(c)

Fig. 2.13: a) Cascata binomial determinística com $m_0 = 0,3$ e $N = 14$. b) Expoentes de Hölder pontuais estimados para a cascata. c) Espectro multifractal da cascata.

obtemos os dados na escala $(N-j-1)$ pela soma consecutiva dos valores do estágio $(N-j)$ da seguinte forma:

$$X_i^{N-j-1} = X_{2i-1}^{N-j} + X_{2i}^{N-j} \tag{2.59}$$

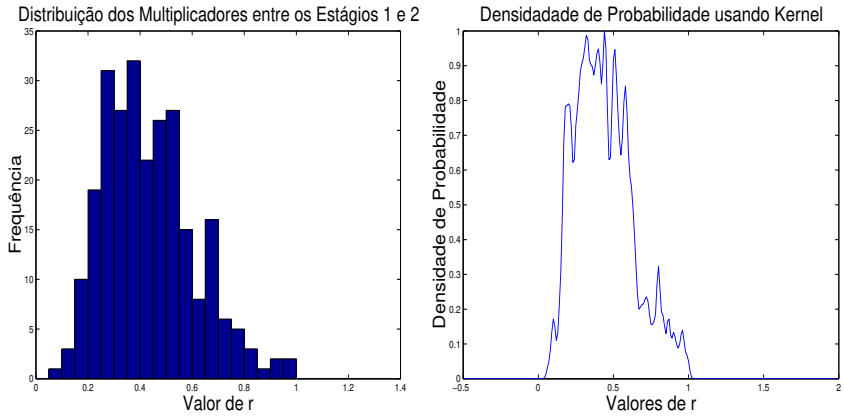
para $i = 1, \ldots, 2^{N-j-1}$. Este procedimento termina quando a agregação dos valores forma apenas um ponto na última escala da cascata. Uma estimativa $r_j^{(i)}$ dos multiplicadores pode ser obtida pela seguinte equação (Feldmann et al., 1998):

$$r_j^{(i)} = \frac{X_i^{N-j}}{X_{2i-1}^{N-j-1}} \tag{2.60}$$

para $i = 1, \ldots, 2^{N-j-1}$. Podemos considerar $r_j^{(i)}$ como amostras da distribuição de multiplicadores $f_{R_j}(r)$ no estágio *j*. A distribuição dos multiplicadores na escala *j* pode ser obtida pelos histogramas de $r_j^{(i)}$. O modelo multiplicador Gaussiano de variância variável (*Variable Variance Gaussian Multiplier*, VVGM) por exemplo, é uma cascata multiplicativa que aproxima os histogramas obtidos por gaussianas (Krishna et al., 2003). A distribuição dos multiplicadores $f_{R_j}(r)$ neste modelo é gaussiana centrada em 0.5 com variâncias que mudam a cada escala. Essas variâncias são estimadas a partir dos histogramas para processos de tempo de chegada de pacote. Na Figura 2.14 são mostrados o histograma obtido usando o método descrito nesta seção e a densidade de probabilidade para os multiplicadores entre os estágios 1 e 2 utilizando o método de Kernel (ver Apêndice C). Neste caso, o processo em questão é o tráfego de chegadas de bytes (série dec-pkt-1). Nota-se que supor a distribuição dos multiplicadores como sendo gaussiana pode não ser realista. De fato, como pode ser visto neste caso, em que a distribuição dos multiplicadores tende a ser lognormal. Levando em consideração esta característica, apresentamos em (Vieira & Ling, 2006b) um modelo multifractal que usa o método de kernel para estimar a distribuição dos multiplicadores da cascata entre os estágios, resultando em um ganho de eficiência de modelagem sobre o modelo VVGM.

2.5.4 Modelo Wavelet Multifractal (MWM)

O modelo *wavelet* multifractal (MWM, Multifractal Wavelet Model) proposto por Rudolf Riedi et al. é um dos principais e mais conhecidos modelos multifractais aplicados à caracterização de tráfego de rede (Riedi et al., 1999). O MWM se baseia em uma cascata multiplicativa no domínio *wavelet* onde a transformada *wavelet* discreta (DWT) é usada dada sua capacidade de representação multiescala de sinais (Daubechies, 1992). Para se gerar um processo segundo o modelo MWM é preciso aplicar a transformada *wavelet* discreta de Haar ao sinal de tráfego de rede, calcular os momentos de segunda ordem dos coeficientes *wavelet* em cada escala, a média e a variância dos coeficientes na escala de maior resolução e calcular p_j, a variável usada para capturar o decaimento de energia dos



(a) (b)

Fig. 2.14: Distribuição dos Multiplicadores entre os estágios 1 e 2 para o traço de tráfego dec-pkt-1

coeficientes *wavelet* com a escala j. O MWM aproxima com eficiência as propriedades do fluxo de tráfego original em termos da distribuição marginal (de fato, produz distruibuição aproximandamente lognormal) e sua estrutura de correlação, tendo grande destaque na modelagem de tráfego (Ribeiro et al., 2000)(Riedi et al., 1999).

Para se compreender o modelo MWM, devemos novamente citar alguns conceitos da transformada *wavelet*. Conforme pode ser verificado no Apêndice A, a transformada *wavelet* discreta é usada para representação multiescala de sinais da seguinte forma:

$$f(t) = \sum_{k} U_{J_0,k}\phi_{J_0,k}(t) + \sum_{j=J_0}^{\infty}\sum_{k} W_{j,k}\varphi_{j,k}(t) \tag{2.61}$$

onde $W_{j,k}$ e $U_{j,k}$ são respectivamente os coeficientes *wavelet* e de escala, dados por:

$$W_{j,k} = \int f(t)\varphi_{j,k}(t)dt \tag{2.62}$$

e

$$U_{j,k} = \int f(t)\phi_{j,k}(t)dt \tag{2.63}$$

A Figura 2.15 exibe as funções escala $\phi_{j,k}(t)$ e *wavelet* de Haar $\varphi_{j,k}(t)$ utilizadas na representação multiescala do sinal. Pode-se demonstrar que os coeficientes de escala dados pela equação (2.63) podem ser recursivamente calculados utilizando a *wavelet* de Haar $\varphi_{j,k}(t)$ através das seguintes equações:

$$U_{j,2k} = 2^{-1/2}(U_{j-1,k} + W_{j-1,k}) \tag{2.64}$$

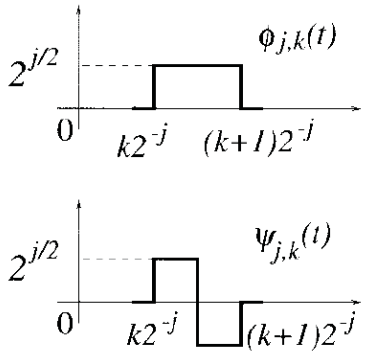


Fig. 2.15: Funções escala $\phi_{j,k}(t)$ e *wavelet* de Haar $\varphi_{j,k}(t)$

$$U_{j,2k+1} = 2^{-1/2}(U_{j-1,k} - W_{j-1,k}) \tag{2.65}$$

Este processo recursivo é repetido até que se atinja a resolução desejada ou equivalentemente, até que se obtenha o número desejado de amostras, formando uma árvore binária de coeficientes de escala. A Figura 2.16 apresenta a árvore binária formada para obtenção dos coeficientes de escala representada pelas equações (2.64) e (2.65). No modelo MWM, a fim de assegurar a não-negatividade do traço de tráfego sintético, determinadas condições devem ser impostas a seus coeficientes *wavelet* e de escala. Os coeficientes $U_{j,k}$ representam a média local do processo em escalas e deslocamentos de tempo diferentes. A condição $X(t) \geq 0$, $\forall t$, impõe que, $U_{j,2k+1} \geq 0$, $\forall j, k$. Impondo a condição $U_{j,k} \geq 0$, $\forall j, k$, pode-se afirmar que $|W_{j,k}| \leq U_{j,k}$, $\forall j, k$. Os coeficientes *wavelet* são gerados a partir da equação:

$$W_{j,k} = U_{j,k}A_{j,k} \tag{2.66}$$

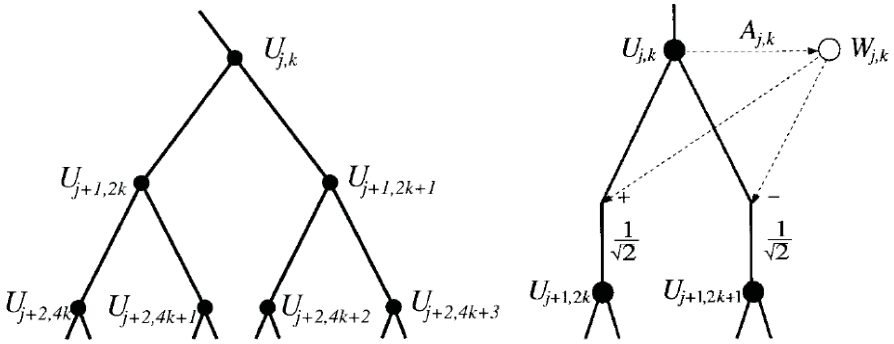
onde $A_{j,k}$ é uma variável aleatória cujo valor está em [-1,1]. Além disso, supõe-se algumas condições para esta variável: os multiplicadores $A_{j,k}$ são independentes e identicamente distribuidos (i.i.d) dentro de cada escala, são também independentes de $U_{j,k}$ e simétricos em torno de zero. A Figura 2.17 mostra a função simétrica Beta $\beta(p_j, p_j)$, que é usada para modelar a distribuição $g_A(a)$ dos multiplicadores. Os multiplicadores $A_{j,k}$ são escolhidos de forma a controlar as energias dos coeficientes *wavelet*. Assim, obtém-se as seguintes relações (Riedi et al., 1999):

$$\frac{E(W_{j-1,k}^2)}{E(W_{j,k}^2)} = \frac{2p_j + 1}{p_{j-1} + 1} \tag{2.67}$$

e

$$(2p_0 + 1)E(W_{0,0}^2) = E(U_{0,0}^2) \tag{2.68}$$

Pode-se notar que p_j é usado para capturar o decaimento da energia dos coeficientes *wavelet* em



(a) (b)

Fig. 2.16: Arvore binária dos coeficientes de escala

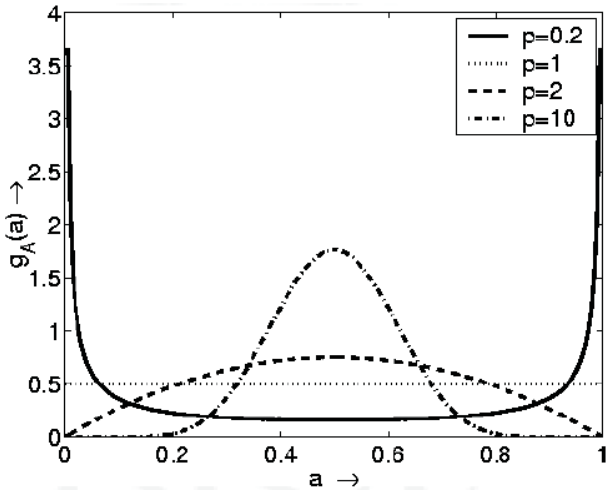


Fig. 2.17: Função beta $\beta(p;p)$: função de distribuição de probabilidade $g_A(a)$ da variável aleatória A (multiplicadores da cascata) para diferentes valores de p.

escala. O coeficiente de escala $U_{0,0}$ na maior escala (mais fina) é modelado como sendo uma variável aleatória normal com média e a variância iguais aos dos coeficientes de escala dos traços de tráfego reais.

O deslocamento k_j dos coeficientes de escala é relacionado ao deslocamento de um de seus dois descendentes diretos como segue (Riedi et al., 1999):

$$k_j + 1 = 2k_j + k'_j \tag{2.69}$$

onde $k'_j = 0$ corresponde ao descendente esquerdo e $k'_j = 1$ ao descendente direito. Dessa forma, os

coeficientes *wavelet* e de escala podem ser expressos respectivamente como:

$$U_{j,k_j} = 2^{-\frac{j}{2}} U_{0,0} \prod_{i=0}^{j-1} [1 + (-1)^{k'_i} A_{i,k_i}] \tag{2.70}$$

e

$$W_{j,k_j} = 2^{-\frac{j}{2}} U_{0,0} \prod_{i=0}^{j-1} [1 + (-1)^{k'_i} A_{i,k_i}] A_{i,k_i} \tag{2.71}$$

Conseqüentemente, considerando um tempo discreto k, tem-se que o processo discreto de tráfego MWM $X[k]$ é obtido pelos coeficientes de escala $U_{j,k}$ na escala mais fina j, da seguinte forma:

$$X[k] = 2^{-\frac{j}{2}} U_{j,k} \tag{2.72}$$

O MWM pode precisamente modelar a dependência de longa duração presente nos dados de tráfego assim como capturar outras características multifractais. Entretanto, uma desvantagem do MWM é o número de parâmetros a serem estimados, que para isso, faz uso de toda série de tráfego. Dessa forma, o MWM se torna inadequado a aplicações em tempo real, uma vez que o custo total para calcular N amostras de um processo MWM é $O(N)$. No próximo capítulo introduzimos um modelo multifractal com poucos parâmetros de entrada, os quais podem ser atualizados adaptativamente.

Capítulo 3

Modelo Multifractal baseado em Wavelets (MMW)

3.1 Introdução

Modelos de tráfego precisos capturam importantes características do tráfego, melhorando sua compreensão e permitindo o estudo dos efeitos dos parâmetros do modelo no desempenho das redes. A constatação da presença da auto-similaridade no tráfego de redes (Leland et al., 1994) foi seguida imediatamente pelo desenvolvimento de modelos fractais de tráfego. O movimento Browniano fracionário (fBm) e seu processo de incrementos, denominado ruído Gaussiano fracionário (fGn), tornaram-se os modelos auto-similares mais utilizados. As séries de tráfego de redes exibem auto-similaridade e dependência de longa duração, mas também correlações de curta duração e comportamento multi-escala incoerentes com modelos auto-similares.

Processos auto-similares apresentam um comportamento em escala único e global, ou seja, apresentam escalonamento monofractal representado pelo parâmetro de Hurst. Processos multifractais são uma generalização dos processos monofractais, possibilitando regularidade e leis de comportamento em escala variantes no tempo e, portanto, proporcionando uma melhor descrição de processos irregulares.

O presente capítulo propõe um modelo multifractal no domínio *wavelet* que explora algumas propriedades das *wavelets*, em especial as da *wavelet* de Haar. São derivados parâmetros do modelo proposto tais como expressões para seu parâmetro de escala global e para sua função de autocorrelação. Para validar o modelo, são apresentados os vários testes realizados, comprovando sua eficiência. O modelo proposto visa capturar as propriedades multifractais representadas pelas funções de escala e pelo fator de momento de séries reais de tráfego, possuindo uma estrutura adequada para modelagem adaptativa de sinais. Com objetivo de proporcionar uma característica de atualização em tempo real

ao modelo proposto, um algoritmo para estimação dos parâmetros multifractais também é proposto. Esta capacidade de adaptação dinâmica do modelo será empregada apenas no Capítulo 5 em uma proposta de cálculo adaptativo de banda efetiva.

3.2 Modelo Multifractal baseado em Wavelets (MMW)

Nesta seção, é proposto um novo modelo de tráfego, que também pode ser visto como uma cascata multifractal tal como o MWM, mas que requer menos parâmetros em seu processo de síntese. Este modelo multifractal denominado de Modelo Multifractal baseado em Wavelets (MMW) é obtido usando apenas a variância do traço de tráfego agregado e a média e a variância dos dados de tráfego. O modelo MMW simplifica o processo de síntese do modelo MWM apresentado no Capítulo 2, ao incorporar o conhecimento dos parâmetros das funções de escala $\tau(q)$ e do fator de momento $c(q)$.

Sabe-se que as propriedades multifractais dos dados reais de tráfego são caracterizadas por suas correspondentes função de escala $\tau(q)$ e fator de momento $c(q)$ descritas pela definição 2.3.1 do Capítulo 2 (Dang. et al., 2002). Então, um modelo multifractal geral deve capturar estas duas propriedades multifractais (Molnár et al., 2002). O modelo proposto objetiva capturar tanto a função de escala como o fator de momento do processo a ser analisado. Segundo os autores em (Dang. et al., 2002), isto pode ser obtido pelo produto de uma cascata e uma variável aleatória i.i.d. positiva *Y*. Dessa forma, o modelo multifractal resultante pode ser visto como o produto da taxa de pico do fluxo *Y*, pela medida de rajada $\mu(\Delta t_N)$ na escala de tempo aplicada Δt_N. A variável *Y* é independente da medida da cascata $\mu(\Delta t_k)$, então a série obtida denotada por $X(\Delta t_N)$ satisfaz a seguinte relação:

$$E(X(\Delta t_N)^q) = E(Y^q)E(\mu(\Delta t_N)^q) = E(Y^q)\Delta t_N^{\tau_0(q)} \tag{3.1}$$

Comparando a equação (3.1) com a equação de definição de processos multifractais (2.16), pode-se notar que as variáveis *R* e *Y* devem ser tais que:

$$\begin{cases} -log_2(E(R^q)) &= \tau_0(q) \\ E(Y^q) &= c(q) \end{cases} \tag{3.2}$$

A medida $\mu(\Delta t_N)$ tem um valor pequeno uma vez que ela é o produto de *N* multiplicadores ($\mu(\Delta t_N) = R(\eta_1)R(\eta_1\eta_2)...R(\eta_1, ..., \eta_N)$), onde $0 < R(\eta_1, ..., \eta_i) < 1$ e i indica o estágio da cascata. Assim, por conveniência, e sem perda de generalidade, multiplicamos os valores de $\mu(\Delta t_N)$ da cascata por 2^N. Uma vez que $E(\mu(\Delta t_N)) = 2^{-N}$, esta multiplicação normaliza o processo, fazendo com que o mesmo tenha média 1. Outro artifício usado é considerar a unidade de intervalo de tempo como unitária (denotada por Δt_0) no estágio *N* da cascata ao invés de $\Delta t_N = 2^{-N}$. Após estas

modificações, tem-se (Dang. et al., 2002):

$$E(\mu(\Delta t_N)^q) = E(Y^q)2^{N(q+log_2E(R^q))}\Delta t_0^{-log_2E(R^q)} \tag{3.3}$$

As variáveis *R* e *Y* agora devem ser escolhidas de forma a atender ao seguinte sistema de equações:

$$\begin{cases} -log_2(E(R^q)) &= \tau_0(q) \\ logE(Y^q) &= logc(q) - (q + log_2(E(R^q))Nlog2 \end{cases} \tag{3.4}$$

A função de escala $\tau(q)$ pode ser precisamente modelada assumindo que *R* é uma variável aleatória em [0,1] com distribuição beta simétrica Beta(α,α) com $\alpha > 0$ (Molnár et al., 2002). Assim, a função $\tau_0(q) := \tau(q) + 1$ relacionada à função de escala $\tau(q)$, pode ser explicitamente escrita como

$$\tau_0(q) = log_2\frac{\Gamma(\alpha)\Gamma(2\alpha + q)}{\Gamma(2\alpha)\Gamma(\alpha + q)}, \tag{3.5}$$

onde $\Gamma(.)$ corresponde à função Gama.

É necessário também se estabelecer uma distribuição para a variável Y. Considerou-se a variável aleatória *Y* como tendo uma distribuição lognormal definida pelos parâmetros ρ e γ e portanto, possuindo momento $E(Y^q) = e^{\rho q+\gamma^2q^2/2}$. Assim, segundo o sistema de equações (3.4), as variáveis ρ e γ devem obedecer a seguinte equação:

$$\rho q + \gamma^2q^2/2 = logc(q) - (q + log_2E(R^q))Nlog2 \tag{3.6}$$

A equação (3.6) permite expressar analiticamente o fator de momento pela seguinte relação:

$$c(q) = e^{\rho q+\gamma^2q^2/2}2^{N\left(q-log_2\frac{\Gamma(\alpha)\Gamma(2\alpha+q)}{\Gamma(2\alpha)\Gamma(\alpha+q)}\right)} \tag{3.7}$$

Dessa forma, analisando-se as equações (3.5) e (3.7), verifica-se que o modelo multifractal proposto é caracterizado por três parâmetros (α,ρ,γ) provindos das expressões analíticas para a função de escala $\tau(q)$ e o fator de momento $c(q)$.

Utilizando propriedades estatísticas das cascatas multiplicativas (Gao & Rubin, 2001)(Waymire & Williams, 1995), obtém-se os seguintes resultados com relação à média e à variância do modelo MMW:

(i) A média do processo MMW $X(\Delta t_0)$, onde Δt_0 denota o intervalo de tempo unitário da série a ser modelada, é dada por:

$$E(X(\Delta t_0)) = e^{\rho+\gamma^2/2} \tag{3.8}$$

(ii) A variância do processo MMW é dada por:

$$var(X(\Delta t_0)) = e^{2\rho+2\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right)^N - e^{2\rho+\gamma^2}. \tag{3.9}$$

Assim como as séries de tráfego real, o processo MMW $X(\Delta t_0)$ tem distribuição lognormal para $N \gg 1$ grande. Esta propriedade pode ser deduzida se fazendo uso do Teorema Central do Limite uma vez que (Papoulis, 1991):

$$X(\Delta t_0) = 2^N Y R(\eta_1)R(\eta_1\eta_2)...R(\eta_1, ..., \eta_N), \tag{3.10}$$

ou seja, o MMW é um processo multiplicativo. Segundo o Teorema Central do Limite um processo que tenha essa característica tende a ter uma distribuição lognormal quanto mais se aumenta o número de estágios da cascata N.

Denotaremos simplesmente por $X(k)$ o processo correspondente ao MMW (equação (3.10)) onde $k = \Delta t_0 \in \mathbb{N}$. Os seguintes lemas serão empregados no estabelecimento do processo de síntese do MMW.

Lema 3.2.1 *Seja $X(k)$ um processo multifractal com média e variância dadas respectivamente pelas equações (3.8) e (3.9) e o processo agregado X^m de $X(k)$ definido como*

$$X^m = \frac{1}{m}\sum_{(k-1)m+1}^{km} X(k). \tag{3.11}$$

A variância do processo agregado $var[X^m]$ onde $m = 2^k$, é determinada pela seguinte equação:

$$var[X^m] = e^{2\rho+2\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right)^{N-k} - (e^{2\rho+\gamma^2}2^{2k-2N}) \tag{3.12}$$

Demonstração Sejam $m = 2^k$ e $x_{(N-k)} = \sum_{(k-1)m+1}^{km} X(k)$. Então, o processo agregado X^m no estágio $(N-k)$ pode ser escrito da seguinte forma:

$$X^m = 2^{-k}x_{(N-k)} \tag{3.13}$$

A variância $var[X^{(m)}]$ do processo agregado X^m do processo MMW $X(k)$ pode ser escrita como:

$var[X^{(m)}] = var(2^{-k}x_{N-k})$

$$var[X^m] = E[\{2^{-k}x_{(N-k)}\}^2] - E^2[2^{-k}x_{(N-k)}] \tag{3.14}$$

Como o processo $X(k)$ corresponde ao produto de uma variável aleatória lognormal Y e uma cascata multiplicativa μ, a seguinte relação é válida:

$$x_{(N-k)} = \sum_{(k-1)m+1}^{km} Y_k \mu(\Delta t_k) \tag{3.15}$$

Os momentos $M(q)$ do processo agregado para cascatas multiplicativas são dados pela equação (2.57) do Capítulo 2. Utilizando esta equação e a relação (3.15), os momentos do processo agregado MMW podem ser expressos por:

$$M(q) = \frac{1}{m}\sum_{i=1}^{m} X(i)^q = \frac{1}{m}\sum_{i=1}^{m} (Y_i \mu(\Delta t_i))^q \tag{3.16}$$

Então, a variância do processo agregado de $X(t)$ se torna:

$$\begin{aligned} var[X^m] &= E[\{2^{-k}\frac{1}{m}\sum_{(k-1)m+1}^{km} Y_k\mu(\Delta t_k)\}^2] - E^2[2^{-k}\frac{1}{m}\sum_{(k-1)m+1}^{km} Y_k\mu(\Delta t_k)] \\ &= E[Y^2]2^{2N}E[R^2]^{N-k} - E^2[Y]\left(\frac{1}{2}\right)^{2(N-k)} \end{aligned}$$

$$var[X^m] = e^{2\rho+2\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right)^{N-k} - (e^{2\rho+\gamma^2}2^{2k-2N}) \tag{3.17}$$

■

Lema 3.2.2 *Supondo que a wavelet de Haar é usada no cálculo dos coeficientes wavelet de um processo $X(t)$ cuja média e variância são dadas respectivamente por (3.8) e (3.9), então o momento de segunda ordem dos coeficientes wavelets pode ser escrito como:*

$$E(w_j^2) = e^{4\rho+4\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right)^2 - 2Z_j \tag{3.18}$$

onde

$$Z_j = 2^{j-1}\left[var[X^{(2^j)}] - e^{2\rho+2\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right) + (e^{2\rho+\gamma^2})\right] \tag{3.19}$$

Além disso, a média e a variância dos coeficientes de escala na escala $j = n$ podem ser expressos respectivamente pelas seguintes equações:

$$E\{U_{n,k}\} = 2^{n/2}(e^{\rho+\gamma^2/2}) \tag{3.20}$$

$$\sigma^2_{U_{n,k}} = e^{4\rho+4\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right)^2 + 2Z_n - (2^n e^{2\rho+\gamma^2}) \tag{3.21}$$

Demonstração Seja $X^m \stackrel{d}{=} \frac{1}{m}\sum_{i=1}^{m} X(i)$ o processo agregado de ordem *m* de $X(i)$, onde $\stackrel{d}{=}$ representa igualdade em distribuição. Para $m = 2$, a variância do processo agregado $X^{(2)}$ pode ser escrita como

$$\sigma^2_{X^{(2)}} = \frac{\sum_{i=0,2,\ldots}^{N-2}(\frac{X(i)+X(i+1)}{2} - \overline{X})^2}{N/2}, \tag{3.22}$$

onde $\overline{X}$ e *N* são a média e o tamanho do processo de tráfego. De uma forma mais geral, seja $m = 2^j$, a variância do processo agregado de ordem 2^j é:

$$\sigma^2_{X^{(2^j)}} = \frac{\frac{1}{2^j}\sum_{i=0,1,\ldots}^{N-1} X(i)^2}{N} + \frac{1}{2^{j-1}}Z_j - \overline{X}^2 \tag{3.23}$$

onde

$$Z_j = \sum_{i=0,2,\ldots}^{N-2} \frac{X(i)+X(i+j)}{N} \tag{3.24}$$

Usando o fato de que:

$$\frac{\frac{1}{2^j}\sum_{i=0,1,\ldots}^{N-1} X(i)^2}{N} = e^{2\rho+2\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right) \tag{3.25}$$

e substituindo a equação (3.8) em (3.23), o termo Z_j pode ser expresso de forma alternativa como:

$$Z_j = 2^{j-1}\left[\sigma^2_{X^{(2^j)}} - e^{2\rho+2\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right) + (e^{2\rho+\gamma^2})\right] \tag{3.26}$$

De acordo com (Vijayan et al., 2002), o momento de segunda ordem dos coeficientes *wavelet* obtidos com a transformada *wavelet* discreta de Haar do processo de tráfego pode ser determinado pela seguinte equação:

$$E(w_j^2) = \frac{\sum_{i=0,1,\ldots}^{N-1} X(i)^2}{N} - 2\sum_{i=0,2,\ldots}^{N-2} \frac{X(i)+X(i+j)}{N} \tag{3.27}$$

Então a equação (3.27) pode ser reescrita como

$$E(w_j^2) = e^{4\rho+4\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right)^2 - 2Z_j, \tag{3.28}$$

onde Z_j é dado pela equação (3.26).

Os coeficientes de escala $U_{n,k}$ na escala $j = n$ são calculados da seguinte maneira (Riedi & Véhel, 1997):

$$U_{n,k} = 2^{-n/2} \sum_{i=2^n k}^{2^n(k+1)-1} X(i) \qquad k = 0, 1, ..., N/(2^n) - 1 \tag{3.29}$$

Então, a média dos coeficientes de escala na escala $j = n$ do processo MMW é dada pela seguinte equação:

$$E\{U_{n,k}\} = 2^{-n/2} \frac{\sum_{i=0}^{N-1} X(i)}{N/(2^n)} = 2^{n/2} E\{X(i)\} \tag{3.30}$$

$$E\{U_{n,k}\} = 2^{n/2}(e^{\rho+\gamma^2/2}) \tag{3.31}$$

Analogamente, a variância dos coeficientes em escala na escala mais grosseira $j = n$ do processo MMW pode ser expressa como:

$$\sigma^2_{U_{n,k}} = \frac{\sum_{i=0,1,...}^{N-1} X(i)^2}{N} + 2Z_n - (E\{U_{n,k}\})^2 \tag{3.32}$$

$$\sigma^2_{U_{n,k}} = e^{4\rho+4\gamma^2} \left(\frac{\alpha+1}{\alpha+1/2} \right)^2 + 2Z_n - (2^n e^{2\rho+\gamma^2}) \tag{3.33}$$

Portanto, inserindo o termo Z_n dado pela equação (3.26) na equação (3.33), obtém-se a expressão desejada para a variância dos coeficientes *wavelet*.

■

Os Lemas 3.2.1 e 3.2.2 enunciam que o momento de segunda ordem dos coeficientes *wavelet* pode ser estimado pelo conhecimento dos parâmetros α, γ, e ρ através das equações (3.12), (3.18) e (3.19). No modelo proposto, assim como é feito no MWM, esses momentos de segunda ordem são usados para calcular as variáveis p_j cuja função é capturar o decaimento de energia dos coeficientes *wavelet* em escala. A equação (2.67) do Capítulo 2 resume o referido procedimento:

$$\frac{E(w^2_{j-1,k})}{E(w^2_{j,k})} = \frac{2p_j + 1}{p_{j-1} + 1} \tag{3.34}$$

Calculado os valores de p_j, o restante do procedimento de geração de amostras do modelo MMW é similar ao do MWM. Assim, as entradas para o algoritmo de síntese do processo MMW são os parâmetros α, γ, e ρ calculados com os valores das funções $\tau(q)$ e $c(q)$ usando o método descrito em (Molnár et al., 2002). A vantagem desse novo modelo multifractal é que, ao invés de se aplicar a DWT em todo a série de tráfego como é feito pelo MWM, é suficiente o conhecimento do conjunto de parâmetros (α, γ, ρ) para a síntese do processo MMW. A implementação completa de um procedimento de modelagem adaptativa utilizando estimação 'on-line' dos parâmetros (α, γ, ρ) é possível graças

ao algoritmo proposto na seção 3.2.3. Supondo conhecidos os parâmetros (α, γ, ρ), o procedimento para geração de tráfego sintético pelo MMW é descrito abaixo:

Algoritmo 3.2.3 *Algoritmo Proposto de Síntese de Tráfego MMW*

Passo 1) Configuração inicial: Conjunto de parâmetros de entrada (α, γ, ρ) para uma série de tráfego de tamanho 2^N amostras;

Passo 2) Inicie com a escala $j = 1$, correspondendo à fração de agregação igual a $2^j = 2$;

Passo 3) Calcule a variância do processo agregado $var[X^m]$ onde $m = 2^j$ por meio da equação (3.12):

$$var[X^{2^j}] = e^{2\rho+2\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right)^{N-j} - (e^{2\rho+\gamma^2}2^{2j-2N})$$

Passo 4) Calcule Z_j usando a variância do processo agregado $var[X^m]$ calculada no passo anterior e a equação (3.19):

$$Z_j = 2^{j-1}\left[var[X^{2^j}] - e^{2\rho+2\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right) + (e^{2\rho+\gamma^2})\right]$$

Passo 5) Estime o momento de segunda ordem dos coeficientes wavelet *$E(w_j^2)$ do processo $X(t)$ via equação (3.18) e pelo valor calculado de Z_j do passo anterior:*

$$E(w_j^2) = e^{4\rho+4\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right)^2 - 2Z_j;$$

Passo 6) Incremente j de 1. Se $j = N$, calcule a média e a variância dos coeficientes em escala na escala mais grosseira usando as equações (3.20) e (3.21); do contrário, vá para o passo 3;

Passo 7) Calcule os valores de p_j usando a equação (3.34);

Passo 8) Aqui começa um procedimento recursivo tendo obtido os valores de $p_j's$. Mude $j = 0$ e calcule os coeficientes de escala na escala mais grosseira $U_{0,0}$;

Passo 9) Gere amostras aleatórias correspondentes a $A_{j,k}$ para $k = 0, 1, \ldots (2^j) - 1$ e calcule os coeficientes wavelet *na escala j $(W_{j,k})$ usando (2.66);*

Passo 10) Na escala j, calcule os coeficientes de escala na escala $j + 1$, $(U_{j+1,2k})$ e $(U_{j+1,2k+1})$ para $k = 0, 1, \ldots, (2^j) - 1$ via equações (2.64) e (2.65);

Passo 11) Incremente j de 1. Se $j < N$, repita os passos (9) e (10).

Passo 12) A série de tráfego gerada é $X(k) = 2^{-j/2}U_{j,k}$, $k = 0, 1, \ldots, 2^N - 1$.

3.2.1 Parâmetro de Escala Global para Tráfego Multifractal

Esta seção é dedicada à derivação de um fator de escala global (definido anteriormente em (Krishna et al., 2001)) associado ao modelo multifractal MMW. A vantagem de se ter um parâmetro de escala global é que é possível se chegar a um modelo de fila aproximado para processos multifractais. Isso proporciona por exemplo, cálculo de probabilidade de perda. Neste sentido, destacam-se os trabalhos de Vieira et al. (Vieira & Ling, 2006a)(Vieira & Ling, 2006b) (Vieira & Ling, 2006c). O desenvolvimento desta teoria requer o uso de alguns resultados oriundos do estudo de processos auto-similares (Norros, 1994) (Beran, 1995).

Seja $X(t)$ um processo auto-similar com parâmetro de Hurst H, com média zero e variância σ^2 tal que a seguinte relação é válida:

$$X \stackrel{d}{=} m^{1-H} X^m \tag{3.35}$$

Denotemos por $Y(t) = X(t) - X(t-1)$, o processo incremento correspondente a $X(t)$. A dependência de longo prazo do processo de incremento $Y(t)$ pode ser obtida da análise da covariância deste processo. Uma vez que o processo agregado de $Y(t)$ tem característica auto-similar, estatísticas dos dados agregados podem ser obtidas da seguinte maneira:

$$\begin{aligned} Y^m &\stackrel{d}{=} \frac{1}{m}\sum_{i=1}^{m} Y(i) \stackrel{d}{=} \frac{1}{m}\{Y(m) + ... + Y(1)\} \\ &\stackrel{d}{=} \frac{1}{m}\{X(m) - X(m-1) + ... + (X(1) - X(0))\} \\ &\stackrel{d}{=} \frac{1}{m}\{X(m) - X(0)\} \stackrel{d}{=} m^{H-1}\{X(1) - X(0)\} \end{aligned}$$

Logo

$$Y^m \stackrel{d}{=} m^{H-1} Y(1) \tag{3.36}$$

A média do processo agregado Y^m é zero e a variância é dada por

$$var[Y^m] = E[(m^{H-1}Y(1))^2] = m^{2H-2}\sigma^2. \tag{3.37}$$

A equação (3.37) relaciona a mudança da variância do processo agregado Y^m em função do parâmetro de agregação *m*. Aplicando logaritmo em ambos os lados da equação (3.37), obtém-se:

$$\log_2\{var[Y^m]\} = (2H-2)\log_2 m + \log_2 \sigma^2 \tag{3.38}$$

Note que o parâmetro de Hurst *H* aparece na equação (3.38). Pretende-se agora obter uma relação

similar à equação (3.38) com intuito de derivar o parâmetro de escala global H_g para o modelo MMW. Este parâmetro é enunciado pela Proposição a seguir.

Proposição 3.2.4 *Seja o modelo multifractal proposto MMW com parâmetros α, m e γ. O parâmetro de escala global para este modelo é dado por*

$$H_g = 1 - \frac{\log_2(\frac{\alpha+1}{\alpha+1/2})}{2}. \tag{3.39}$$

Demonstração A equação (3.12) descreve o comportamento da variância $var[X^m]$ do processo agregado em função do parâmetro de agregação m para o processo multifractal MMW $X(k)$ descrito no Lema 3.2.1. Quando o número de estágios *N* na geração da cascata é grande, o termo $(e^{2\rho+\gamma^2}2^{2k-2N})$ pode ser seguramente ignorado. Aplicando logaritmo na equação (3.12), obtemos:

$$\log_2 var[X^m] =$$

$$= \log_2 e^{2\rho+2\gamma^2} + \log_2 \left(\frac{\alpha+1}{\alpha+1/2}\right)^N + \log_2 \left(\frac{\alpha+1}{\alpha+1/2}\right)^{-\log_2 m} \tag{3.40}$$

Tendo em mãos a equação (3.40), pode-se definir um parâmetro de escala global H_g para o tráfego multifractal análogo ao parâmetro de Hurst H no caso monofractal. Comparando as equações (3.38) e (3.40), pode-se observar a correspondência entre os termos envolvidos. Assim, o parâmetro de escala global H_g é dado pela seguinte equação:

$$H_g = 1 - \frac{\log_2(\frac{\alpha+1}{\alpha+1/2})}{2} \tag{3.41}$$

■

Como conclusão, pode-se afirmar que para processos multifractais como o modelo de tráfego MMW, existe um parâmetro de escala global similar ao parâmetro de Hurst para o caso monofractal, apesar de possuirem diferentes expoentes de Hölder locais (Riedi et al., 1999).

3.2.2 Função de Autocorrelação do Modelo Multifractal MMW

Para se descrever um processo por completo é necessário o conhecimento de sua função de distribuição acumulada conjunta (Papoulis, 1991). Entretanto, esta é muitas vezes, difícil de ser obtida. A função de autocorrelação provê informação parcial mas que é mais fácil de ser obtida. Muitos teoremas sobre existência de continuidade, de integrabilidade, diferenciabilidade e ergodicidade dependem da função de autocorrelação (Lee & Fapojuwo, 2005). Seja um processo em tempo discreto

$X(n)$ e os instantes de tempo n e k, a função de autocorrelação $cor[X(n), X(n+k)]$ para este processo é definida como

$$cor[X(n), X(n+k)] = E[X(n)X(n+k)]. \tag{3.42}$$

Pode-se constatar pela função de autocorrelação de um processo, a presença ou não de dependência de longo prazo. Além disso, a função de autocorrelação reflete a estatística de segunda ordem de uma série temporal. Esta seção trata da obtenção de uma expressão para a função de autocorrelação para o MMW. A partir de algumas propriedades deste modelo multifractal, sua função de autocorrelação pode ser obtida de forma analítica.

Teorema 3.2.5 *Seja o processo multifractal MMW $X(\Delta t_0)$ com parâmetros α, ρ, γ e N estágios, onde Δt_0 denota o intervalo de tempo unitário de modelagem dos dados. A função de autocorrelação deste processo é dada pela seguinte equação:*

$$E[X(\Delta t_0)_n X(\Delta t_0)_{n+k}] = e^{2\rho+\gamma^2}\Big(\frac{\alpha(\alpha+1)^{N-1}}{(\alpha+1/2)^N} k^{-\log_2(\frac{\alpha+1}{\alpha+1/2})}\Big) \tag{3.43}$$

Demonstração Primeiramente, vamos reescrever a função de autocorrelação (3.42) para o processo $X(\Delta t_0)$ nos instantes de tempo discretos n e k da seguinte maneira:

$$cor[X(\Delta t_0)_n, X(\Delta t_0)_{n+k}] = E[X(\Delta t_0)_n X(\Delta t_0)_{n+k}] \tag{3.44}$$

Como o MMW consiste da multiplicação de uma cascata multiplicativa $\mu(\Delta t_N)$ por uma variável aleatória lognormal Y, a função de autocorrelação de um processo MMW onde $k = 2^p$, $p = 1, 2, \ldots,$ pode ser escrita como:

$$E[X(\Delta t_0)_n X(\Delta t_0)_{n+k}] = E(Y^2)\{2^{2N} E[\mu(\Delta t_N)_n . \mu(\Delta t_N)_{n+k}] - 1\} + E[X(\Delta t_0)_n] E[X(\Delta t_0)_{n+k}] \tag{3.45}$$

As medidas $\mu(\Delta t_N)_n$ e $\mu(\Delta t_N)_{n+k}$ podem ser expressas em função de $\mu(\Delta t_{N-p-1})$ da seguinte forma:

$$\begin{aligned} \mu(\Delta t_N)_n &= \mu(\Delta t_{N-p-1}) r_{N-p} \prod_{i=N-p+1}^{N} r_{i,j_1} \\ \mu(\Delta t_N)_{n+k} &= \mu(\Delta t_{N-p-1})(1 - r_{N-p}) \prod_{i=N-p+1}^{N} r_{i,j_2} \end{aligned} \tag{3.46}$$

onde $r_{i,j}$ representa o multiplicador no estágio i da cascata. Assim, temos:

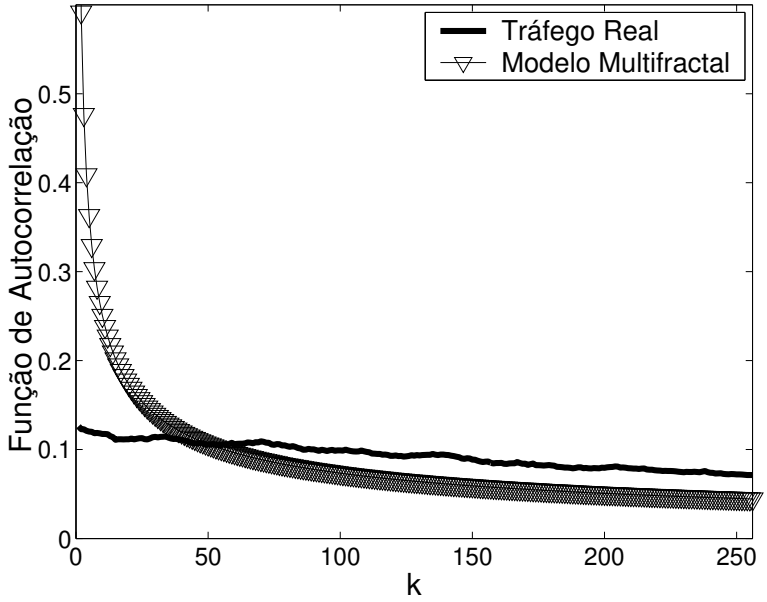


Fig. 3.1: Função de autocorrelação: traço de tráfego dec-pkt-1

$$E[\mu(\Delta t_N)_n \mu(\Delta t_N)_{n+k}] = E[\mu(\Delta t_{N-p-1})^2] E[r_{N-p}(1-r_{N-p})] E\left[\prod_{i=N-p+1}^{N} r_{i,j_1}\, r_{i,j_2}\right]$$

$$= E(R^2)^{N-p-1}\left[\frac{1}{2} - E(R^2)\right]\left(\frac{1}{2}\right)^{2p} \tag{3.47}$$

Inserindo a equação (3.47) em (3.45), obtemos a função de autocorrelação para o modelo multifractal proposto:

$$E[X(\Delta t_0)_n X(\Delta t_0)_{n+k}] = e^{2\rho+\gamma^2}\left(\frac{\alpha(\alpha+1)^{N-1}}{(\alpha+1/2)^N}\left[\frac{\alpha+1}{\alpha+1/2}\right]^{-p} - 1\right) + e^{2\rho+\gamma^2}$$

$$E[X(\Delta t_0)_n X(\Delta t_0)_{n+k}] = e^{2\rho+\gamma^2}\left(\frac{\alpha(\alpha+1)^{N-1}}{(\alpha+1/2)^N} k^{-\log_2(\frac{\alpha+1}{\alpha+1/2})}\right) \tag{3.48}$$

■

Quando N, k são grandes, a função de autocorrelação (3.43) é governada por $k^{-\log_2(\frac{\alpha+1}{\alpha+1/2})}$, ou seja, o modelo tem dependência de longo prazo com parâmetro de Hurst igual ao parâmetro de escala global $H_g = H = 1 - \frac{\log_2(\frac{\alpha+1}{\alpha+1/2})}{2}$. A Figura 3.1 ilustra a função de autocorrelação para a série de tráfego dec-pkt-1 (as séries utilizadas são descritas na seção 3.3.1) calculada pela equação (3.42)

e a função de autocorrelação do modelo expressa por (3.48). Pela inspeção da Figura 3.1 pode-se observar que a função de autocorrelação do modelo expressa por (3.48) apresenta dependência de longa duração. Portanto, o MMW descrito neste capítulo captura a dependência de longa duração e exibe outras características multifractais, como por exemplo, momentos estatísticos representados pela equação (2.16).

3.2.3 Estimação Adaptativa dos Parâmetros (α, γ, ρ)

Os parâmetros de entrada α, γ, e ρ para o modelo multifractal proposto (MMW) podem ser estimados através das funções $\tau(q)$ e $c(q)$ de processos de tráfego reais. Esta seção propõe um algoritmo para estimação adaptativa do conjunto de parâmetros (α, γ, ρ). Este algoritmo permite o cálculo adaptativo das propriedades do modelo multifractal, o que será útil tanto na formulação do preditor de tráfego adaptativo *fuzzy* proposto no Capítulo 4, quanto para o cálculo adaptativo de banda efetiva no Capítulo 5.

O MMW possui uma estrutura analítica adequada à modelagem adaptativa. Para se efetuar esta modelagem adaptativa é preciso atualizar adaptativamente os valores dos parâmetros α, γ e ρ do modelo através dos valores das funções $\tau(q)$ and $c(q)$ do tráfego real. Um método de estimação destas funções baseado na equação (2.16) é o seguinte (Dang et al., 2003): Dado o processo de incrementos $\{X_1, X_2, \ldots X_n\}$, e seja o seu processo agregado correspondente X^m ao nível m definido como:

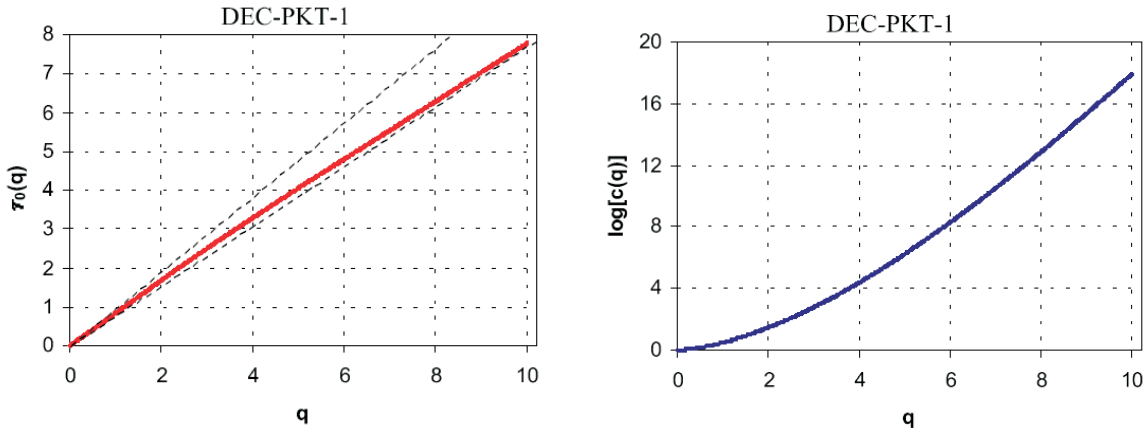
$$X_t^m = X_{(t-1)m+1} + X_{(t-1)m+2} + \ldots + X_{(t)m} \quad t, m = 1, 2, \ldots \tag{3.49}$$

Se a seqüência X_t tem propriedades em escala (*scaling behavior*), então, o momento absoluto $E(|X_t^m|^q)$ versus m em um gráfico log-log deve ser uma reta como a seguinte:

$$\log E(|X_t^m|^q) = \tau_0(q) \log m + \log c(q) \tag{3.50}$$

A inclinação dessa reta provê uma estimativa de $\tau_0(q)$ e o ponto de interseção no eixo vertical corresponde ao valor de $\log c(q)$. Propõe-se estimar os valores de $\tau_0(q)$ e $\log c(q)$ aplicando o método de mínimos quadrados recursivos (Young, 1984). Em seguida, de posse desses valores, usa-se o algoritmo de Levenberg-Marquardt (Marquardt, 1963) para o cálculo de (α, γ, ρ). A partir da equação (3.50) e segundo o método de mínimos quadrados, as seguintes estimativas podem ser feitas para as funções $\tau_0(q)$ e $\log c(q)$ considerando uma série com n amostras:

$$\hat{\tau_0}(q) = \frac{\sum_{m=1}^{n} \log E(|X^m|^q) \log m - nE\{\log m\} E\{\log E(|X^m|^q)\}}{\sum_{m=1}^{n} \log m^2 - nE\{\log m\}^2} \tag{3.51}$$



(a) Função de escala

(b) Logaritmo do Fator de momento

Fig. 3.2: Estimação da função de escala e do fator de momento para a série dec-pkt-1

$$\log \hat{c}(q) = \frac{E\{\log E(|X^m|^q)\} \sum_{m=1}^{n} (\log m)^2 - E\{\log m\} \sum_{m=1}^{n} \log m \log E(|X^m|^q)}{\sum_{m=1}^{n} \log m^2 - nE\{\log m\}^2} \tag{3.52}$$

A determinação do parâmetro α é realizada a partir de $\hat{\tau_0}(q)$ (equação 3.51) usando o algoritmo de Levenberg-Marquardt que consiste de um método de estimação não-linear de parâmetros pela minimização da seguinte função (ver Apêndice D):

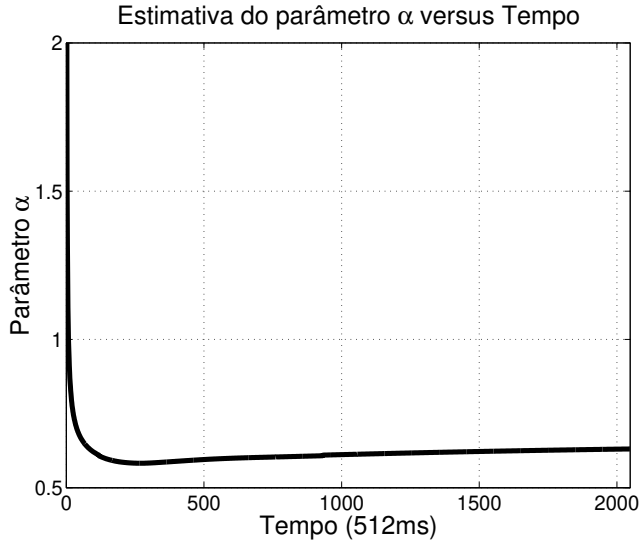
$$\phi = \sum_{q=1}^{n} \left(\log_2 \left(\frac{\Gamma(\alpha)\Gamma(2\alpha+q)}{\Gamma(2\alpha)\Gamma(\alpha+q)} \right) - \hat{\tau_0}(q) \right)^2 \tag{3.53}$$

A equação de atualização do algoritmo de Levenberg-Marquardt para estimação do parâmetro α é (Marquardt, 1963):

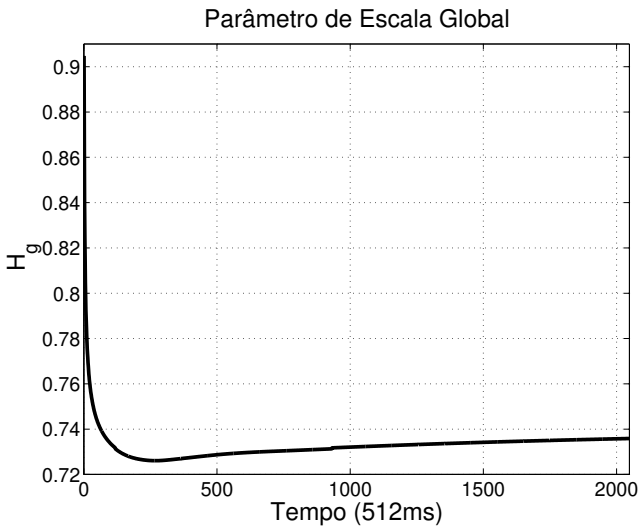
$$\alpha_{i+1} = \alpha_i - (H_{es} + \eta diag(H))^{-1} \nabla \phi(\alpha_i) \tag{3.54}$$

onde H_{es} é a matriz Hessiana ($H_{es} = \nabla^2 \phi(\alpha_i)$), η é um parâmetro de controle, α_i é o valor do parâmetro α na i-ésima iteração do algoritmo e ∇ representa o operador gradiente de uma função.

De forma similar, o algoritmo de Levenberg-Marquardt pode ser aplicado para se estimar os valores dos parâmetros γ e ρ da função $c(q)$. Ao se assumir um cenário adaptativo, considerou-se $t = 1$ e $m = 1, \ldots, n$ na equação (3.50), representando o instante de tempo m no qual os parâmetros são atualizados. Denotaremos esse instante de tempo discreto por k. A partir disso e utilizando o algoritmo de mínimos quadrados recursivos (Young, 1984), é introduzido a seguir um algoritmo adaptativo para atualização dos valores de $\tau_0(q)$ e $\log c(q)$ para estimação do conjunto de parâmetros (α, γ, ρ):



(a) Estimação do parâmetro α: série de tráfego dec-pkt-2



(b) Parâmetro de escala global: série de tráfego dec-pkt-2

Fig. 3.3: Estimação de parâmetros multifractais

Algoritmo 3.2.6 *Algoritmo Proposto para o Cálculo Adaptativo dos Parâmetros Multifractais (α, γ, ρ)*

Passo 1) Seja $p_{1,0} = 1$, $\underline{\hat{a}}_0^q = [0\ 0]$, $k = 0, ..., K$ *e* $q = 1, \ldots, q_2$*;*

Passo 2) Calcule $\underline{\hat{a}}_k^q = [\hat{\tau}_0(q) \log \hat{c}(q)]$ *para todo valor de* q *usando as seguintes equações recursivas:*

$$p_{q,k} = p_{q,k-1} - p_{q,k-1}\underline{x}_k[1 + \underline{x}_k^T p_{k-1}\underline{x}_k] - \underline{x}_k^T p_{q,k-1} \tag{3.55}$$

$$\underline{\hat{a}}_k^q = \underline{\hat{a}}_{k-1}^q - p_{q,k}[x_k x_k^T \underline{\hat{a}}_{k-1}^q - \underline{x}_k y_k^q] \tag{3.56}$$

onde $y_k^q = \log E(|X^{(k)}|)^q$ *e* $\underline{x}_k^T = [1 \log 2 ... \log k]$*;*

Passo 3) Estime o valor do parâmetro α *de* τ_0 *usando o algoritmo de Levenberg-Marquardt de acordo com a equação de atualização:*

$$\alpha_{i+1} = \alpha_i - (H_{es} + \eta diag(H))^{-1} \nabla \phi(\alpha_i) \tag{3.57}$$

Passo 4) Aplique novamente o algoritmo de Levenberg-Marquardt para estimar os valores dos parâmetros ρ *e* γ *da função* $c(q)$*.*

A Figura 3.3(a) mostra o valor de α atualizado adaptativamente pelo algoritmo 3.2.6. Conforme foi demonstrado nas seções 3.2.1 e 3.2.3, é possível calcular o parâmetro de escala global H_g ana-

liticamente em função do parâmetro α pela equação $H_g = 1 - \frac{\log_2(\frac{\alpha+1}{\alpha+1/2})}{2}$, o que neste caso, resulta em $H_g = 0.7338$. Pela Figura 3.3(b), pode-se confirmar que o valor do parâmetro de escala global calculado adaptativamente pelo algoritmo proposto é próximo a este valor teórico esperado ($H_g = 0.7338$)(Vieira & Ling, 2006a).

3.3 Validação do Modelo Multifractal Proposto

Com o objetivo de comparar as características estatísticas do MMW e do tráfego real, assim como do MWM, foram calculados para os mesmos, alguns parâmetros tais como: média, variância, parâmetro de Hurst, função de autocorrelação, momentos de ordem *q* e da análise multifractal, o espectro multifractal. Além disso, para verificar a precisão do modelo MMW em representar tráfego real foram realizadas simulações para analisar o tamanho de fila e a perda de *bytes* em um sistema alimentado pelo traço de tráfego sintético MMW.

Neste capítulo utilizou-se nas simulações traços de tráfego TCP/IP (dec-pkt-1.tcp, dec-pkt-2.tcp, dec-pkt-3.tcp e lbl-tcp-3) obtidos da Digital Equipment Corporation (Erramilli et al., 2000). Considerou-se amostras de tráfego em uma escala de agregação de 512ms para os traços de tráfego TCP/IP, escala na qual os traços apresentam forte característica multifractal (Erramilli et al., 2000). Também foram usados traços de tráfego Ethernet obtidos da Bellcore que apresentam características auto-similares e multifractais (Leland et al., 1994).

Iniciaremos a validação do modelo proposto pela análise das estatísticas de primeira ordem (média e variância) e do parâmetro de Hurst. A Tabela 3.1 exibe a média, a variância e o parâmetro de Hurst para a série de tráfego Ethernet BcAug na escala de 1000ms. Por este teste preliminar pode-se perceber que alguns parâmetros estatísticos básicos dos dados sintéticos gerados segundo o MMW se aproximam dos da série de tráfego real. O parâmetro de escala global dado pela equação (3.39) da série de tráfego BcAug é 0.8186, que é próximo do valor do parâmetro de Hurst calculado pelo método de *wavelets* e mostrado na Tabela 3.1 (ver Apêndice B) (Veitch & Abry, 1999).

Tab. 3.1: Média, Variância e Parâmetro de Hurst.

	Média	Variância	Parâmetro de Hurst.
BcAug	1.3725^5	7.6090^9	0,8850
MMW	1.3725^5	6.5069^{10}	0,8011
MWM	1.3725^5	7.4048^9	0,8399

A Tabela 3.2 compara os valores do parâmetro de Hurst calculado pelo método de *wavelets* e os valores obtidos pelo parâmetro de escala global (equação (3.39)), para duas séries de tráfego.

Verifica-se que os valores do parâmetro de escala global proposto H_g são novamente próximos aos dos valores de H estimados pelo método de *wavelets*. Portanto, o parâmetro H_g pode ser visto como uma forma analítica de cálculo do parâmetro de Hurst que será usado no Capítulo 6 na estimação de probabilidade de perda.

Tab. 3.2: Média, Variância, Parâmetro de Hurst e H_g.

Série de Tráfego	Média	Variância	P.Hurst	H_g
lbl-tcp-5	$2,6146.10^3$	$1,0033.10^7$	0,7811	0,8062
Bc-Aug	$1,3725.10^5$	7.6090^9	0,8850	0,8797

Nas próximas seções será demonstrado através de simulações que o MMW é um modelo multifractal preciso e cujo desempenho de modelagem é comparável ao do MWM. Mas antes disso, se faz necessário uma breve descrição das séries de tráfego utilizadas neste trabalho.

3.3.1 Séries de Tráfego Utilizadas

Utilizou-se nesta tese séries de tráfego com origens distintas tais como traços de tráfego da Bellcore[1], da Petrobrás, da Digital Equipment Corporation (DEC)[2] e Lawrence Berkeley Laboratory (LBL)[3]. Algumas séries de fato já foram citadas no decorrer do texto desta tese.

As séries temporais retiradas das medições da Bellcore têm sido utilizadas em diversos estudos (Veitch et al., 2005)(Veitch & Abry, 1999). Estes arquivos são trilhas contendo duas colunas de dados, a primeira coluna descreve o tempo de início do *frame* e a segunda indica o tamanho de *bytes* do *frame*.

Foram também incluídos neste estudo arquivos de tráfego capturados na rede da Petrobrás entre os anos de 2000 e 2003 através de um analisador de dados da Acterna^{TM} modelo DA350, com a resolução (*time stamp*) de 32 microsegundos (Perlingeiro & Ling, 2005). A notação adotada para designar os diferentes arquivos de tráfego reais analisados é a seguinte (LRPRC, 2002): Os arquivos de tráfego agregado designados com a letra "S" representam traços de tráfego capturados junto aos servidores de aplicações. Arquivos de tráfego agregado designados pela letra "I" foram capturados em roteador de acesso Internet. Os arquivos de tráfego agregado designados por meio da letra "R" foram capturados em roteador de tráfego IP corporativo. Os arquivos designados pelas letras "FTP" e "MTX" são arquivos de tráfego de fonte única de dados e áudio/vídeo, respectivamente. A Figura 3.4

[1] http://www.acm.org/sigcomm/ITA
[2] http://ita.ee.lbl.gov/html/contrib/DEC-PKT.html
[3] http://www.acm.org/sigcomm/ITA

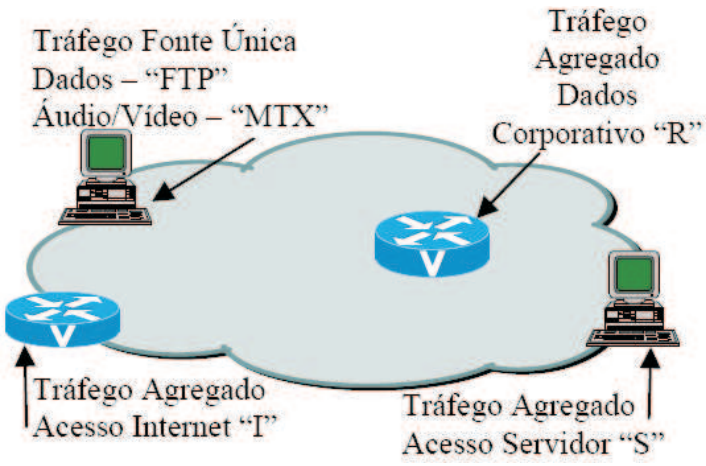


Fig. 3.4: Cenário de captura do tráfego

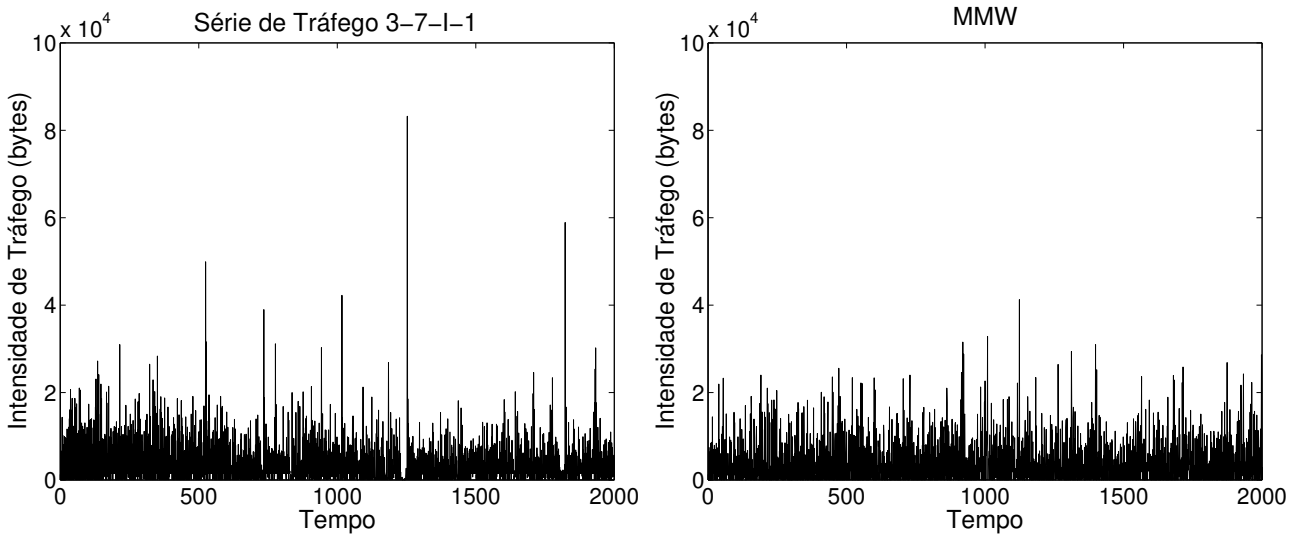
mostra o cenário de captura dos arquivos de tráfego real, sendo que alguns deles são utilizados neste estudo. A série de tráfego da Petrobrás 3-7-I-1 é apresentada na Figura 3.5, assim como a sua série sintética correspondente gerada segundo o MMW.

Com relação às análises de tráfego WAN TCP/IP, utilizou-se traços que correspondem ao registro do tráfego de pacotes IP transmitidos em períodos de uma hora coletados pela Digital Equipment Corporation (DEC) e pelo Lawrence Berkeley Laboratory (LBL) em seus respectivos pontos de acesso de Internet. Entre estes traços de tráfego, podem ser citados os traços dec-pkt-1, dec-pkt-2, dec-pkt-3, lbl-pkt-3, lbl-pkt-5 e lbl-pkt-4, bastante mencionados na literatura (Erramilli et al., 2000) (Veitch et al., 2005). A Figura 3.6 exibe a série de tráfego lbl-pkt-3 e a sua série MMW sintética correspondente.

Pode-se constatar comportamento multifractal em diversas séries utilizadas neste trabalho. A Figura 3.7 mostra o Diagrama Multiescala Linear de 2 séries utilizadas. Note que o Diagramas Multiescala Linear da série lbl-tcp-5 tem um comportamento menos 'horizontal' (constante) do que para a série dec-pkt-2, indicando que a série lbl-tcp-5 é multifractal. O mesmo não pode ser precisamente afirmado para a série dec-pkt-2, principalmente para escalas de tempo maiores do que 100ms.

3.3.2 Coeficiente de Correlação

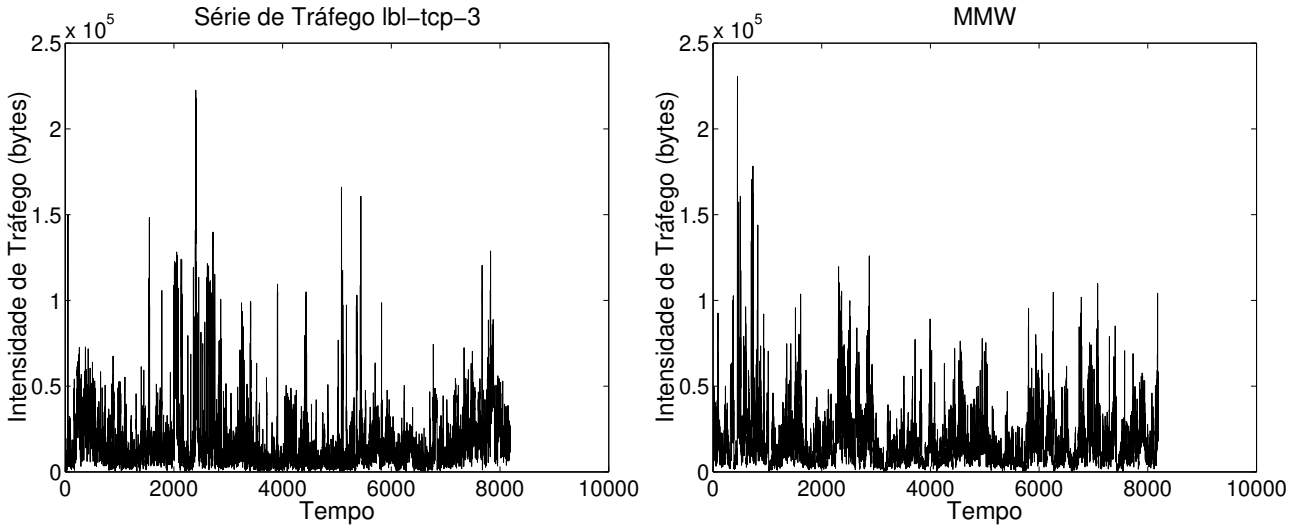
A função de autocorrelação e o coeficiente de correlação refletem as estatísticas de segunda ordem de uma série, dando uma idéia a respeito da longa-dependência nos dados. O coeficiente de correlação pode ser visto como a covariância normalizada de um processo. Seja uma série $y(t)$ com média μ_t e desvio-padrão σ_t, e a mesma série deslocada no tempo $y(t+k)$ com média μ_{t+k} e desvio-padrão



(a) Série de tráfego 3-7-I-1

(b) Série de tráfego gerada segundo o MMW

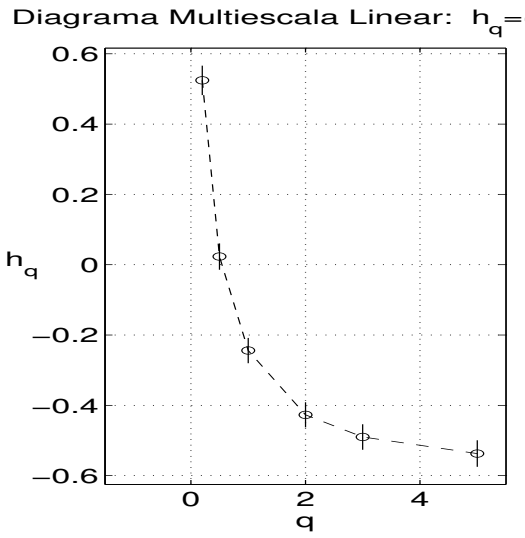
Fig. 3.5: Séries de tráfego real (Petrobrás) e sintética na escala de 100ms



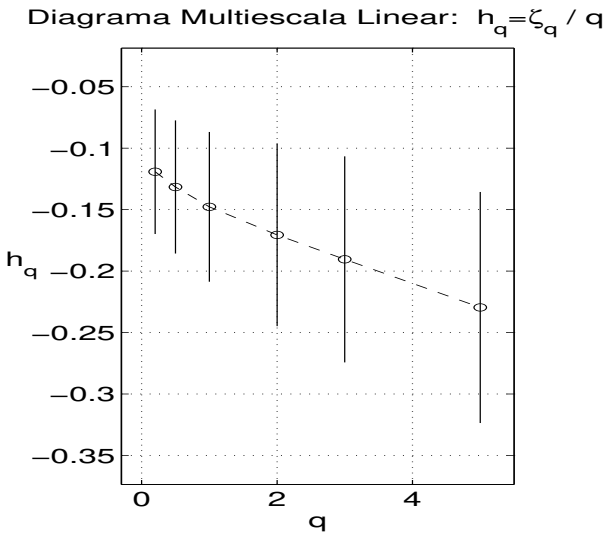
(a) Série de tráfego lbl-tcp-3

(b) Série de tráfego gerada segundo o MMW

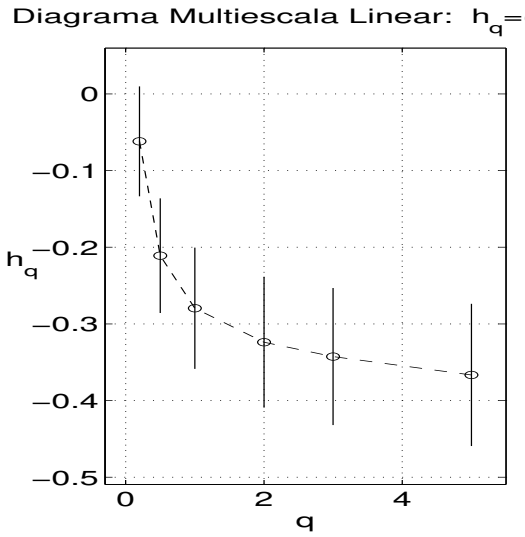
Fig. 3.6: Séries de tráfego real e sintética na escala de 512ms



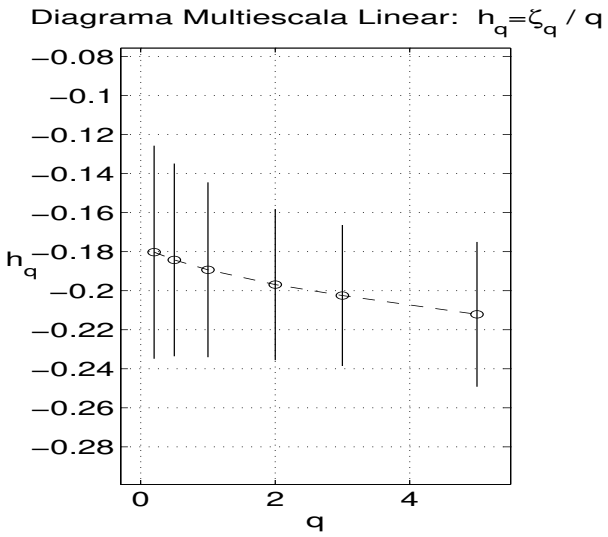
(a) Traço de tráfego lbl-tcp-5 na escala de 1ms



(b) Traço de tráfego lbl-tcp-5 na escala de 100ms



(c) Traço de tráfego dec-pkt-2 na escala de 1ms



(d) Traço de tráfego dec-pkt-2 na escala de 100ms

Fig. 3.7: Diagrama Multiescala Linear

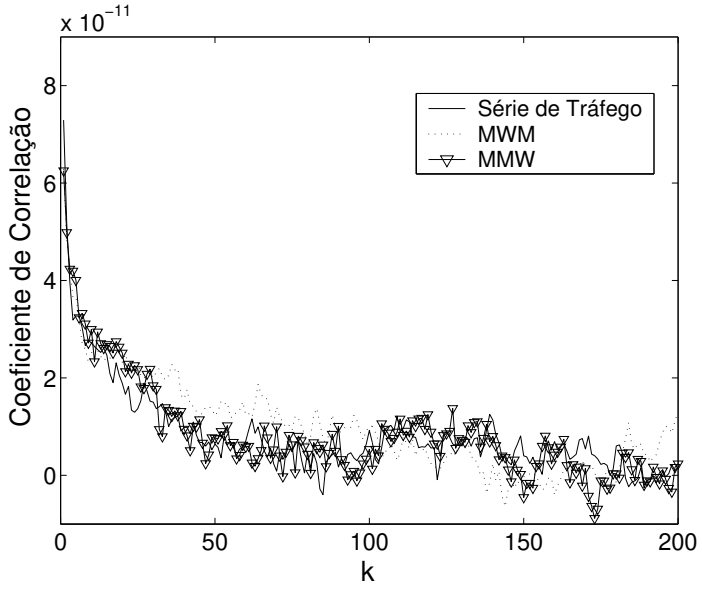


Fig. 3.8: Comparação dos coeficientes de correlação

σ_{t+k}. O coeficiente de correlação para este $y(t)$ é dado por

$$\rho(k) = \frac{E[(y(t+k) - \mu_{t+k})(y(t) - \mu_t)]}{\sigma_{t+k}\sigma_t}. \tag{3.58}$$

A Figura 3.8 compara os coeficientes de correlação calculados usando a equação (3.58) para a série de tráfego Bc-Aug e os traços sintéticos obtidos com os modelos MMW e MWM. O processo gerado segundo o MMW apresenta coeficientes de correlação próximos aos dos calculados diretamente com o tráfego real de rede, mesmo para valores maiores de k. O decaimento da função de autocorrelação evidencia a existência de dependência a longo prazo na série sintética MMW, assim como é observada nas séries de tráfego reais.

3.3.3 Momentos de Ordem q

O tráfego agregado pode influenciar fortemente o comportamento de fila e o desempenho dos multiplexadores de rede. Além disso, agregação de fluxos com diferentes intensidades ocorre em vários pontos da rede. Portanto, é importante investigar não somente o comportamento de fluxos individuais de tráfego mas também o do tráfego agregado. Cabe lembrar que uma das deficiências dos modelos monofractais é o de não conseguir capturar os momentos de mais alta ordem do tráfego de rede. Nesta seção, são analisados os momentos do tráfego agregado do MMW. Seja o processo

agregado X_k^m definido como

$$X_k^m = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i^N, \tag{3.59}$$

onde $i = 1, 2, ..., 2^N$, $k = 1, 2, ..., L$ e $L = 2^N/m$ é o número total de agregações para um valor fixo de *m*. Estima-se o momento de ordem *q* do tráfego agregado da seguinte forma:

$$\hat{\mu}^{(m)}(q) = \frac{1}{L} \sum_{k=1}^{L} |X_k^m|^q \tag{3.60}$$

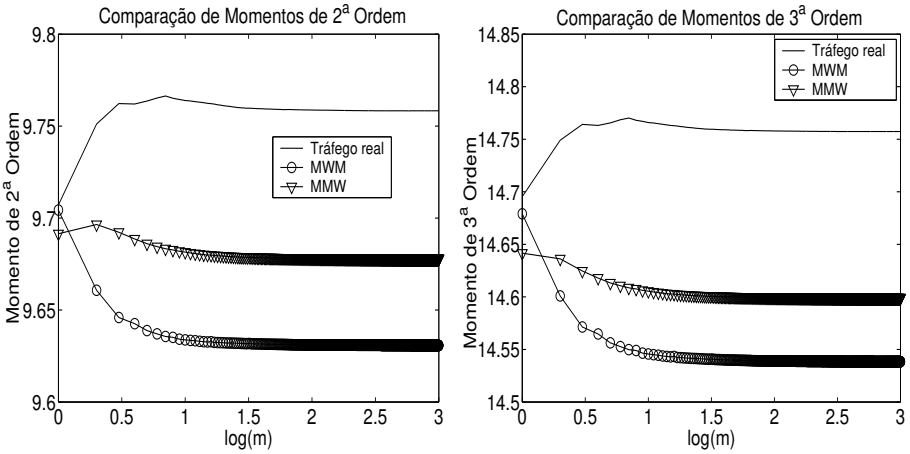
A Figura 3.9 apresenta os resultados das simulações realizadas para o traço de tráfego dec-pkt-1 na escala de agregação de 512ms, expressando a variação de $\log_{10}(\hat{\mu}^{(m)}(q))$ versus $log_{10}(m)$ para $q = 2, 3, 4$ e 5. Para os modelos multifractais considerados, as curvas dos momentos de ordem q do tráfego agregado são obtidas pela média de 100 realizações de processo. Pode-se notar que os modelos multifractais analisados possuem momentos para o tráfego agregado similares, entretanto seus valores são menores do que os das séries de tráfego reais. Isto significa que os dados de tráfego reais podem ter uma estrutura ligeiramente mais complexa do que os traços sintéticos de ambos modelos MMW e MWM.

3.3.4 Espectro Multifractal

Em contraste a outros modelos de tráfego, processos multifractais contêm uma multiplicidade de expoentes de Hölder locais dentro de qualquer intervalo finito. Os expoentes de Hölder descrevem as características em escala locais de um processo em um determinado ponto no tempo. Conforme já mencionado no Capítulo 2, o conceito de expoente de Hölder está relacionado a singularidade local de um processo, ou seja, caracteriza a sua suavidade (quantidade de rajadas) em um certo instante de tempo (Zhang et al., 2003). A distribuição destes expoentes pode ser representada por uma densidade normalizada denominada espectro multifractal. Em uma interpretação alternativa, o espectro multifractal descreve a dimensão fractal do conjunto de instantes possuindo um dado expoente local (Riedi et al., 1999). Nesta seção, o espectro multifractal $f(\alpha)$ de um processo $X(t)$ é calculado como a transformada de Legendre da função de escala $\tau(q)$ pela relação:

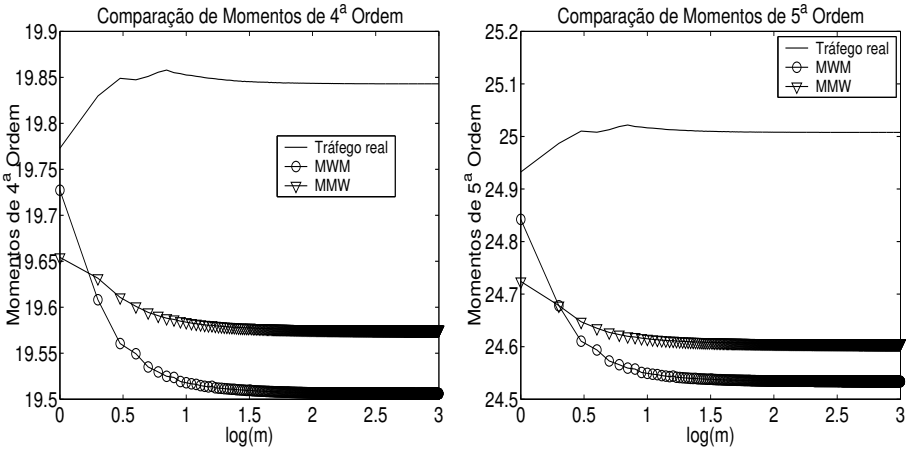
$$f(\alpha) = \min_q \{q\alpha - \tau(q)\} \tag{3.61}$$

Através deste método foram obtidos os espectros multifractais para os processos gerados segundo os modelos MMW e MWM. Como pode ser visto pelas Figuras 3.10(a) e 3.10(b), os espectros multifractais dos traços de tráfego dec-pkt-1 e dec-pkt-2 apresentam $\alpha < 1$, o que indica uma alta incidên-



(a) Momento de $2^{\underline{a}}$ ordem - Traço de tráfego dec-pkt-1

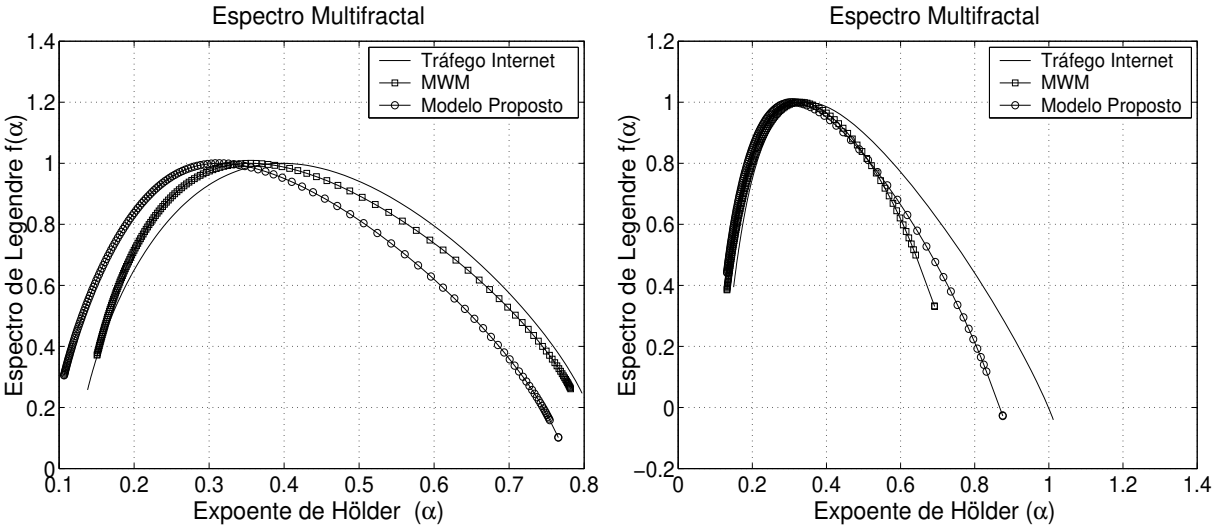
(b) Momento de $3^{\underline{a}}$ ordem - Traço de tráfego dec-pkt-1



(c) Momento de $4^{\underline{a}}$ ordem - Traço de tráfego dec-pkt-1

(d) Momento de $5^{\underline{a}}$ ordem - Traço de tráfego dec-pkt-1

Fig. 3.9: Momentos de ordem q do tráfego agregado



(a) Série de tráfego dec-pkt-1

(b) Série de tráfego dec-pkt-2

Fig. 3.10: Espectro Multifractal de Legendre

cia de rajadas multiescala. Comparativamente, pode-se observar que o modelo MMW captura com eficiência o espectro multifractal e por conseqüência a 'multifractalidade' dos traços de tráfego reais.

3.3.5 Testes de Desempenho e Verificação de Comportamento de Fila

Com intuito de comparar o desempenho do MMW em modelar tráfego real, considerou-se como modelo de simulação do enlace, um servidor com *buffer* finito sendo alimentado pelo processo MMW conforme é ilustrado pela Figura 3.11. Analisou-se principalmente a ocupação média de *bytes* no *buffer* e a porcentagem de perda associadas a diferentes utilizações do *buffer*. A utilização do *buffer* λ é definida como sendo a razão entre o tempo total de serviço fornecido aos pacotes pelo tempo total de uso do *buffer* (Krishna et al., 2003). Um valor de λ próximo de 1 indica que o *buffer* está constantemente sendo usado e conseqüentemente há uma maior probabilidade de descarte de dados. Um valor abaixo de 0.4, significa que o *buffer* está sendo subutilizado, um valor entre 0.6 e 0.7 é considerado adequado em redes reais (Krishna et al., 2003). A utilização do *buffer* λ depende da capacidade do servidor, ou seja, da taxa em que os dados são transmitidos.

Nesta seção, dado um tamanho de *buffer* x, estimou-se a probabilidade de perda de *bytes* $P(Q > x)$, sendo Q o processo de tamanho da fila no *buffer*, pela seguinte expressão:

$$P(Q > x) \cong \frac{N_x}{N_t} \tag{3.62}$$

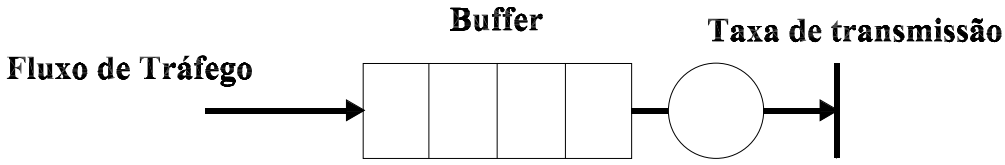


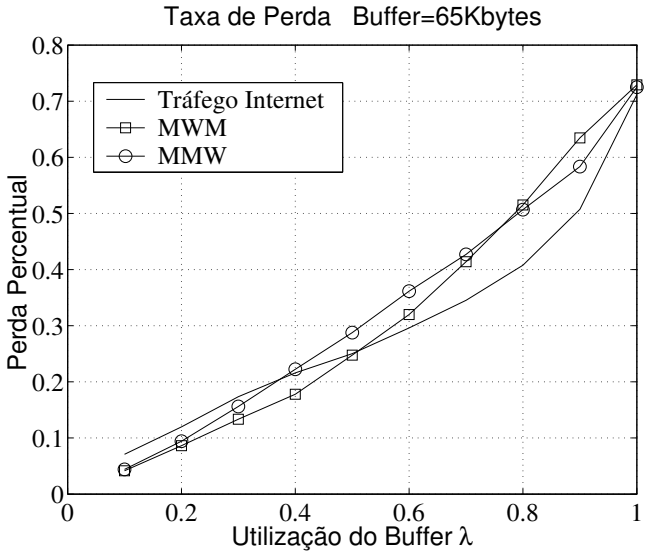
Fig. 3.11: Modelo de simulação usado para o enlace com servidor único e *buffer* finito

onde N_x corresponde ao número de *bytes* descartados que não podem ser armazenados no *buffer* de tamanho x e N_t é o número total de *bytes* a ser atendido.

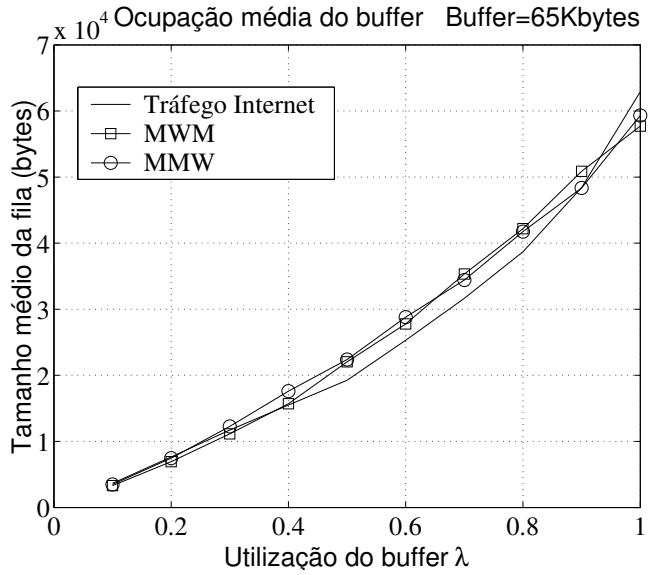
A Figura 3.12(a) mostra os resultados de perda de *bytes* em função da utilização do *buffer* para um *buffer* finito de tamanho 64Kbytes e a Figura 3.12(b) exibe a ocupação média de *bytes* no *buffer* versus a utilização do mesmo para o traço de tráfego lbl-tcp-5 na escala de tempo de 512ms. A relação da taxa de perda de *bytes* e o tamanho do *buffer* pode ser vista na Figura 3.13, onde a capacidade do servidor foi estabelecida como sendo a média da série de tráfego lbl-tcp-3. Os resultados das simulações com várias séries de tráfego confirmam que o modelo proposto reproduz fielmente o comportamento de fila do tráfego Internet para qualquer utilização ou tamanho do *buffer*.

3.4 Considerações Finais

Neste capítulo foi proposto um modelo multifractal baseado em *wavelets* explorando propriedades dos coeficientes *wavelet* de processos multifractais para caracterização do tráfego de rede. Testes estatísticos e de desempenho mostraram que os resultados para o MMW proposto se comparam aos do MWM, bem como aos das simulações com tráfego real de rede. Porém, as vantagens do MMW vão além pois este requer menos parâmetros e os quais podem ser atualizados em tempo real ('on-line'). A função de autocorrelação e os espectros multifractais dos traços sintéticos MMW comprovam a existência de dependência de longo prazo e de características multifractais, as quais são encontradas em traços reais de tráfego de rede. Os resultados de desempenho de fila complementaram a análise realizada sobre os traços sintéticos gerados, concluindo que o MMW é adequado para modelagem de



(a) lbl-tcp-5 Traffic Trace



(b) lbl-tcp-5 Traffic Trace

Fig. 3.12: a) Taxa de perda b) Tamanho médio da fila

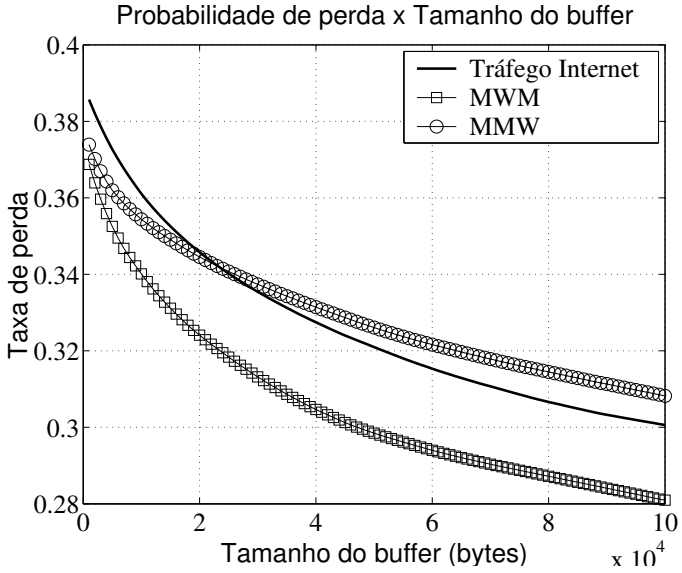


Fig. 3.13: Probabilidade de perda vs Tamanho do *buffer*

tráfego real e pode ser perfeitamente aplicado em avaliações de desempenho de redes.

Devido às suas propriedades, o MMW generaliza uma série de modelos de tráfego de redes, sejam eles de curta ou longa-dependência. Ao se derivar o parâmetro de escala global para este modelo, verifica-se que o tráfego de redes, mesmo ao possuir características multifractais, possui uma lei de escala global. Este parâmetro de escala global será mencionado novamente nesta tese, principalmente como peça elementar para as propostas de cálculo de probabilidade de perda do Capítulo 7. Deve-se salientar que os valores para o parâmetro de escala global H_g se mostraram próximos ao parâmetro de Hurst. No entanto, o parâmetro H_g é obtido analiticamente a partir de um modelo multifractal, o MMW.

A função de autocorrelação pode descrever algumas características de um processo, como o grau de correlação entre as amostras e a presença de dependência de longo prazo. A função de autocorrelação obtida para o MMW confirma analiticamente a dependência de longo prazo presente nos dados gerados segundo este modelo. Esta função ainda será de grande valia no Capítulo 5 no desenvolvimento de um preditor de tráfego que leva em consideração a função de autocorrelação do MMW.

O MMW possui poucos parâmetros de entrada, os quais são estimados a partir da série de tráfego analisada. A estimação adaptativa destes parâmetros permite que alguns resultados sejam obtidos de forma adaptativa. Conforme já foi mencionado, esta característica será explorada na obtenção de algoritmos para o cálculo adaptativo da banda efetiva do MMW.

Capítulo 4

Predição de Séries Temporais: Filtros Adaptativos e Sistemas Fuzzy

4.1 Introdução

As redes atuais de comunicação oferecem grande flexibilidade e vantagem ao utilizar a multiplexação estatística de diferentes tipos de fluxos de tráfego com características estatísticas distintas e dinâmicas. Uma questão importante em gerenciamento de redes é como determinar as bandas a serem alocadas a esses fluxos de modo a garantir qualidade de serviço (QoS) neste ambiente dinâmico. Neste contexto, um modelo preciso de tráfego de rede é necessário para o controle de congestionamento em redes modernas de alta velocidade, evitando que os recursos disponíveis sejam superestimados ou subestimados. Para prover garantia de qualidade de serviço (QoS) a usuários, como perda de *bytes* e atraso, os fluxos de tráfego nos roteadores precisam ser adequadamente controlados. Com o uso de algoritmos de predição e modelos apropriados ao tráfego, pode-se atingir um melhor aproveitamento da banda (capacidade do servidor), e parte desta não utilizada pode ser alocada para outros usuários ou serviços.

Tradicionalmente abordagens baseadas em teoria de filas Markovianas para o cálculo de retardos na fila e probabilidades de perda de pacote são adotadas quando o processo de tráfego de chegada e tamanho de pacotes são descritos por modelos Markovianos e de curta-dependência (Kelly, 1996). Estas soluções podem ser usadas para alocação de banda e dimensionamento de *buffer*. Entretanto, os fluxos de tráfego nas modernas redes multiserviços podem exibir características não consideradas por estes modelos, tornando mais difícil e desafiadora uma apropriada gerência de redes. Um desses comportamentos do tráfego é a característica de invariância em escala que está associada a processos auto-similares e tem relação com a quantidade de rajadas presentes, sendo este último fator crucial na avaliação de QoS (Park & Willinger, 2000). Um aumento do número de rajadas, em geral pode con-

duzir a uma diminuição na qualidade de serviço. Embora a característica auto-similar e a dependência de longo prazo possam ser capturadas por processos monofractais (Leland et al., 1994), em (Riedi et al., 1999) (Feldmann et al., 1998) os autores evidenciam que as características do tráfego, como por exemplo suas propriedades em escala e irregularidades locais, podem ser ainda mais complexas. Na busca de uma descrição mais completa do tráfego, modelos multifractais têm sido usados na condução de importantes investigações a respeito do desempenho de redes (Molnar & Terdik, 2001) (Ribeiro et al., 2000).

Além desses estudos de modelagem de tráfego, existem outros que têm se concentrado em modelos não-paramétricos os quais não fazem considerações prévias sobre as características do tráfego citadas anteriormente e podem ser aplicados à predição de tráfego. Entre estes modelos estão as redes neurais e modelagem *fuzzy* (Vieira et al., 2003a) (Vieira & Lee, 2004a). Novas modelagens inteligentes que empregam técnicas motivadas por sistemas biológicos e inteligência humana têm sido introduzidos, como algoritmos genéticos e sistemas *fuzzy*. Avanços neste sentido podem ser encontrados nos trabalhos de Vieira et al. (Vieira et al., 2003b) (Vieira & Lee, 2004c) (Vieira & Lee, 2004b). Quanto a modelagem *fuzzy*, esta é um exemplo típico de técnica que faz uso do processo de dedução e raciocínio humano. A modelagem *fuzzy* tem sido bastante empregada em vários campos de pesquisa, desde que a teoria *fuzzy* foi inicialmente desenvolvida (Bezdek, 1993). A razão para estas pesquisas é que os modelos *fuzzy* possuem algumas vantagens com relação a determinados sistemas sobre modelos lineares, como por exemplo, na descrição de processos reais desconhecidos com características não-lineares e variantes no tempo, como o tráfego de redes (Chen et al., 2000).

Este capítulo descreve maneiras de se modelar um sistema para que, com informações da série temporal de entrada, seja possível fazer predições de seus valores futuros. A modelagem desse sistema é obtida pela relação entre valores prévios e futuros de uma série temporal. Ao se modelar um sinal, imita-se o comportamento do sistema que o gerou, sem o conhecimento de seus mecanismos de funcionamento. O propósito desta modelagem pode ser preditivo ou apenas de caracterização, ou seja, uma forma de se conhecer melhor e reproduzir o sistema. Serão tratados neste capítulo, modelos preditivos entre eles, o filtro de Wiener, filtros adaptativos e sistemas *fuzzy*, os quais encontram relações entre valores de uma série temporal.

4.2 Predição e Controle de Tráfego

Congestionamento ocorre quando a demanda de tráfego não pode ser atendida pelos recursos disponíveis da rede. A engenharia de tráfego visa assegurar às rede de computadores capacidade para suportar a demanda de tráfego com a qualidade de serviço exigida e prevenir congestionamento. As características multifractais presentes no tráfego dificultam a adoção de modelos mais simples que

sejam apropriados em capturar eficientemente estas propriedades. O controle de congestionamento aplicado às redes deve tentar contornar o problema trazido por estas características, como perda de pacotes e atrasos.

À medida que as redes se tornam maiores e mais heterogêneas, com maiores velocidades e tráfego integrado, o problema de controle de congestionamento se torna mais crítico. As redes deverão lidar com diferentes tipos de serviço e garantias de QoS para classes de tráfego. O controle de tráfego deve levar em conta que em uma rede multiserviço se tem esta variedade de tráfego, alguns cuja integridade temporal deve ser preservada, como é o caso de serviços de vídeo e telefone, outros em que a perda de dados é mais relevante do que o atraso e tráfego elástico em que a taxa varia. Sem mecanismos de controle, a entrega de pacotes e qualidade de serviço podem ficar seriamente comprometidos na ocorrência de congestionamento. Como exemplo, os mecanismos atuais de controle de congestionamento da Internet poderão resultar em desempenho insatisfatório (baixa utilização e alta taxa de perda), à medida que o número de usuários e o tamanho da rede aumentam (Ryu, 2003). Nas redes atuais TCP/IP, um aviso de esgotamento de tempo ('timeout') ou a chegada de reconhecimento duplicado são usados como indicadores de congestionamento (Ryu, 2003). Quando congestionamento ocorre, o protocolo TCP controla sua taxa de envio, que é determinada de acordo com a taxa de chegada dos reconhecimentos (ACK-*Acknowledgement*) de pacotes anteriores. A taxa de chegada de pacotes é então determinada em função da presença ou ausência de enlaces congestionados no caminho entre a origem e o destino.

Há alguns mitos e pensamentos errôneos. Os problemas de congestionamento não podem ser resolvidos com : aumento do tamanho do *buffer*, processadores mais rápidos ou enlaces de alta velocidade (Jain, 1990). Em primeiro lugar, mesmo *buffer* com memória infinita são suscetíveis a congestionamento. Depois, enlaces de alta velocidade não podem ser usados com enlaces de baixa velocidade, e mesmo se todos enlaces e processadores possuírem a mesma velocidade, congestionamento pode ocorrer. Os problemas de congestionamento de longa, média e curta duração podem possuir diferentes soluções (Ryu, 2003). Para congestionamento de curta duração, controles a nível de enlace e rede tal como criação de classes de prioridades são requeridos, enquanto para congestionamento com mais longa duração, um controle a nível de sessão ou um esquema de criação de recursos deve ser usado. Se o congestionamento se estende indefinidamente, talvez a melhor saída seja realmente instalar recursos extras (Jain, 1990).

De acordo com as ações tomadas pelo controle de tráfego, este pode receber diferentes classificações. O controle de congestionamento pode ser preventivo, em que ações são tomadas para evitar o congestionamento baseadas principalmente no contrato de tráfego, além de ações como: policiamento de usuários, controle de prioridades e moldagem (*shaping*) de tráfego. Porém, quando o processo é de longa dependência por exemplo, é mais difícil de se garantir que o tráfego respeite o contrato na

admissão da conexão. Além do mais, o algoritmo de balde furado, que comumente é indicado como método de policiamento, não é uma solução satisfatória para tráfego auto-similar (Roberts, 2000). Outro tipo de controle é o reativo, em que se tenta minimizar a duração e o efeito do congestionamento, entre os métodos existentes há o controle de fluxo, descarte seletivo de células e notificação de congestionamento. A alta velocidade das redes pode tornar o controle reativo ineficiente, porém ele evita maiores danos no caso de tráfego auto-similar. O congestionamento nas redes pode ser evitado ao se alocar recursos de acordo com o valores de tráfego preditos. Ao se alocar taxa aos fluxos em uma rede de acordo com os valores de intensidade de tráfego preditos, os protocolos da rede podem tomar ações antes que o congestionamento ou a diminuição da qualidade de serviço aconteçam. Este tipo de controle é denominado de controle preditivo que tem como base algoritmos de predição de tráfego e serão tratado neste e no próximo capítulo.

4.3 Predição e Filtragem Linear

Ao se formular um modelo preditor deve-se considerar que os mecanismos do sistema a ser modelado não são conhecidos e que a informação deve ser extraída dos dados de entrada, ou seja, da série temporal correspondente aos dados de tráfego e das saídas desse sistema. Nesta tese, a série temporal cujos valores serão preditos consiste de números de *bytes* de dados medidos em diferentes instantes de tempo. A predição dos fluxos de tráfego pode ser usada em várias áreas da engenharia de tráfego de redes, tais como alocação dinâmica de banda (Adas, 1998) (Chong et al., 1995), suavização de tráfego (Yuang, 1997) , controle de congestionamento (Ramamurthy & Sengupta, 1996), controle de admissão (Shiomoto et al., 1999), etc.

Seja $x(t)$ uma série temporal e $\underline{x}(t) = [x(t), x(t-1), ..., x(t-L-1)]^T$ um vetor com L amostras passadas do processo x no instante t. Em 1927, Yule propôs que a predição de um valor da série $x(t)$ um passo a frente, ou seja, um instante de tempo posterior, poderia ser escrita da seguinte forma (Yule, 1927):

$$\hat{x}(t+1) = F[\underline{x}(t)] = F[x(t), x(t-1), ..., x(t-L-1)]^T \tag{4.1}$$

onde F é a função que representa a dinâmica do sistema. Esse método também justificado por Takens e usado para sistemas sem ruído é adotado em diversas áreas (Takens, 1981).

Segundo a equação (4.1), o valor predito pode ser expresso como função de valores anteriores. É dito que a predição é linear se a função F pode ser descrita como uma combinação linear das amostras passadas. Estruturas lineares são aquelas cujo mapeamento obedece ao princípio da superposição, ou seja, cuja resposta a uma combinação linear de entradas é a combinação linear das respostas a cada entrada, enquanto as estruturas não-lineares não obedecem a este princípio. Os modelos de predição linear têm uma estrutura mais simples do que os modelos não-lineares.

Através da equação (4.1) define-se o modelo AR (Auto-Regressivo) de ordem P de processos estocásticos representado pela seguinte equação:

$$x(n) = \sum_{k=1}^{P} w(k)x(k-n) + e(n), \tag{4.2}$$

onde $w(k)$ representa os coeficientes do modelo no instante de tempo discreto n. A equação (4.2) formaliza que a predição de um valor no instante n pode ser obtida da soma ponderada dos valores passados de uma série mais o erro de predição $e(n)$.

Na predição não-linear o mapeamento entrada-saída é mais rico do que na predição linear, que consiste de um hiperplano simples. Há vários modelos não-lineares na literatura, entre eles, pode-se citar o modelo Auto-Regressivo Não-Linear (NAR- *Nonlinear AutoRegressive*) cuja equação é a seguinte:

$$x(n) = G(x(n-1), x(n-2), ..., x(n-p)) + e(n), \tag{4.3}$$

onde G agora é uma função não-linear das amostras passadas do processo.

Outro modelo não-linear e que oferece uma melhor descrição para diferentes tipos de sinais é o modelo NARMA (*Nonlinear Autoregressive Moving Average*) (Box & Jenkins, 1976). Este modelo se baseia em uma função mais geral que incorpora erros passados:

$$\hat{x}(n) = F[x(n-1), ..., x(n-p)], e(n-1), ...e(n-p)]. \tag{4.4}$$

Pela equação (4.4) nota-se que mais informação pode ser disponibilizada a um modelo de tal forma que este apresente predições mais precisas. Neste capítulo será dado ênfase aos algoritmos de predição lineares LMS (*Least Mean Square*) e RLS (*Recursive Least Squares*) e com relação a modelos não-lineares de predição, serão abordados os modelos *fuzzy*, em especial o proposto por Takagi-Sugeno (Sugeno & Yasukawa, 1993).

4.3.1 Filtros de Wiener

O problema de filtragem consiste em reduzir o efeito do ruído presente no sinal de entrada produzindo na saída do filtro um sinal de interesse. Filtros em tempo discreto buscam oferecer soluções em tempo discreto ao problema de filtragem. Na década de 40, Wiener começou a aplicar a teoria de predição linear em várias situações (Haykin, 1989). Ótimo no sentido do erro quadrático médio, o filtro de Wiener minimiza o valor quadrático médio do sinal de erro, definido como a diferença entre o sinal de saída do filtro e algum sinal desejado. Uma importante característica da teoria de Wiener

é que esta requer apenas o conhecimento das estatísticas de primeira e segunda ordens do sinal de interesse.

Predição é um assunto de especial interesse em processamento de sinais, e consiste na particularização do problema mais genérico da filtragem linear, sendo o filtro projetado para tal fim denominado filtro preditor. Seja $\{x(n)\}$ um processo estocástico discreto e estacionário em sentido amplo. O projeto de um preditor linear que minimize o erro quadrático médio de estimação consiste em encontrar um filtro de Wiener discreto que a partir da combinação linear de amostras do sinal de entrada $x(n-1)$, $x(n-2)$, ..., $x(n-M)$, estime o valor da amostra $x(n+\Delta-1)$, onde *M* é número de amostras de entrada do filtro e Δ é o número de passos futuros.

O preditor de Wiener é, portanto, um filtro linear transversal cujos coeficientes são otimizados objetivando-se minimizar o valor do erro quadrático médio, conforme ilustrado na Figura 4.1.

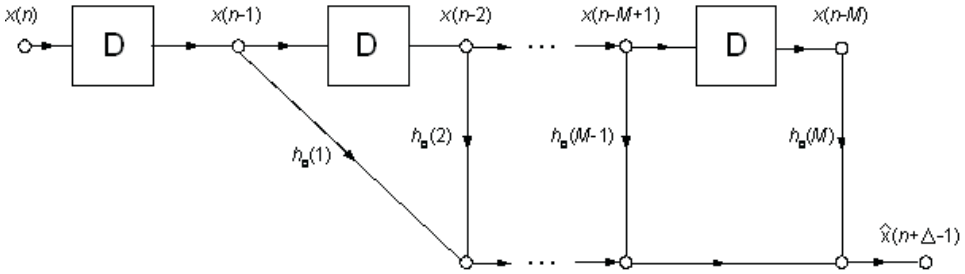


Fig. 4.1: Filtro linear transversal

A relação entrada-saída do filtro é descrita através da seguinte soma-convolução:

$$\hat{x}(n+\Delta-1) = \sum_{k=1}^{M} h_o(k)x(n-k) \tag{4.5}$$

onde $h_o(k)$ é o vetor de coeficientes ótimos do filtro, $x(n-k)$ é a amostra do processo *k*-instantes de tempo anteriores, e $\hat{x}(n+\Delta-1)$ é a saída do filtro correspondente à estimação da resposta desejada $x(n+\Delta-1)$.

A resposta desejada $x(n+\Delta-1)$ (predição) será denotada por $d(n+\Delta-1)$. Assumindo que o sinal de entrada $x(n)$ e a resposta desejada $d(n)$ sejam conjuntamente estacionários em sentido amplo, ou seja, estes processos são separadamente estacionários em sentido amplo e sua função de autocorrelação cruzada $r_{dx}(k)$ não depende do tempo n, as seguintes afirmações podem ser feitas:

1) A função de autocorrelação $r_x(k-m)$ entre as amostras de entrada do filtro pode ser dada por:

$$r_x(k-m) = E\left[x(n-m)x(n-k)\right], \quad k,m = 1,2,3,\ldots,M \tag{4.6}$$

2) A média $E[d^2(n)]$ é igual ao valor médio quadrático do sinal de resposta desejada $d(n)$, ou

seja:

$$r_d(0) = E[d^2(n)] \tag{4.7}$$

3) A correlação cruzada $r_{dx}(k)$ entre a resposta desejada $d(n+\Delta-1)$ e as entradas do filtro é descrita como:

$$r_{dx}(k) = E\left[d(n+\Delta-1)x(n-k)\right], \;\; k = 1, 2, 3, \ldots, M \tag{4.8}$$

O erro de estimação consiste na diferença entre a resposta desejada e a saída do preditor:

$$e(n+\Delta-1) = d(n+\Delta-1) - \hat{x}(n+\Delta-1) \tag{4.9}$$

Seja $\varepsilon = E\left[e(n+\Delta-1)\right]^2$ o erro quadrático médio. O erro quadrático médio atinge seu valor mínimo quando suas derivadas em relação a $h(k)$ para $k = 1, 2, 3, \ldots, M$ são simultaneamente zero. Ou seja,

$$\frac{\partial \varepsilon}{\partial h(k)} = 0, \tag{4.10}$$

para $k = 1, 2, 3, \ldots, M$.

Fazendo a derivada de ε em relação a $h(k)$ igual a zero, obtém-se a equação para os coeficientes ótimos do filtro. Tal equação é uma versão em tempo discreto da equação de *Wiener-Hopf*, sendo dada por

$$\sum_{m=1}^{M} h_o(m) r_x(k-m) = r_{dx}(k), \;\; k = 1, 2, 3, \ldots, M \tag{4.11}$$

Obtidos os coeficientes ótimos do filtro a partir da equação (4.11), o erro quadrático médio mínimo $\varepsilon_{\min}$ produzido por este filtro é dado por

$$\varepsilon_{\min} = r_d(0) - \sum_{k=1}^{M} h_o(k) r_{dx}(k). \tag{4.12}$$

A equação (4.11) é comumente chamada na literatura de *sistema de equações normais*. Este sistema de equações pode ser expresso de uma maneira compacta através da utilização de notação matricial. Para tal fim, sejam as seguintes definições:

Definição 4.3.1 *- Define-se o vetor coeficiente como uma matriz Mx1 dada por:*

$$\boldsymbol{h}_0 = \begin{bmatrix} h_o(1) \\ h_o(2) \\ \vdots \\ h_o(M) \end{bmatrix} \tag{4.13}$$

Definição 4.3.2 *- Define-se o vetor de correlação cruzada como uma matriz Mx1, na qual seus elementos consistem nas correlações entre a resposta desejada $x(n+\Delta-1)$ e o valor presente nas entradas do filtro, $x(n-1)$, $x(n-2)$, ..., $x(n-M)$, dada por:*

$$\boldsymbol{r}_{dx}= \begin{bmatrix} r_{dx}(1) \\ r_{dx}(2) \\ \vdots \\ r_{dx}(M) \end{bmatrix} \tag{4.14}$$

Definição 4.3.3 *Define-se a matriz de correlação como uma matriz MxM onde os elementos consistem nos valores quadráticos médios das entradas do filtro, $x(n-1)$, $x(n-2)$, ..., $x(n-M)$, assim como as correlações entre essas entradas. A matriz de correlação é dada por:*

$$\boldsymbol{R}_x= \begin{bmatrix} r_x(0) & r_x(1) & \cdots & r_x(M-1) \\ r_x(1) & r_x(0) & \cdots & r_x(M-2) \\ \vdots & \vdots & & \vdots \\ r_x(M-1) & r_x(M-2) & \cdots & r_x(0) \end{bmatrix} \tag{4.15}$$

Assim, utilizando-se as definições 4.3.1, 4.3.2 e 4.3.3, as equações normais (equação (4.11)) podem ser reescritas na seguinte forma matricial:

$$\mathbf{R}_x\mathbf{h}_o = \mathbf{r}_{dx} \tag{4.16}$$

Portanto, desde que a matriz $\mathbf{R}_x$ seja não singular, o vetor de coeficientes ótimos do filtro pode ser obtido através da solução da seguinte equação matricial:

$$\mathbf{h}_o = \mathbf{R}_x^{-1}\mathbf{r}_{dx} \tag{4.17}$$

A expressão do erro quadrático médio mínimo (equação (4.12)) também pode ser escrita na forma matricial que segue:

$$\bar{\varepsilon}_{min} = \mathbf{r}_d(0) - \mathbf{r}_{dx}^T\mathbf{h}_o \tag{4.18}$$

onde o T sobrescrito significa a operação de transposição de matriz.

A já mencionada característica dos filtros de Wiener, de requerer apenas o conhecimento das estatísticas de primeira e segunda ordens do sinal de interesse, é confirmada através da observação das equações (4.11) e (4.17). Nestas equações, apenas o valor quadrático médio dos valores presentes nas entradas, a correlação entre estes valores, e a correlação cruzada entre o valor desejado e os valores nas entradas do filtro são necessários para a obtenção dos coeficientes. Com base no Filtro

de Wiener, Hirchoren et al. propuseram um preditor linear em tempo discreto para o ruído Gaussiano fracionário, confirmando que é possível a inserção de informação sobre a característica fractal do tráfego em preditores lineares (Hirchoren & Arantes, 1998) (Hirchoren, 1999).

4.3.2 Preditor Linear Adaptativo LMS

A implementação de filtros de Wiener utiliza as equações normais no cálculo dos coeficientes do filtro, sendo necessário o conhecimento da correlação entre as entradas do filtro, e também a correlação cruzada entre as entradas do filtro e a resposta desejada. Encontrados os coeficientes do filtro $h_o(k)$ a partir das equações normais, a saída do filtro de Wiener é dada por

$$y(n) = \sum_{k=1}^{M} h_o(k)x(n-k). \tag{4.19}$$

Entretanto, quando o filtro é utilizado em um ambiente desconhecido, versões adaptativas, capazes de aprender com o ambiente e de ajustarem seus coeficientes de uma maneira recursiva, devem ser buscadas.

Conhecido como LMS, o algoritmo *least mean-square* consiste de uma implementação adaptativa do filtro de Wiener, proporcionando uma solução recursiva para as equações normais. O algoritmo LMS é obtido através da variação dos coeficientes do filtro $h(k)$ no tempo discreto n. Os coeficientes do filtro são então denotados por $h(k, n+1)$ e calculados de maneira recursiva por:

$$h(k, n+1) = h(k, n) - \frac{1}{2}\mu\nabla_k(n), \tag{4.20}$$

onde μ é uma constante e $\nabla_k(n)$ é o gradiente da superfície de desempenho de erro em relação ao *k*-ésimo coeficiente do filtro. Ou seja:

$$\nabla_k(n) = \frac{\partial\varepsilon(n)}{\partial h(k, n)}, \tag{4.21}$$

onde $\varepsilon(n) = E[e^2(n)]$ e $e(n) = d(n) - y(n)$ é o sinal erro.

A derivada da função custo $\varepsilon(n)$ em relação aos coeficientes do filtro é obtida como

$$\frac{\partial\varepsilon}{\partial h(k)} = -2E\left[e(n)x(n-k)\right] = -2r_{ex}(k), \tag{4.22}$$

onde $r_{ex}(k)$ é a correlação cruzada entre o sinal erro $e(n)$ e as entradas do filtro. Portanto, pode-se

reescrever a equação (4.20) como

$$h(k, n+1) = h(k, n) + \mu r_{ex}(k). \tag{4.23}$$

A equação (4.23) mostra então que se pode obter o novo valor do *k*-ésimo coeficiente do filtro aplicando-se uma correção ao seu valor anterior $h(k, n)$. Idealmente, o valor médio do termo de correção em (4.23) deve se aproximar de zero quando n se aproxima de infinito.

Para a implementação do filtro adaptativo utiliza-se uma estimação instantânea para a correlação cruzada $r_{ex}(k)$, dada por

$$\hat{r}_{ex}(k) = e(n)x(n-k). \tag{4.24}$$

Seja $\hat{h}(k, n)$ a correspondente estimativa do coeficiente do filtro, escreve-se a equação de atualização do algoritmo LMS como

$$\hat{h}(k, n+1) = \hat{h}(k, n) + \mu.e(n)x(n-k), k = 1, 2, ..., \boldsymbol{M}\,. \tag{4.25}$$

O fator μ é chamado de parâmetro de adaptação. O sinal erro *e*(*n*) utilizado em (4.25) é dado por:

$$e(n) = d(n) - \sum_{k=1}^{M} \hat{h}(k, n)x(n-k), k = 1, 2, ..., \boldsymbol{M}\,. \tag{4.26}$$

Neste trabalho, todos os coeficientes do filtro no instante $n = 0$ são estabelecidos iguais a zero. Quanto ao valor designado ao parâmetro de adaptação μ, este deve ser tal que se garanta a convergência do algoritmo LMS. Portanto, o valor do parâmetro de adaptação do algoritmo LMS é dependente da série aplicada ao filtro.

4.3.3 Preditor Linear Adaptativo RLS

O filtro preditor RLS (*Recursive Least Squares*) é derivado a partir da minimização de uma função de erro quadrático mínimo ponderado. Seja um filtro de Wiener com coeficientes

$$\mathbf{w}_n = [w_n(0), w_n(1), ..., w_n(p)]^T \tag{4.27}$$

que minimizam no instante de tempo n, o erro quadrático mínimo ponderado dado por

$$\varepsilon(n) = \sum_{i=0}^{n} \lambda^{n-1} |e(i)|^2\,, \tag{4.28}$$

onde $0 < \lambda \leq 1$ é denominado de fator de esquecimento (ou fator de ponderação exponencial) e

$$e(i) = d(i) - y(i) = d(i) - \mathbf{w}_n^T \mathbf{x}(i) \tag{4.29}$$

Note que $e(i)$ é a diferença entre o sinal desejado $d(i)$ e a saída do filtro $y(i) = \mathbf{w}_n^T \mathbf{x}(i)$ no instante de tempo i. Para se encontrar os coeficientes w_n que minimizam $\varepsilon(n)$ se faz $\frac{\partial \varepsilon(n)}{\partial w_n} = 0$. Pode-se demonstrar que esse procedimento tem como resultado o conhecido sistema de equações normais determinísticos (Hayes, 1996):

$$\mathbf{R}_x(n)\mathbf{w}_n = \mathbf{r}_{dx}(n) \tag{4.30}$$

onde $\mathbf{R}_x(n)$ é a matriz de autocorrelação determinística de $x(n)$ ponderada exponencialmente dada da seguinte forma:

$$\mathbf{R}_x(n) = \sum_{i=0}^{n} \lambda^{n-1} \mathbf{x}^*(i)\mathbf{x}^T(i) \tag{4.31}$$

em que $\mathbf{x}^*(i)$ é o complexo conjugado de $\mathbf{x}(i)$, o vetor de dados representado por:

$$\mathbf{x}(i) = [x(i), x(i-1), ..., x(i-p)]^T \tag{4.32}$$

sendo $\mathbf{r}_{dx}(n)$ a correlação cruzada determinística entre $d(n)$ e $x(n)$,

$$r_{dx}(n) = \sum_{i=0}^{n} \lambda^{n-1} d(i)x^*(i). \tag{4.33}$$

Os coeficientes do filtro RLS são ótimos para uma dada série, ao invés de estatísticamente ótimos para uma determinada classe de processos. O erro mínimo do algoritmo RLS é expresso pela seguinte equação:

$$\{\varepsilon(n)\}_{\min} = \|\mathbf{d}(n)\|_\lambda^2 - \mathbf{r}_{dx}^H(n)\mathbf{w}_n \tag{4.34}$$

onde $\|\mathbf{d}(n)\|_\lambda^2$ é a norma ponderada do vetor $\mathbf{d}(n) = [d(n), d(n-1), \ldots, d(0)]^T$ e $\mathbf{r}_{dx}^H(n)$ representa o complexo conjugado do vetor transposto de $\mathbf{r}_{dx}(n)$, i.e. , $\mathbf{r}_{dx}^H(n) = (\mathbf{r}_{dx}(n)^T)^*$.

Como $\mathbf{R}_x(n)$ e $\mathbf{r}_{dx}(n)$ dependem do tempo n, pode-se derivar uma solução recursiva para as equações normais determinísticas. Assim, os coeficientes do filtro podem ser calculados recursivamente pela seguinte equação (Hayes, 1996):

$$\mathbf{w}_n = \mathbf{w}_{n-1} + \alpha(n)\mathbf{g}(n) \tag{4.35}$$

onde $\alpha(n) = d(n) - \mathbf{w}_{n-1}^T \mathbf{x}_n$ e $\mathbf{g}(n)$ é o vetor de ganho dado por:

$$\mathbf{g}(n) = \frac{1 - \lambda^{-1}\mathbf{P}(n-1)\mathbf{x}^*(n)}{1 - \lambda^{-1}\mathbf{x}^T(n)\mathbf{P}(n-1)x^*(n)} \tag{4.36}$$

e

$$\mathbf{P}(n) = \mathbf{R}_x^{-1}(n) = \lambda^{-1}\left[\mathbf{P}(n-1) - g(n)\mathbf{x}^T(n)\mathbf{P}(n-1)\right] \tag{4.37}$$

A variável $\alpha(n)$ representa o erro a priori, ou seja, é o erro que ocorreria se os coeficientes do filtro não fossem atualizados. Por outro lado, o erro a posteriori $e(n)$ é o erro que ocorre após a atualização do vetor de coeficientes,

$$e(n) = d(n) - \mathbf{w}_n^T\mathbf{x}_n. \tag{4.38}$$

Nota-se que no cálculo do vetor de ganho $\mathbf{g}(n)$ e na matriz de autocorrelação inversa $\mathbf{P}(n)$ é necessário calcular o seguinte produto:

$$\mathbf{z}(n) = \mathbf{P}(n-1)\mathbf{x}^*(n) \tag{4.39}$$

Para se finalizar a apresentação das equações que envolvem o algoritmo RLS, é preciso verificar sua inicialização. Uma maneira muito usada de se inicializar o algoritmo é fazer $\mathbf{P}(0) = \delta^{-1}I$ onde δ é uma constante positiva de valor pequeno, assim como inicializar os pesos com valor zero, $\mathbf{w}_0 = 0$. Dessa forma, se obtém o algoritmo RLS descrito a seguir:

Algoritmo 4.3.1

Inicialização:

$\boldsymbol{w}_0 = 0$

$\boldsymbol{P}(0) = \delta^{-1}I$

Estimação:

$\boldsymbol{z}(n) = \boldsymbol{P}(n-1)\boldsymbol{x}^*(n)$

$\boldsymbol{g}(n) = \frac{1-\lambda^{-1}\boldsymbol{P}(n-1)\boldsymbol{x}^*(n)}{1-\lambda^{-1}\boldsymbol{x}^T(n)\boldsymbol{P}(n-1)x^*(n)}$

$\alpha(n) = d(n) - \boldsymbol{w}_{n-1}^T\boldsymbol{x}_n$

$\boldsymbol{w}_n = \boldsymbol{w}_{n-1} + \alpha(n)\boldsymbol{g}(n)$

$\boldsymbol{P}(n) = \lambda^{-1}\left[\boldsymbol{P}(n-1) - g(n)\boldsymbol{x}^T(n)\boldsymbol{P}(n-1)\right]$

O algoritmo RLS em geral converge mais rápido do que o LMS. Por outro lado, sem o fator de esquecimento λ, o RLS pode ter desempenho inferior para processos não-estacionários. Isso se deve ao fato de que para $\lambda = 1$, todos os dados são igualmente ponderados na estimação das correlações. A seguir, serão apresentadas estruturas de modelagem e predição mais complexas do que as lineares, baseadas principalmente na lógica *fuzzy*.

4.4 Sistemas Fuzzy

A Teoria de Conjuntos Fuzzy foi concebida por L. A. Zadeh com o objetivo de traduzir em termos matemáticos as informações de caráter impreciso ou vago expressas por um conjunto de regras lingüísticas (Zadeh, 1965). Os sistemas *fuzzy* podem combinar de forma efetiva tanto a informação numérica como a lingüística para obter sistemas eficientes e aplicáveis a diversas áreas como economia, computação, engenharia, medicina, etc.

Os sistemas *fuzzy* ou nebulosos são estruturas não-lineares para as quais a Teoria de Conjuntos Fuzzy (uma generalização da teoria de conjuntos clássica) e a Lógica Fuzzy fornecem a base matemática para lidar com regras lingüísticas.

Os sistemas *fuzzy* apresentam portanto uma característica única que os distinguem dos demais: podem manipular simultaneamente tanto dados numéricos quanto informações provenientes de conhecimento lingüístico. Outra característica fundamental dos sistemas *fuzzy* é a capacidade de aproximação universal, isto é, a capacidade de aproximar com precisão arbitrária qualquer mapeamento não-linear contínuo definido sobre uma região compacta do domínio (Wang & Mendel, 1992).

4.4.1 Conjuntos Fuzzy

Um conjunto *fuzzy* A em um universo X é definido por uma função de pertinência $\mu_A(x) : X \rightarrow [0, 1]$ e representado por um conjunto de pares ordenados:

$$A = \{\mu_A(x), x\}, \qquad x \in X \tag{4.40}$$

onde $\mu_A(x)$ indica o quanto x é compatível com o conjunto A. Um determinado elemento pode pertencer a mais de um conjunto *fuzzy* com diferentes graus de pertinência.

Neste caso, a função de pertinência pode ser vista como uma medida do grau de similaridade entre os elementos do universo de discurso e o conjunto *fuzzy*. A Figura 4.2 ilustra a diferença entre as funções de pertinência para uma variável precisa ('crisp') e para uma *fuzzy*. As funções de pertinência mais usadas são as triangulares, trapezoidais, lineares por partes e gaussianas.

O formato e os parâmetros das funções de pertinência mais adequados são dependentes da aplicação, podendo ser escolhidos arbitrariamente, ou através de técnicas de otimização (Ross, 1997).

A exemplo do que ocorre com conjuntos ordinários, há uma série de definições e operações envolvendo conjuntos *fuzzy*. Sejam A e B dois conjuntos *fuzzy*, algumas dessas operações e propriedades são:

- Conjunto vazio: $\mu_A(x) = 0, \forall x \in X$;
- Igualdade de conjuntos: $A = B \Leftrightarrow \mu_A(x) = \mu_B(x)$;

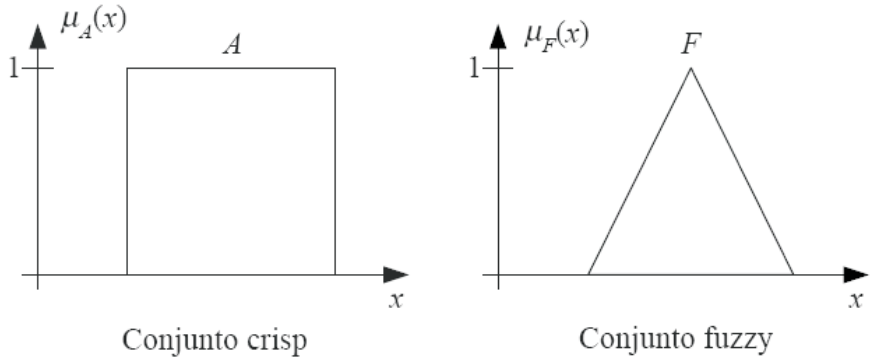


Fig. 4.2: Conjunto *crisp* e conjunto fuzzy

- Complemento: $\bar{\mu}_A(x) = 1 - \mu_A(x)$;
- Inclusão: $A \subseteq B \Leftrightarrow \mu_A(x) \leq \mu_B(x)$;
- União: $\mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\}$;
- Intersecção: $\mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\}$.

Com o objetivo de generalização, foram definidos operadores para a união e intersecção baseados nos conceitos de norma e co-norma triangular. As normas triangulares são modelos genéricos das operações de união e intersecção da teoria de conjuntos *fuzzy* e da conjunção e disjunção na lógica correspondente, devendo apresentar as propriedades de comutatividade, associatividade, monotonicidade e satisfazer as condições de contorno (Pedrycz & Gomide, 1998). As normas triangulares são chamadas de t-normas e s-normas, podendo ser formalmente definidas como:

Definição 4.4.1 *Uma t-norma é uma operação binária *:* $[0,1]^2 \rightarrow [0,1]$ *onde* $\forall x, y, z, w \in [0,1]$ *tal que as seguintes propriedades são satisfeitas:*

1)Comutatividade: $x * y = y * x$

2) Associatividade: $(x * y) * z = x * (y * z)$

3) Monotonicidade: se $x \leq y$, $w \leq z$, *então* $x * w \leq y * z$

4) Condições de contorno: $x * 0 = 0$ *e* $x * 1 = x$

Definição 4.4.2 *Uma s-norma, também conhecida como co-norma-t, é uma operação binária* $\oplus$*:* $[0,1]^2 \rightarrow [0,1]$ *onde* $\forall x, y, z, w \in [0,1]$ *tal que as seguintes propriedades são satisfeitas:*

1)Comutatividade: $x \oplus y = y \oplus x$

2) Associatividade: $(x \oplus y) \oplus z = x \oplus (y \oplus z)$

3) Monotonicidade: se $x \leq y$, $w \leq z$, *então* $x \oplus w \leq y \oplus z$

4) Condições de contorno: $x \oplus 0 = x$ *e* $x \oplus 1 = 1$

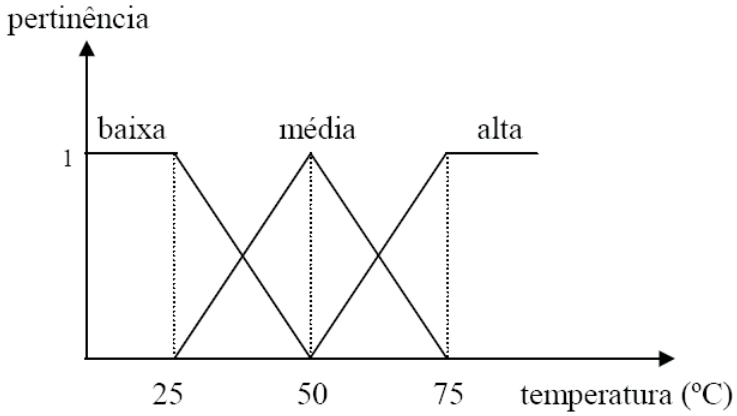


Fig. 4.3: Funções de pertinência para a variável temperatura

Pode-se observar claramente que t-normas incluem a operação *min* (intersecção padrão) e s-normas a operação *max* (união padrão). Várias normas-t e co-normas-t são encontradas na literatura, mas têm sido utilizados preponderantemente os operadores *min* (mínimo) e o produto algébrico para intersecção e o operador *max* (máximo) para a união (Pedrycz & Gomide, 1998).

4.4.2 Variáveis Lingüísticas

Uma variável lingüística é uma variável cujos valores são palavras ou sentenças representados por conjuntos *fuzzy*. Por exemplo, a temperatura de um determinado processo pode ser uma variável lingüística assumindo valores: baixa, média, e alta como mostra a Figura 4.3. Estes valores são descritos por intermédio de conjuntos *fuzzy*, representados por funções de pertinência.

A principal função das variáveis lingüísticas é fornecer uma maneira sistemática para uma caracterização aproximada de fenômenos complexos ou mal definidos.

4.4.3 Relações Fuzzy

No caso de conjuntos ordinários, uma relação exprime a presença ou a ausência de uma associação (ou interação) entre elementos de dois ou mais conjuntos. Relações *fuzzy* generalizam o conceito de relações e representam o grau de associação entre elementos de dois ou mais conjuntos *fuzzy*. Dados dois universos X e Y, a relação *fuzzy* R é um conjunto *fuzzy* definido no espaço resultante do produto cartesiano $X \times Y$, caracterizada por uma função de pertinência $\mu_R(x, y) \in [0, 1]$, onde $x \in X$ e $y \in Y$.

A composição de relações representa um papel muito importante em sistemas de inferência *fuzzy*. Sejam $R(U; V)$ e $S(V; W)$ duas relações *fuzzy* em $U \times V$ e $V \times W$, e que possuem um conjunto em

comum. A composição destas relações é um conjunto *fuzzy* definido como:

$$R \circ S \Rightarrow \mu_{R \circ S}(x, z) = \sup_{y \in V} [\mu_R(x, y) * \mu_S(y, z)] \tag{4.41}$$

onde $x \in U$, $y \in V$, $z \in W$ e a t-norma (representada por *) é normalmente o mínimo ou o produto, embora seja permitido usar outras t-normas. No caso de universos finitos, a operação *sup* é o máximo (Pedrycz & Gomide, 1998).

4.4.4 Lógica Fuzzy

Regras são expressas através de implicações lógicas na forma "*se ... então*", representando uma relação $R_{A \rightarrow B}$ entre um ou mais antecedentes e um ou mais conseqüentes. A função de pertinência associada a esta relação é definida por intermédio do operador de implicação $f_{\rightarrow}$, que deve ser escolhido apropriadamente. Os conceitos de lógica *fuzzy* nasceram inspirados na lógica tradicional, embora modificações tenham se tornado necessárias para adaptá-los aos requisitos de aplicações em engenharia.

A extensão da lógica tradicional para a lógica *fuzzy* foi efetuada através da simples substituição das funções características (ou funções de pertinência bivalentes) da primeira por funções de pertinência *fuzzy*, à semelhança da extensão de conjuntos ordinários para conjuntos *fuzzy*. Assim, a declaração condicional:

Se x é A, *então* y é B, $x \in X$ e $y \in Y$

equivale à implicação $A \rightarrow B$ e tem uma função de pertinência $\mu_{A \rightarrow B}(x, y)$ que mede o grau de veracidade da relação de implicação entre x e y. Sejam os conjuntos *fuzzy* A^*, B^*, A e B e a regra de inferência para lógica *fuzzy* dada da seguinte forma:

Premissa 1: x é A^*

Premissa 2: se x é A então y é B

Conseqüência: y é B^*

Na lógica *fuzzy*, uma regra será disparada se houver um grau de similaridade diferente de zero entre a Premissa 1 e o antecedente da regra (conjunto A); o resultado será um conseqüente com grau de similaridade não nulo em relação ao conseqüente da regra (conjunto B). Formalmente, a função de pertinência do conseqüente, $\mu_{B^*}(y)$, é obtida a partir do conceito de regra de inferência composicional $B^* = A^* \circ R$, na qual a conexão entre as duas proposições é representada explicitamente por uma relação R. Deste modo, para obter o conjunto *fuzzy* resultante da ativação de uma regra, ou seja, $\mu_{B^*}(y)$, deve-se calcular a composição entre um conjunto *fuzzy* $\mu_{A^*}(x)$ e a relação *fuzzy* $\mu_{A \rightarrow B}(x, y)$

dada por (Mendel, 1995):

$$\mu_{B^*}(y) = \sup_{x \in A^*} [\mu_{A^*}(x) * \mu_{A \to B}(x, y)]. \tag{4.42}$$

Uma vez definido o mecanismo através do qual são obtidos os conjuntos *fuzzy* resultantes da ativação das regras, resta saber como obter a função de pertinência da implicação $\mu_{A \to B}(x, y)$. De um modo geral, a implicação $\mu_{A \to B}(x, y)$ pode ser representada utilizando uma t-norma (Mendel, 1995) (Ross, 1997):

$$\mu_{A \to B}(x, y) = \mu_A(x) * \mu_B(y). \tag{4.43}$$

4.4.5 Sistema de Inferência Fuzzy

Na grande maioria das aplicações, as entradas para o sistema de inferência *fuzzy* são não-*fuzzy*, ou precisas, resultantes de medições ou observações. Em virtude disto, é necessário efetuar-se um mapeamento destes dados precisos para os conjuntos *fuzzy* relevantes, o que é realizado no estágio de fuzzificação ('fuzzificador') apresentado na Figura 4.4. Neste estágio ocorre também a ativação das regras relevantes para uma dada situação.

Uma vez obtido o conjunto *fuzzy* de saída através do processo de inferência, no estágio de defuzzificação é efetuada uma interpretação dessa informação. Isto se faz necessário pois na prática geralmente são requeridas saídas precisas.

Existem vários métodos de defuzzificação na literatura; dois dos mais empregados são o centro de gravidade e a média dos máximos (Ross, 1997). Neste último, a saída *crisp* é obtida tomando-se a média entre os dois elementos extremos no universo que corresponde aos maiores valores da função de pertinência do conseqüente. No método de centro de gravidade, a saída é o valor no universo que divide a área sob a curva da função de pertinência em duas partes iguais.

As regras podem ser fornecidas por especialistas, em forma de sentenças lingüísticas, e se constituem em um aspecto fundamental no desempenho de um sistema de inferência *fuzzy*. Alternativamente ao uso de especialistas para a definição da base de regras, existem métodos de extração de

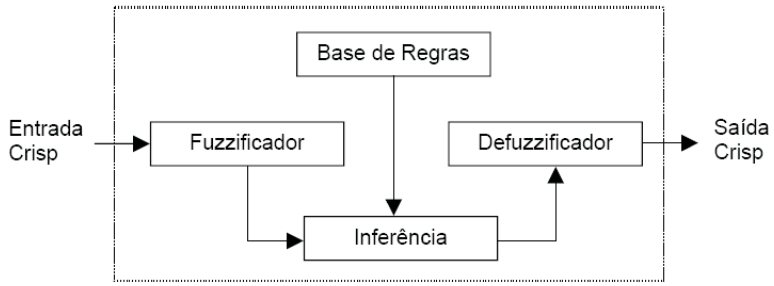


Fig. 4.4: Estrutura geral de um sistema de inferência *fuzzy*

regras de dados numéricos. Estes métodos são particularmente úteis em problemas de classificação e predição de séries temporais.

No estágio de inferência ocorrem as operações com conjuntos *fuzzy* propriamente ditas: combinação dos antecedentes das regras, implicação e relações. O conjunto *fuzzy* de entrada, relativos aos antecedentes das regras e o conjunto de saída, referente ao conseqüente, podem ser definidos previamente ou, alternativamente, gerados automaticamente a partir dos dados. Um aspecto importante é a definição dos conjuntos *fuzzy* correspondentes às variáveis de entrada (antecedentes) e às de saída (conseqüentes), pois o desempenho do sistema de inferência dependerá do número de conjuntos e de sua forma. Pode-se efetuar uma sintonia "manual" das funções de pertinência dos conjuntos, mas é mais comum empregarem-se métodos automáticos. A integração entre sistemas de inferência *fuzzy* e redes neurais, originando os sistemas neuro-*fuzzy* e os algoritmos genéticos tem se mostrado adequados para a sintonia de funções de pertinência, assim como para a geração automática de regras (Ross, 1997). A próxima seção, foca a modelagem *fuzzy* TSK (Takagi-Sugeno-Kang) a qual por suas interessantes características tem sido utilizada em diversas aplicações (Sugeno & Yasukawa, 1993)(Takagi & Sugeno, 1985)(Euntai et al., 1997).

4.5 Modelo Fuzzy TSK

A modelagem de tráfego de redes consiste na obtenção de uma estimativa $\hat{f}$ de uma função f a partir de um conjunto de observações correspondentes à série de tráfego

$$\{(\vec{x}(1), y_1), (\vec{x}(2), y_2), ..., (\vec{x}(N), y_N)\}, \tag{4.44}$$

com $\vec{x}(k) \in \mathbb{R}^n$ e $y_k \in \mathbb{R}$, onde N é o número de dados do treinamento, $\vec{x}(k) = [x_1(k), x_2(k), ..., x_n(k)]$ é o vetor da i-ésima entrada, e y_k é a saída desejada para a entrada $\vec{x}(k)$. Supõe-se que essas observações sejam obtidas por uma função desconhecida $y(k) = f(\vec{x}(k))$ no instante de tempo discreto k que relaciona os dados de entrada com a saída desejada. Idealmente, deseja-se construir uma função $\hat{f}$ que represente esta relação, que pode ser a predição de um valor futuro dada uma entrada atual.

Modelos *fuzzy* têm sido eficientemente empregados na obtenção da função $\hat{f}$, ou seja, na modelagem de sistemas, uma vez que são aproximadores universais. Estes modelos podem ser divididos em 3 classes: Modelos linguísticos (Modelos de Mamdani) (Yager & Filev, 1994), Modelos relacionais *fuzzy* (Yager & Filev, 1994) e Modelos do tipo Takagi-Sugeno-Kang (TSK) (Takagi & Sugeno, 1985). Este último sistema de inferência concebido por H. Takagi e M. Sugeno também denominado de modelo auto-regressivo *fuzzy*, é bastante difundido e difere do de Mamdani por exemplo, na parte do conseqüente, que é uma função linear das variáveis dos antecedentes, expressa da seguinte forma:

se x_1 é A_1 e x_2 é A_2 , então $z = f(x_1, x_2)$. A função f é, em geral, um polinômio e o sistema de inferência é geralmente referenciado em função do grau deste polinômio. Por exemplo: em um sistema de inferência Takagi-Sugeno-Kang de ordem zero, a saída z é uma constante. Essencialmente, em um sistema TSK deste tipo, o espaço não-linear é subdividido em várias regiões lineares, o que evidentemente facilita a modelagem de vários tipos de processos.

Os modelos *fuzzy* do tipo TSK ganharam bastante atenção por seu destacado desempenho. No modelo TSK original os usuários definiam subespaços *fuzzy* e os parâmetros conseqüentes eram obtidos por estimação através do filtro de Kalman. Várias alternativas para a modelagem das regras TSK foram propostas com intuito de ajustar também a parte de premissa (Yager & Filev, 1994)(Euntai et al., 1997) (Kroll, 1996). Como cada subespaço associado a uma função linear é usado para caracterizar uma regra *fuzzy*, supondo ter uma geometria simples no mapeamento entrada-saída (elipsóide), algoritmos de classificação *fuzzy* como o FCM (Fuzzy c-Means) são adequados para definir os subespaços *fuzzy* (Kroll, 1996). Entretanto, em geral as propostas que aplicam FCM são baseadas apenas na classificação ('clustering') do espaço de entrada dos dados de treinamento sem levar em conta o espaço de saída dos dados. Como alternativa, em (Euntai et al., 1997) os autores propõem identificar simultaneamente os subespaços *fuzzy* e as funções da parte conseqüente utilizando o algoritmo de classificação *fuzzy* FCRM (Fuzzy C-Regression Model), que será descrito na próxima seção.

Como já mencionado, na modelagem *fuzzy* TSK, as informações do processo são divididas em *clusters* (agrupamentos), onde cada *cluster* é descrito por um modelo local. O tráfego de redes pode ser representado pela combinação de modelos locais AR (Auto-Regressivos) via regras *fuzzy*, o que é feito no modelo *fuzzy* TSK (Chen et al., 2000)(Takagi & Sugeno, 1985). Tipicamente, um modelo *fuzzy* TSK consiste de regras *fuzzy* do tipo SE-ENTÃO que tem a seguinte forma:

$$\textbf{Regra } R^i : \begin{cases} \textbf{Se} \quad x_1 \text{ é } A_1^i(\vec{\theta_1^i}) \quad e \quad x_2 \text{ é } A_2^i(\vec{\theta_2^i}), \quad ... \quad , x_n \text{ é } A_n^i(\vec{\theta_n^i}) \\ \textbf{Então} \quad h^i = f_i(x_1, x_2, ..., x_n; \vec{a}^i) \end{cases} \tag{4.45}$$

para $i = 1, 2, ..., C$, onde C é o número total de regras, $A_j^i(\vec{\theta_j^i})$ é o conjunto *fuzzy* da *i*-ésima regra para a componente de entrada x_j com vetor de parâmetros $\overrightarrow{\theta_j^i}$, e $\vec{a}^i = (a_0^i, a_1^i, ..., a_n^i)$ é o conjunto de parâmetros das partes conseqüentes. O conseqüente de cada regra R^i, ou seja, a saída de cada regra nebulosa R^i é uma expressão funcional $h^i = f_i(x_1, x_2, ..., x_n)$. Neste trabalho, as funções de pertinência são consideradas gaussianas, por isso as representaremos por $A_j^i(\theta_{j1}^i, \theta_{j2}^i)$, ou seja, são descritas em função dos parâmetros θ_{j1}^i e θ_{j2}^i. Assim, estas funções de pertinência são dadas pela seguinte equação:

$$A_j^i(\theta_{j1}^i, \theta_{j2}^i) = \exp\left\{-\frac{(x_j - \theta_{j1}^i)^2}{2(\theta_{j2}^i)^2}\right\} \tag{4.46}$$

Logo, θ_{j1}^i e θ_{j2}^i são parâmetros ajustáveis da função de pertinência $A_j^i(\theta_{j1}^i, \theta_{j2}^i)$ da regra *fuzzy* i. A

modelagem *fuzzy* TSK faz uso das regras *fuzzy* R^i para estimação da saída do modelo y (que pode ser a predição de um valor futuro) pela seguinte equação:

$$\hat{y} = \frac{\sum_{i=1}^{C} h^i w^i}{\sum_{i=1}^{C} w^i} \tag{4.47}$$

onde h^i é a saída da i-ésima regra e $w^i = \prod_{j=1...n} A_j^i(\theta_{j1}^i, \theta_{j2}^i)$ é o ativador da i-ésima regra. Pela análise da equação (4.47), nota-se que o modelo *fuzzy* TSK é realmente equivalente a uma combinação não-linear de modelos auto-regressivos locais. A Figura 4.5 exemplifica o funcionamento do modelo *fuzzy* TSK em analogia à estrutura conexionista de uma rede neural, o que motiva vários autores de a chamarem de modelo neuro-*fuzzy*.

Na modelagem TSK, os parâmetros das partes de premissa (isto é, θ_{j1}^i e θ_{j2}^i) e das partes conseqüentes ($\vec{a}^i$) devem ser determinados. Para identificar os parâmetros conseqüentes ($\vec{a}^i$), as regras mais adequadas para os dados de entrada devem ser encontradas. Para isto, o modelo de regressão nebuloso de classificação (FCRM), que é uma versão modificada do FCM pode ser aplicado, desenvolvendo *clusters* com formas hiperplanas (Chen et al., 2000). Entretanto, o FCRM identifica os parâmetros do modelo nebuloso de modo aproximado, assim um procedimento de ajuste fino é em seguida aplicado baseado em algoritmo de gradiente descendente. Ambos procedimentos, de ajuste aproximado e fino, podem ser repetidos para achar um número apropriado de *clusters*. Será mostrado como determinar os parâmetros conseqüentes ótimos a_j^i e os parâmetros de premissa θ_{j1}^i e θ_{j2}^i a fim de minimizar um índice de desempenho, tomado como sendo o EQMN (erro quadrático médio normalizado) de predição do modelo *fuzzy* (Chen et al., 2000).

4.5.1 Ajuste Aproximado pelo Algoritmo FCRM

O modelo AR linear local no *i*-ésimo *cluster* do modelo TSK representado por h^i pode ser reescrito na seguinte forma matricial:

$$h^i(k+1) = Z^T(k)\mathbf{A}_i \tag{4.48}$$

onde $Z(k) = [1, y(k), y(k-1), ..., y(k-j+1)]^T$ e $\mathbf{A}_i = [a_{i,0}, a_{i,1}, \ldots, a_{i,n}]^T$, sendo n a ordem deste modelo AR local. Em resumo, segundo o algoritmo FCRM, o vetor de parâmetros $\mathbf{A}_i$ do modelo AR *fuzzy* pode ser obtido de forma aproximada pelos seguintes passos:

Algoritmo 4.5.1 *Algoritmo FCRM*

Passo 1) Inicie com o contador de iteração $M = 0$. Defina uma matriz U de dimensão CxN onde C é o número total de regras e N é o número total de amostras da série, da seguinte forma (Kim et al., 1997):

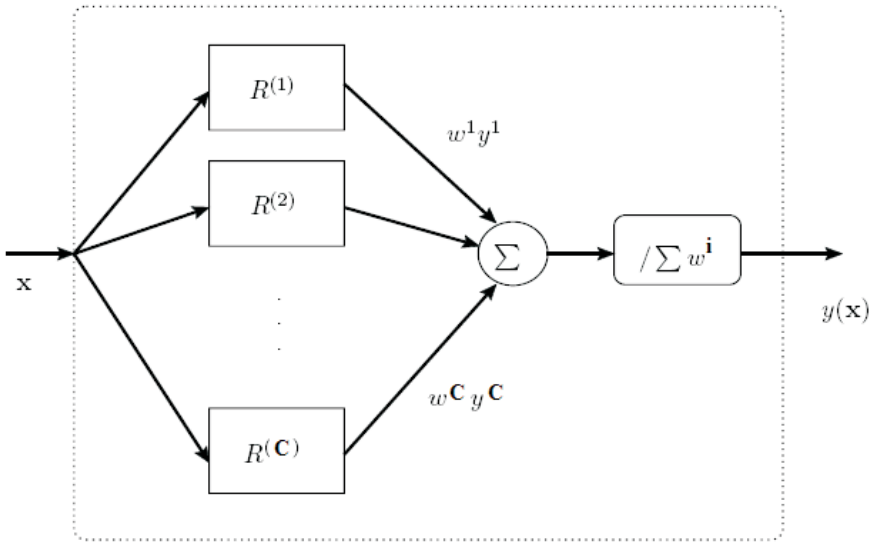


Fig. 4.5: Sistema fuzzy do tipo TSK.

$$U = \begin{bmatrix} u_{1,1} & u_{1,2} & \ldots & u_{1,N} \\ u_{2,1} & u_{2,2} & \ldots & u_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ u_{C,1} & u_{C,2} & \ldots & u_{C,N} \end{bmatrix}$$

onde

$$0 \leq u_{i,k} \leq 1, 1 \leq i \leq C, 1 \leq k \leq N \tag{4.49}$$

e

$$\sum_{i=1}^{C} u_{i,k} = 1, \forall k = 1, ..., N \tag{4.50}$$

A variável $u_{i,k}$ deve ser tal que satisfaça as restrições (4.49) e (4.50).

Passo 2) Na M-ésima iteração a função custo do algoritmo FCRM é dada por:

$$J = \sum_{i=1}^{C} \sum_{k=1}^{N} u_{i,k}^{m} d_{i,k}^{2} \tag{4.51}$$

onde $d_{i,k} = \left\| y(k) - Z^T(k-1)\boldsymbol{A}_i \right\|$. A condição necessária para que a equação (4.51) atinja seu mínimo é que (Chen et al., 2000):

$$u_{i,k} = \frac{1}{\sum_{i=1}^{C} \left(\frac{d_{i,k}}{d_{k,i}} \right)^{\frac{2}{n-1}}} \tag{4.52}$$

Portanto, calcula-se uma nova matriz U pela equação (4.52).

Passo 3) Se a função custo for menor do que um certo valor o algoritmo é finalizado, senão vá para o passo 4.

*Passo 4) Usando os $u_{i,k}$'s obtidos no passo 2, calcule o vetor de parâmetros $\boldsymbol{A}_i(k+1)$ pelo algoritmo de mínimos quadrados recursivos ponderado (WRLS-*Weighed Recursive Least Squares*) descrito abaixo:*

$$\boldsymbol{A}_i(k+1) = \boldsymbol{A}_i(k) + H(k)[y(k+1) - Z^T(k)\boldsymbol{A}_i(k)] \tag{4.53}$$

$$H(k) = \frac{S(k)Z(k)}{\frac{1}{u_{l.k}} + Z^T(k)S(k)Z(k)} \tag{4.54}$$

$$S(k+1) = (I - H(k)Z^T(k)).S(k) \tag{4.55}$$

onde k=1,2,..., N e i=1,2,..., C. No início do algoritmo WRLS, usa-se $S(0) = \alpha I$ e $\alpha > 100$, onde I é a matriz identidade.

Passo 5) Vá ao passo 2 e incremente M de 1. Se o algoritmo tiver atingido a iteração desejada pare.

Realizados esses passos, pode-se obter estimativas de θ^i_{j1} e θ^i_{j2} para as funções de pertinência gaussianas pelas equações (Kim et al., 1997):

$$\theta^i_{j1} = \frac{\sum_{k=1}^{N} u_{i,k} y(k-j)}{\sum_{k=1}^{N} u_{i,k}} \tag{4.56}$$

$$\theta^i_{j2} = \sqrt{2.\frac{\sum_{k=1}^{N} u_{i,k}(y(k-j) - \theta^i_{j1})^2}{\sum_{k=1}^{N} u_{i,k}}} \tag{4.57}$$

4.5.2 Ajuste Fino utilizando o Algoritmo de Gradiente Descendente

Os parâmetros conseqüentes ótimos representados pelo vetor $\mathbf{A}_i$ podem ser determinados pelo método de mínimos quadrados que é aplicado em problemas de estimação linear. Já, a determinação dos parâmetros θ^i_j's ótimos é um problema de estimação não-linear. Embora sejam problemas de estimação diferentes, o algoritmo de gradiente descendente pode ser usado para refinar tanto os parâmetros conseqüentes quanto os de premissa:

1. Procedimento de ajuste dos parâmetros de premissa:

Os parâmetros de premissa θ^i_{jq}'s ($q = 1, 2$) do modelo AR *fuzzy* podem ser ajustados pela equação:

$$\Delta\theta^i_{jq}(k+1) = \eta(y(k) - \hat{y}(k)).(y_i(k) - \hat{y}(k)).\frac{1}{\sum_{i=1}^{C} w^i}\frac{\partial w^i}{\partial \theta^i_{jq}}, \tag{4.58}$$

onde η é uma taxa de aprendizagem, *y(k)* é a intensidade de tráfego atual, $\hat{y}(k)$ é a saída do modelo nebuloso e $e(k) = y(k) - \hat{y}(k)$, o erro de predição do modelo.

2. Procedimento de ajuste dos parâmetros conseqüentes:

Os parâmetros conseqüentes $\mathbf{A}_i = [a_{i,0}, a_{i,1}, \ldots a_{i,n}]^T$ do modelo AR nebuloso de tráfego são ajustados de forma precisa pela equação:

$$\Delta a_{i,j}(k+1) = \gamma(y(k) - \hat{y}(k)).w^i\frac{y(k-j)}{\sum_{i=1}^{C} w^i}, \tag{4.59}$$

onde γ é uma outra taxa de aprendizagem.

Uma vez que os parâmetros θ^i_{jq}'s e $a_{i,j}$'s foram estimados a partir do tráfego real uma predição a um passo $\hat{y}(n+1)$ pode ser obtida pelos dados de tráfego em instantes anteriores utilizando as seguintes equações:

$$\hat{y}(n+1) = \frac{\sum_{l=1}^{c} w^i h^i(n+1)}{\sum_{l=1}^{c} w^i}, \tag{4.60}$$

$$w^i = \prod_{j=1\ldots n} A^i_j(\theta^i_{j1}, \theta^i_{j2}) \tag{4.61}$$

Inicialmente o ajuste aproximado realizado pelo algoritmo FCRM tem o papel mais importante na estimação dos parâmetros do modelo. Com a convergência do cálculo dos parâmetros do modelo, pode-se aplicar simplesmente o procedimento de ajuste fino (Chen et al., 2000). Ao se atualizar o modelo apenas com o procedimento de ajuste fino que é baseado no algoritmo de gradiente descendente, se diminui o custo computacional de treinamento do modelo. A modelagem auto-regressiva nebulosa tem convergência excelente sendo capaz de predizer o tráfego de redes como será demonstrado na próxima seção e segundo os trabalhos de Vieira et al. (Vieira & Lee, 2004a)(Vieira et al., 2004a) (Vieira et al., 2004b).

4.6 Resultados Experimentais

A análise da precisão de algoritmos de predição de tráfego de redes é importante porque estes algoritmos podem ser usados no controle de tráfego e um valor mais exato da taxa necessária a ser alocada pode evitar subestimação e ou superestimação de recursos da rede. Nesta seção, verifica-se que os erros de predição da modelagem *fuzzy* TSK são comparáveis aos de outros métodos tais como redes neurais (Vieira et al., 2003a). Para tal verificação, seja o erro quadrático médio normalizado definido a seguir.

Definição 4.6.1 - *Seja σ_x^2 a variância do processo X, dada por $\sigma_x^2 = E\left[(\mu - x)^2\right]$ onde μ é a média do processo. Define-se o erro quadrático médio normalizado do tipo 1 como*

$$EQMN1 = \frac{EQM}{\sigma_x^2} = \frac{E\left[(\hat{x} - x)^2\right]}{E\left[(\mu - x)^2\right]}. \tag{4.62}$$

Mais especificamente, o erro quadrático médio normalizado do tipo 1 pode ser escrito também da seguinte forma:

$$EQMN1 = \frac{1}{\sigma^2 p}\sum_{n=1}^{p}[x(n) - \hat{x}(n)]^2, \tag{4.63}$$

onde *x(n)* é o valor real da série de tráfego a ser predita, $\hat{x}(n)$ é o valor predito, σ^2 é a variância da série real sob o intervalo de predição e *p* é o número de amostras de teste.

Com intuito de demonstrar que a modelagem auto-regressiva *fuzzy* prediz de forma satisfatória amostras futuras do tráfego de rede, utilizou-se nos teste de predição, os traços de tráfego TCP/IP dec-pkt-1.tcp e dec-pkt-2.tcp. Considerou-se 2048 amostras de tráfego em uma escala de agregação de 512ms para estes traços de tráfego TCP/IP. Também foram utilizados traços de tráfego Ethernet obtidos da Bellcore que apresentam características auto-similares e multifractais. Estes traços são: Bc-Octext na escala de tempo de agregação de 1min com 2046 pontos e o traço de tráfego Bc-Octint com 1759 amostras na escala de tempo de 1s.

Foram realizadas predições a um passo destas séries de tráfego usando 2 regras nebulosas e 5 coeficientes $a_{i,j}$, $j = 1, 2 \ldots 5$. Estabeleceu-se para as taxas de aprendizagem η e γ os valores 0,001 e 0,01, respectivamente. A predição efetuada pela modelagem *fuzzy* TSK consiste de uma fase de treinamento com uma quantidade de amostras estipulada das séries e em seguida, aplica-se a estrutura obtida para efetuar a predição a um passo dessas mesmas amostras, conforme representado na Figura 4.6. A Tabela 4.1 mostra o EQMN1 de predição para estas séries temporais e o intervalo considerado no cálculo das predições. A Figura 4.7 ilustra a predição a um passo para a série Bc-Octint realizada pelo modelo *fuzzy* TSK treinado com o algoritmo FCRM e posteriormente com o algoritmo de gradiente descendente. O mesmo é mostrado na Figura 4.8, mas para série dec-pkt-2

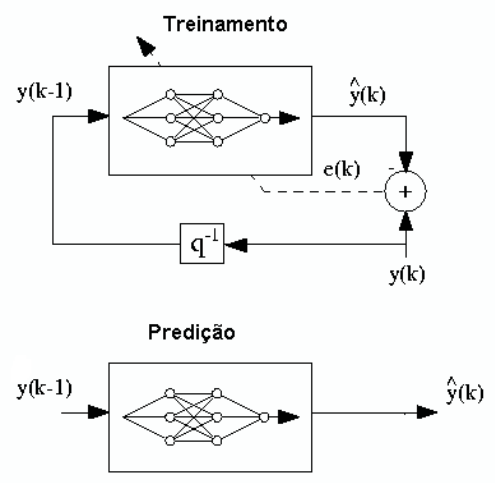


Fig. 4.6: Fases de treinamento e teste do modelo *fuzzy*

Tab. 4.1: Comparação entre EQMN1

Traço de Tráfego	Intervalo	EQMN1
Bc-Octint	801–1701	0,3107
Bc-Octext	1000-2000	0,4152
Dec-pkt-1	1–2048	0,6987
Dec-pkt-2	1-2048	0,5836

onde um passo de predição equivale a 512ms.

A Tabela 4.2 faz uma comparação entre os valores de EQMN obtidos utilizando a modelagem auto-regressiva nebulosa, uma rede neural MLP (Multilayer Perceptron) e uma rede recorrente treinada com algoritmo RTRL (Real Time Recurrent Learning) (Vieira et al., 2003a). Pode-se notar que a modelagem auto-regressiva nebulosa obtém erros de predição comparáveis aos métodos citados. Dessa forma, os erros quadráticos médios obtidos com o modelo preditor *fuzzy* treinado com o algoritmo FCRM são realmente satisfatórios.

Tab. 4.2: EQMN1

Traço de Tráfego	Intervalo	EQMN-*Fuzzy*	EQMN-MLP	EQMN-RTRL
Bc-Octint	801–1701	0,3107	1,21	0,3946
Bc-Octext	1000-2000	0,4152	0,4037	0,3850

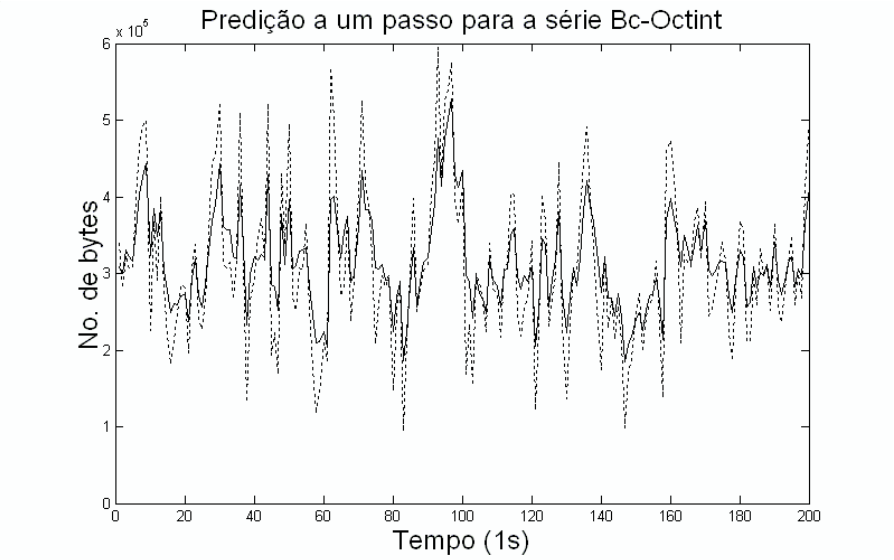


Fig. 4.7: Predição a um passo pela modelagem nebulosa (linha sólida). Série temporal de tráfego Bc-Octint (linha pontilhada)

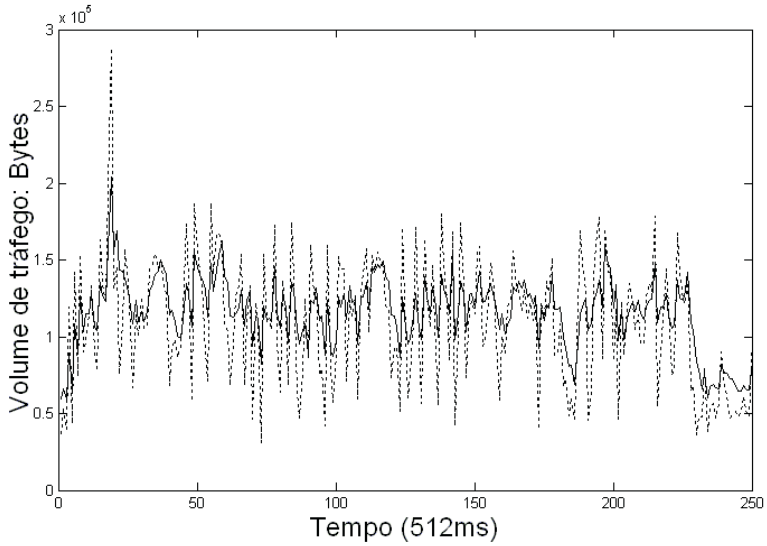


Fig. 4.8: Predição a um passo pelo modelo nebuloso (linha sólida). Traço de tráfego Dec-pkt-2 (linha pontilhada)

4.7 Considerações Finais

Neste capítulo foram apresentados, de modo sucinto, os conceitos fundamentais ligados à predição de séries temporais. Primeiramente, os preditores lineares representados pelo algoritmo LMS e RLS foram descritos. Ao se considerar preditores não-lineares, ênfase foi dada aos modelos *fuzzy* e seus conceitos como conjuntos *fuzzy*, lógica *fuzzy* e mecanismos de inferência.

Após os trabalhos iniciais de Mamdani, surgiram inúmeras outras aplicações da lógica *fuzzy*, inclusive em escala industrial (Bezdek, 1993). A utilidade da teoria de conjuntos *fuzzy* em reconhecimento de padrões e análise de *clusters* foi reconhecida desde os anos 60. Atualmente a literatura a respeito é bastante vasta e a modelagem *fuzzy* TSK tem destaque. Entretanto, diferentemente dos algoritmos de predição LMS e RLS, esta modelagem, da forma como foi introduzida, não é adaptativa. Ou seja, a criação de regras e o ajuste dos parâmetros são realizados com o uso de toda a série a ser analisada. No próximo capítulo, é proposto um modelo preditor *fuzzy* baseado no modelo *fuzzy* TSK, entretanto com característica adaptativa. Em outras palavras, a cada nova amostra de tráfego disponível, os parâmetros do modelo proposto são ajustados, não necessitando de acumular amostras e fazendo com que o ajuste dos parâmetros seja feito dinamicamente.

Capítulo 5

Preditor Fuzzy Adaptativo com Funções de Base Ortonormais Baseadas na Autocorrelação do MMW

5.1 Introdução

Em mecanismos dinâmicos e preditivos de controle de tráfego, o preditor de tráfego desempenha um papel fundamental. Nestes casos, o desempenho dos mecanismos de controle está intimamente ligado à precisão dos preditores de tráfego neles empregados. A utilização de modelos estatísticos no projeto de preditores de tráfego pode proporcionar melhor desempenho de predição, sendo este desempenho fortemente dependente de quão preciso é o modelo em capturar as reais propriedades estatísticas do tráfego. Devido às características multifractais do tráfego de redes, sua predição é uma tarefa desafiadora, porém deve ser precisa para evitar estimação errônea do desempenho da rede (Sang & Li, 2002).

Redes neurais podem aprender com os dados, mas a interpretação do significado associado a cada neurônio e a cada peso não é trivial. Diferente das regras *fuzzy*, cujo processo de inferência é compreensível. Devido as suas características complementares, os conceitos de redes neurais e sistemas *fuzzy* têm sido mesclados, formando redes neuro-*fuzzy* (Ross, 1997). Em modelagem neuro-*fuzzy* há dois importantes problemas a serem lidados: identificação da estrutura e dos parâmetros do modelo. O primeiro está relacionado com a construção da estrutura *fuzzy* inicial e o outro com o refinamento das regras *fuzzy* com algoritmos de aprendizagem.

Entre os modelos *fuzzy* propostos, particularmente o modelo nebuloso descrito em (Kim et al., 1997) é capaz de descrever um determinado sistema desconhecido com um pequeno número de regras nebulosas como o modelo de Takagi e Sugeno (Takagi & Sugeno, 1985) e é de fácil de implementação

como o modelo de Sugeno e Yasukawa (Sugeno & Yasukawa, 1993). Porém, este modelo *fuzzy* não é adaptativo, ou seja, para modelagem de uma série de dados são exigidos todos os pontos da série temporal. Tendo em vista isso, este capítulo é dedicado ao desenvolvimento e validação de um modelo preditor adaptativo *fuzzy* com funções de base ortonormais (FBO). É proposto um novo algoritmo de treinamento para este modelo denominado de algoritmo ARFA (Agrupamento Regressivo Fuzzy Adaptativo) que cria adaptativamente agrupamentos *fuzzy* à medida que dados de tráfego de entrada são disponibilizados. O algoritmo ARFA é capaz de prever eficientemente intensidades do tráfego de redes, como será demonstrado pelas simulações deste capítulo.

Muitos esquemas de alocação de banda citados na literatura fazem simplificações sobre a dinâmica do tráfego de redes e se baseiam no controle reativo. Por exemplo, Groskinsky et. al (Groskinsky et al., 2001) investigam esquemas adaptativos de controle de capacidade do servidor usando um modelo de fluído para os fluxos e, Wang et al. (Wang et al., 1996) apresentam vários esquemas de alocação de banda para fluxos de tráfego heterogêneos. Entretanto, algumas propostas de predição de banda baseadas em redes neurais e *fuzzy* voltadas ao controle preditivo merecem atenção pois mostram que este tipo de controle pode ser bastante eficiente, uma vez que tenta reduzir o congestionamento antes que este aconteça e se adaptam bem às variações bruscas do tráfego de redes (Zhang, Wu & H.Xi, Zhang et al.)(Aquino & Barria, 2006).

Neste capítulo, é apresentado um novo esquema de alocação de banda usando os valores preditos obtidos pelo preditor *fuzzy* proposto que usa funções de base ortonormais calculadas a partir da função de autocorrelação do MMW. Este esquema de alocação de banda possui duas partes, uma relacionada à predição de tráfego e a outra, baseada no resultado de predição, à provisão de banda propriamente dita. O capítulo está organizado da seguinte forma. A seção 5.2 introduz os conceitos ligados ao algoritmo de controle de malha aberta baseado em predição. As seções 5.3 e 5.4 apresentam as bases para o desenvolvimento do nosso modelo. Mais especificamente, na seção 5.5, apresenta-se o modelo preditor *fuzzy*-FBO e o algoritmo de treinamento adaptativo ARFA e na subseção 5.5.1 é feita uma comparação de desempenho de predição do modelo proposto com outros métodos existentes na literatura. Na seção 5.6, o preditor de tráfego proposto é empregado em um esquema de alocação de banda, onde é verificado sua eficiência em relação a outros esquemas que também usam predição em sua composição. Por fim, na seção 5.7, as conclusões obtidas são apresentadas.

5.2 Controle Adaptativo de Taxa por Predição

Embora a maior parte do tráfego Internet ainda seja de melhor-esforço, aplicações emergentes com restrições de QoS estão se tornando cada vezes mais comuns. Um componente que falta nas arquiteturas propostas como o Diffserv e MPLS, é a capacidade de garantir parâmetros quantitativos

de QoS para tráfego agregado. Garantias de QoS têm tipicamente sido realizadas por métodos de alocação de banda estáticos que assumem algumas propriedades estocásticas para os fluxos de tráfego (Kelly, 1996)(Kim & Shroff, 2001) (Guérin et al., 1991). Entretanto, tal alocação estática é ineficiente para tráfego agregado de rede porque:

1) O modelo de tráfego pode não capturar precisamente o comportamento estatístico do tráfego;

2) A agregação de fluxos de tráfego com diferentes estatísticas pode resultar em outro fluxo de tráfego com características desconhecidas;

3) Mesmo conhecendo as características do tráfego, estas podem ser alteradas ao passar por multiplexadores e *buffers* na rede;

4) Os parâmetros declarados pelo usuários podem não representar com precisão o tráfego real;

Uma alternativa à alocação de banda estática para atender QoS por nó com eficiência é o controle adaptativo de banda (CAB) (Ryu, 2003). Algoritmos CAB ajustam dinamicamente a banda alocada para assegurar que esta seja suficiente para atender aos requisitos impostos. Esses algoritmos se baseiam em medidas do estado da rede e medidas de informação do tráfego e, principalmente, em predição de tráfego. Como resultado, menos banda é gasta devido à sobrealocação, o que aumenta a utilização da rede. Estes esquemas de alocação de banda têm as seguintes características: o impacto no desempenho devido à imprecisão dos parâmetros de tráfego declarados pelos usuários será reduzido porque o controle se adapta para as condições reais do tráfego e não é necessária uma parametrização a priori dos fluxos de tráfego.

De acordo com a técnica de controle utilizada, os esquemas de controle adaptativo de banda (CAB) podem ser classificados como sendo de malha aberta e malha fechada. O controle em malha fechada ou controle com realimentação surge naturalmente no contexto onde perda de pacote, tamanho médio da fila, ou outros estados do sistema são regularmente observados para prover realimentação para o ajuste da banda alocada. Estes esquemas de controle podem ser classificados também de acordo com o parâmetro de QoS garantido, o que inclui tamanho médio de fila (Pitsillides et al., 2001), perda (Siripongwutikorn et al., 2002) e retardo (Kesidis, 1999). Quanto ao tamanho médio da fila se inclui os esquemas de gerenciamento ativo de fila (AQM-Active Queue Management), onde o roteador intencionalmente descarta pacotes com uma probabilidade que aumenta com o tamanho médio da fila (Fengyuan, 2002). Por outro lado, o controle de malha aberta envolve a predição da taxa de tráfego de entrada usando amostras passadas. A taxa de serviço do servidor é então ajustada ao valor da taxa predita para obter perda zero de pacote ou baixo retardo. Conseqüentemente, a maioria dos trabalhos existentes de controle de malha aberta busca atender a esses objetivos ao invés de garantir QoS quantitativamente (Chong et al., 1995)(Adas, 1998) (Duffield et al., 1999).

O controle adaptativo de banda (CAB) é atrativo particularmente para tráfego de vídeo, cuja principal característica é que a sua taxa de dados pode variar significativamente de um quadro para outro

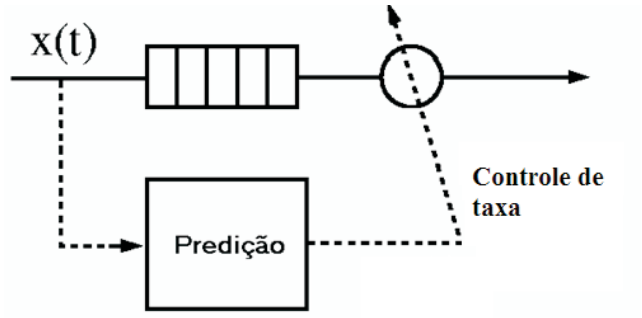


Fig. 5.1: Controle de taxa por predição

(Fitzek & Reisslein, 2000). Vários artigos foram escritos sobre transmissão de vídeo VBR (Variable Bit Rate) em redes ATM com garantia de QoS, sendo constatado que sua estrutura de autocorrelação possui decaimento lento (Beran et al., 1995)(Pancha & Eizarki, 1994). Portanto, fluxos de vídeo podem exigir um controle mais elaborado, como por exemplo, controle do tipo CAB.

Este capítulo se concentra em um controle adaptativo de taxa em malha aberta do tipo perda zero baseado em predição de tráfego usando um modelo *fuzzy*. O preditor *fuzzy* adaptativo desenvolvido neste capítulo é aplicável não só para tráfego com determinadas características, mas qualquer tipo de tráfego de redes, uma vez que este não requer o conhecimento a priori da estatística dos dados. Isso é desejável porque muitos trabalhos têm mostrado que o tráfego de redes tem comportamento não-linear, não-estacionário e complexo (Neves et al., 1995)(Hiramatsu, 1990). Devido a essas características, a alocação dinâmica de recursos em redes é interessante (Chong et al., 1995)(Adas, 1996). A idéia do controle de malha aberta adaptativo é estimar a taxa mínima para garantir a transmissão completa do tráfego em intervalos regulares (Figura 5.1). A taxa de serviço é dinamicamente ajustada baseada nesta estimativa. O uso de uma taxa fixa de serviço para tráfego com cauda pesada e altamente correlacionado pode causar um número excessivo de perda de células e pacotes se esta taxa não estiver próxima ao pico de tráfego (Adas & Mukerjee, 1995). Esta tese trata apenas de algoritmos para disponibilizar recursos (capacidade do servidor) ao invés de ajustar a taxa da fonte, como pode ser feito ajustando a taxa de vídeo comprimido pela mudança nos parâmetros do codificador de vídeo. Importantes resultados sobre controle da taxa da fonte podem ser encontrados nos artigos de Sousa et al. (Sousa et al., 2006a) (Sousa et al., 2006b).

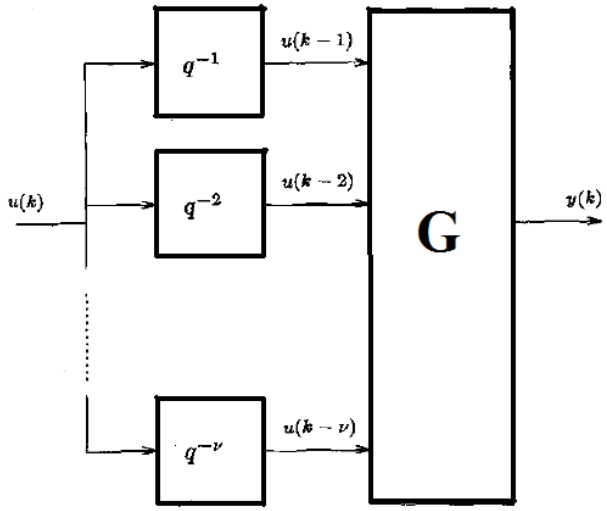


Fig. 5.2: Diagrama de bloco de um modelo não-linear de média móvel

5.3 Funções de Base Ortonormais em Modelagem *Fuzzy*

Modelos *fuzzy* dinâmicos se baseiam em geral na topologia do modelo NARX (Non-linear Auto-Regressive with eXogeneuos input) onde as saídas são escritas em função das entradas e saídas passadas (Brockwell & Davis, 1996)(Oliveira et al., 1999). No modelo NARX há a realimentação dos erros de saída, o que pode degradar a precisão da predição, especialmente para horizontes maiores de predição. Uma estratégia eficiente de resolver esse problema se baseia no uso de funções de base ortonormais. Funções ortonormais como as de Laguerre e Kautz têm sido utilizadas na modelagem linear e não-linear de sistemas (Oliveira et al., 1999). Modelos baseados em tais funções podem ser vistos como uma generalização de modelos baseados em resposta ao impulso. Entretanto, o primeiro é capaz de descrever um sistema dinâmico com um número menor de parâmetros.

No caso de sistemas não-lineares com memória finita, podemos representar a relação entrada-saída através de um modelo não-linear de média móvel (NLMA- Non-Linear Moving Average) como (Oliveira et al., 1999):

$$y(k) = G(u(k-1), \ldots, u(k-v)), \tag{5.1}$$

onde $u(k)$ corresponde ao processo de entrada no instante k.

O sinal de entrada do modelo (5.1) pode ser considerado como o desenvolvimento da seqüência $u(k)$ em uma base como ilustram as Figuras 5.2 e 5.3, tal que o i-ésimo componente de entrada seja dado por

$$u_i(k) = \Phi_{lag,i}(q^{-1})u(k) = q^{-i}u(k) = u(k-i), \tag{5.2}$$

onde q^{-i} é o operador de deslocamento com $\Phi_{lag,i}(q^{-1}) = q^{-i}$.

No modelo NLMA, a saída de um sistema é modelada com arbitrária precisão de acordo com o número de funções na base. Entretanto, pode-se ter um modelo com um número elevado de funções. Uma maneira de reduzir o número de funções de base para um determinado erro é inserindo conhecimento a priori da dinâmica do sistema na base. As funções de base ortonormais de Laguerre e de Kautz têm sido freqüentemente utilizadas com esse intuito (Ninness & Gustafsson, 1995) (Wahlberg & Ljung, 1992).

A base de Laguerre é usada em vários contextos de identificação e controle de sistemas não-lineares (Oliveira et al., 1999) (Dumont & Fu, 1993). Neste estudo, considera-se apenas a base de Laguerre, especialmente porque esta é completamente descrita por um único pólo (pólo de Laguerre), tornando mais simples a obtenção das funções de base ortonormais. O conjunto de funções de transferência associadas a esta base é dada pela seguinte equação:

$$\Phi_{lag,i}(q^{-1}) = \sqrt{1-p^2}\frac{q^{-1}(q^{-1}-p)^{i-1}}{(1-pq^{-1})^i}, \qquad i = 1, \ldots \tag{5.3}$$

onde $p \in \{R : -1 < p < 1\}$ é o pólo das funções de Laguerre. Pode-se notar que fazendo $p = 0$, resulta na base $\Phi_{lag,i}(q^{-1}) = q^{-i}$. Portanto, a base $\Phi_{lag,i}(q^{-1})$ do modelo é um caso especial da base de Laguerre.

A saída do modelo expressa pela equação (5.1), pode ser reescrita como

$$y(k) = H(l_i(k), \ldots, l_n(k)), \tag{5.4}$$

onde $l_i(k) = \Phi_{lag,i}(q^{-1})u(k)$ é a saída da i-ésima função de Laguerre no instante de tempo k, n é o número de funções utilizadas e H é um operador não-linear. O diagrama em bloco deste modelo poder ser visto na Figura 5.3.

As funções de Laguerre são recursivas, ou seja, a i-ésima função pode ser escrita usando a $(i-1)$-ésima função. Os sinais $l_i(k)$ podem ser obtidos por equações de estado da seguinte forma (Oliveira et al., 1999):

$$\mathbf{l}(k+1) = A\mathbf{l}(k) + b\mathbf{u}(k) \tag{5.5}$$

$$y(k) = H(\mathbf{l}(k)) \tag{5.6}$$

onde $\mathbf{l}(k) = [l_i(k) \ldots l_n(k)]^T$. A matriz A e o vetor b dependem da ordem do modelo n e do valor do

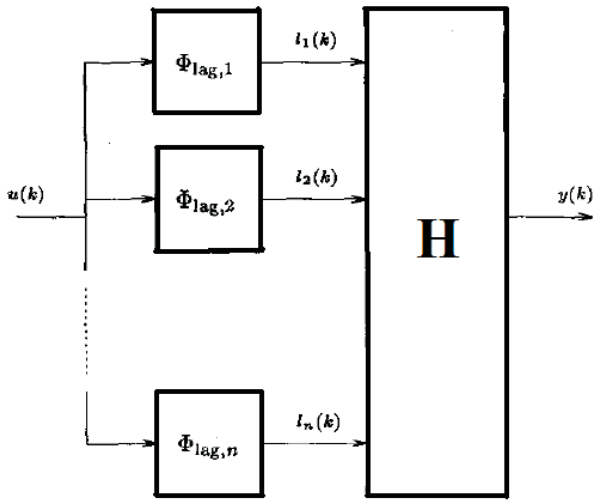


Fig. 5.3: Diagrama de bloco de um modelo não-linear de bases ortonormais

pólo p como segue:

$$A = \begin{pmatrix} p & 0 & 0 & \dots & 0 \\ 1-p^2 & p & 0 & \dots & 0 \\ (-p)(1-p^2) & 1-p^2 & p & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (-p)^{n-2}(1-p^2) & (-p)^{n-3}(1-p^2) & (-p)^{n-4}(1-p^2) & \dots & p \end{pmatrix} \tag{5.7}$$

$$b = \sqrt{(1-p^2)}[\ 1 \quad -p \quad (-p)^2 \quad \dots \quad (-p)^{n-1}\ ]^T \tag{5.8}$$

O modelo não-linear representado pelas equações (5.5) e (5.6) constitui um mapeamento linear entre a entrada $u(k)$ e as funções de Laguerre $l_i(k)$, mais o mapeamento entre $l_i(k)$ e a saída do sistema $y(k)$, o que é exemplificado pela Figura 5.4. Obtém-se dessa forma, um modelo mais preciso do que o modelo NLMA tradicional com o mesmo número de funções.

Dado um número de funções de base n (ordem do modelo), um valor adequado do pólo p da base ortonormal acarreta uma melhor representação do sistema a ser modelado. Em geral, o pólo p é selecionado usando conhecimento a priori da dinâmica dominante do sistema. Em relação ao número de funções n, a escolha ideal é aquela que faz o erro de modelagem tender a zero. Na prática, esta escolha depende da complexidade do sistema; sistemas mais complexos exigem mais funções. Na próxima seção, são utilizadas propriedades do MMW na proposição de um pólo adequado para as funções de base ortonormais.

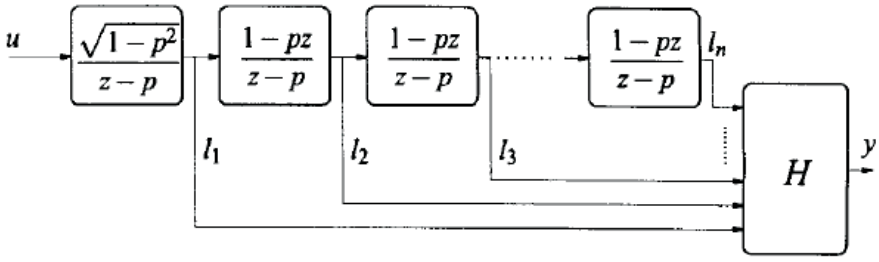


Fig. 5.4: Modelo OBF com dinâmica de Laguerre

5.4 Pólo baseado na Função de Autocorrelação do MMW

A modelagem *fuzzy* pode ser aplicada na implementação do operador não-linear H na equação (5.6), que em conjunto com as funções de base de Laguerre resulta em uma modelagem *fuzzy*-FBO. Para realizar tal modelagem, é necessário se estabelecer um procedimento para o cálculo do pólo de Laguerre. Esta seção introduz uma proposta de cálculo analítico para este pólo que utiliza a função de autocorrelação do MMW apresentada na seção 3.2.2 do Capítulo 3.

A ordem n do modelo assim como o pólo p das funções ortonormais são parâmetros de projeto. O pólo p, entretanto, pode ser selecionado de forma a minimizar o erro associado a um número estipulado n de funções. Soluções analíticas ótimas têm sido apresentadas na literatura para sistemas lineares e para modelos de Volterra (Fu & Dumont, 1993) dentro de um contexto de funções ortonormais. A maneira mais simples porém efetiva de selecionar este pólo é através de conhecimento a priori da dinâmica dominante do sistema (Ninness & Gustafsson, 1995). Considerou-se neste estudo a função de autocorrelação do modelo multifractal MMW para se obter o pólo dominante do sistema utilizando resultados propostos por Levinson e Durbin em seu algoritmo de determinação de pólos (Hayes, 1996). As entradas para o algoritmo de Levinson-Durbin são os valores atualizados de autocorrelação e como saída, um modelo AR de um pólo pode ser obtido, o qual é incorporado ao modelo *fuzzy*-FBO adaptativo. Seja então a seguinte proposição que estabelece uma relação entre o parâmetro α do MMW e o valor do pólo p:

Proposição 5.4.1 *O pólo p do modelo* fuzzy *para o cálculo das funções de base considerando a função de autocorrelação do MMW (equação (3.43)) pode ser dado por*

$$p = -\frac{1}{2^{\log_2(\frac{\alpha+1}{\alpha+1/2})}}. \tag{5.9}$$

Demonstração O algoritmo de Levinson-Durbin, descrito no Apêndice E, é usado para resolver as equações de modelagem por pólos e as equações normais do método da autocorrelação (Hayes,

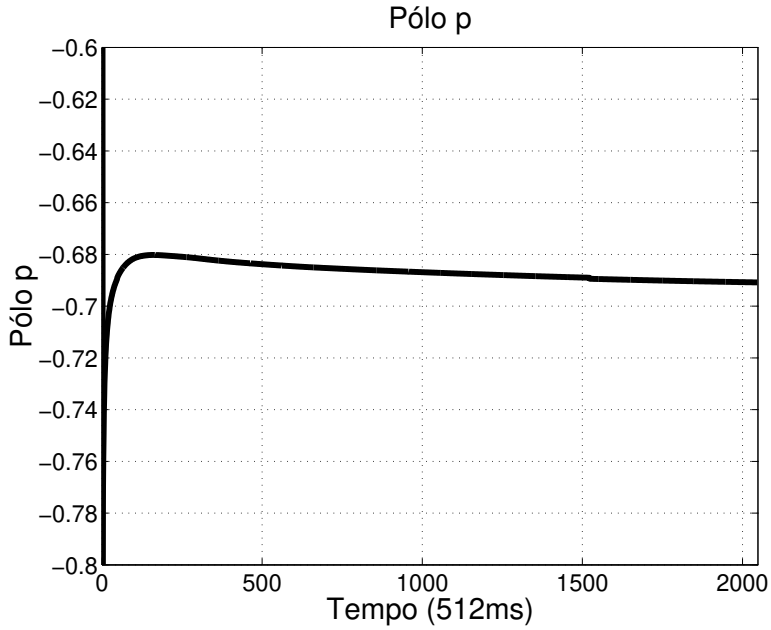


Fig. 5.5: Pólo: dec-pkt-1

1996). Seja $X(t)$ um processo qualquer, sua função de autocorrelação é representada por $r_x(k) = E[X(t+k)X(t)]$. No primeiro passo do algoritmo, calcula-se o valor do coeficiente de reflexão Γ_{j+1} $j = 1, \ldots, n$ para um modelo AR de ordem n. Este coeficiente é usado no cálculo dos pólos do modelo AR. O segundo passo do algoritmo consiste em calcular os valores dos coeficientes $a_{n+1}(j)$ do modelo AR a partir de $a_n(j)$ de tal forma que

$$r_x(k) + \sum_{j=1}^{n} a_n(j) r_x(k-j) = 0; \qquad k = 1, 2, \ldots, n. \tag{5.10}$$

No passo final, atualiza-se o valor do erro ϵ_{j+1}. Como o objetivo é o de encontrar um modelo com 1 pólo $(j = 1)$, pode-se demonstrar que o pólo é dado, considerando um passo do algoritmo, por:

$$\Gamma_{j+1} = -\frac{r_j}{\epsilon_j} \tag{5.11}$$

Através da função de autocorrelação do MMW derivada no Capítulo 3 (3.43), da equação (5.11), e sabendo que $p = \Gamma_1$, tem-se a seguinte equação para o pólo p:

$$p = -\frac{1}{2^{\log_2(\frac{\alpha+1}{\alpha+1/2})}} \tag{5.12}$$

■

Uma vez que este pólo é dado em função do parâmetro multifractal α, o mesmo pode ser estimado em tempo real usando o algoritmo de estimação adaptativa de parâmetros multifractais proposto na seção 3.2.3 do Capítulo 3. A Figura 5.5 mostra o valor do pólo dominante p em função do tempo para o traço de tráfego dec-pkt-1 na escala de tempo de 512ms.

5.5 Modelo fuzzy-FBO Adaptativo

A estimação do valor do pólo permite que a interpolação *fuzzy* de modelos locais, que é a idéia por trás dos modelos do tipo TSK, seja estendida a um contexto de funções de base ortonormais. Esta seção propõe um algoritmo de agrupamento regressivo *fuzzy* adaptativo (ARFA) para o modelo preditor *fuzzy*-FBO que simultaneamente identifica os subespaços *fuzzy* e as funções da parte conseqüente, levando em consideração portanto, a relação entre variáveis de entrada e de saída. O ARFA possui dois estágios: o primeiro consiste de um estágio não-supervisionado para localização dos centros dos agrupamentos, e no segundo aplica-se um algoritmo de gradiente descendente de forma a ajustar os parâmetros envolvidos no primeiro estágio. Nossa proposta se baseia na versão em espaço de estados dos modelos TSK, ou seja, cada regra do modelo *fuzzy* representa um modelo de espaço de estados diferente:

$$R^i: \quad Se \quad l_1(k) \quad \acute{e} \quad L_1^i, \quad \ldots \quad e \quad l_n(k) \quad \acute{e} \quad L_n^i$$

$$\text{Então} \begin{cases} \mathbf{l}_i(k+1) = A_i \mathbf{l}_i(k) + b_i \vec{x}(k) \\ y_i(k) = H_i(\mathbf{l}_i(k)) \end{cases} \tag{5.13}$$

onde R^i corresponde à i-ésima regra, a matriz A_i e o vetor b_i dependem do pólo $p_i(k)$ e $H_i(\mathbf{l}_i(k))$ é o mapeamento que relaciona a saída $y_i(k)$ do modelo local i a seu correspondente estado de Laguerre (funções de base ortonormais) $\mathbf{l}_i(k) = [l_1(k)\, l_2(k) \ldots\, l_n(k)]$, sendo $\vec{x}(k) = [x_1(k)\, x_2(k)\, \ldots\, x_n(k)]$ o vetor de entrada e L_j^i a função de pertinência *fuzzy* para a regra i associada a j-ésima variável de premissa. As variáveis de premissa são os estados de Laguerre do modelo *fuzzy* TSK-FBO resultante. A saída do modelo fuzzy-FBO é dada por

$$y(k) = \frac{\sum_{i=1}^{C} y_i(k) w_i(\mathbf{l}_i(k))}{\sum_{i=1}^{C} w_i(\mathbf{l}_i(k))}, \tag{5.14}$$

onde C é o número de regras (modelos locais) e os pesos $w_i(\mathbf{l}(k))$ da regra i são dados por

$$w_i(\mathbf{l}(k)) = \prod_{j=1}^{n} L_j^i(l_j(k)). \tag{5.15}$$

Um caso particular que simplifica a estrutura do modelo é quando se tem pólos iguais para os C modelos locais ($p_1(k) = \ldots = p_C(k)$). Com esta condição temos $A = A_i$ e $b = b_i$ para $i = 1, \ldots, C$, ou seja, o modelo TSK-FBO pode ser representado conforme as equações (5.5) e (5.6) e H dado pelas equações (5.15) e (5.14). Esta hipótese é equivalente a se dizer que os modos dominantes do sistema não mudam significativamente em diferentes regiões de operação. No caso dessa hipótese não ser verdadeira, a única conseqüência é que um maior número de funções pode ser requerido para prover uma modelagem com uma determinada precisão desejada. Entretanto, este caso particular simplifica consideravelmente o modelo e um valor adequado para o pólo de Laguerre *p* pode amenizar o erro de modelagem. Pode-se verificar que o modelo TSK aproxima com precisão sistemas discretos, causais, invariantes no tempo e estáveis com entrada e saídas limitadas e que não possuam descontinuidades (Takagi & Sugeno, 1985). Como será mostrado, comparado ao modelo TSK, o modelo TSK-FBO proposto adicionalmente representa de maneira eficiente processos variantes no tempo (tráfego de rede) com o uso do algoritmo de treinamento ARFA proposto e pela estimação adaptativa de parâmetros.

O algoritmo ARFA para treinamento do modelo *fuzzy*-FBO leva em conta a distribuição dos dados considerando o erro de regressão e a distância entre os dados de entrada e os *clusters*. Seja a função custo do algoritmo ARFA definida como:

$$J = \sum_{i=1}^{C} \sum_{k=1}^{N} u_{ik}^2 (r_{ik} d_{ik})^2 \tag{5.16}$$

sujeito a

$$\sum_{i=1}^{C} u_{ik} = 1, \quad para \quad 1 \leq k \leq N \tag{5.17}$$

onde u_{ik} é o grau de ativação da i-ésima regra para o k-ésimo padrão de treinamento, C é o número de regras *fuzzy* e N é o número total de dados de treinamento. Na equação (5.16), r_{ik} é o erro entre a k-ésima saída desejada $y(k)$ do sistema modelado e a saída da i-ésima regra $f_i(\vec{x}(k); \vec{a}^i(k))$ com a k-ésima entrada, isto é,

$$r_{ik} = y(k) - f_i(\vec{x}(k); \vec{a}^i(k)), \tag{5.18}$$

com $i = 1, 2, ..., C$ e $k = 1, 2, ..., N$. Na mesma equação (5.16), d_{ik} é a distância entre o vetor de entrada $\vec{x}(k)$ no instante discreto k e o centro do i-ésimo cluster β_i, ou seja,

$$d_{ik} = \vec{x}(k) - \beta_i. \tag{5.19}$$

Para minimizar a função custo J em (5.16), o método dos multiplicadores de Lagrange pode ser

aplicado. A função de Lagrange é definida como (Bazaraa et al., 1993):

$$L = \sum_{i=1}^{C}\sum_{k=1}^{N} u_{ik}^2 (r_{ik} d_{ik})^2 - \sum_{k=1}^{N} \lambda_k (\sum_{i=1}^{C} u_{ik} - 1) \tag{5.20}$$

As condições para minimização desta função fornecem as bases para o desenvolvimento do nosso algoritmo de treinamento. As condições necessárias para minimizar J são:

$$\frac{\partial L}{\partial \vec{a}^i(k)} = \sum_{k=1}^{N} (u_{ik})^2 (d_{ik})^2 \frac{\partial r_{ik}^2}{\partial \vec{a}^i(k)} = 0 \tag{5.21}$$

$$\frac{\partial L}{\partial u_{ik}} = 2u_{ik}(r_{ik} d_{ik})^2 - \lambda_k = 0 \tag{5.22}$$

$$\frac{\partial L}{\partial \beta_i} = \sum_{k=1}^{N} (u_{ik})^2 (r_{ik})^2 \frac{\partial d_{ik}^2}{\partial \beta_i} = 0 \tag{5.23}$$

A derivada parcial $\frac{\partial r_{ik}^2}{\partial \vec{a}^i(k)}$ da equação (5.21) pode ser obtida através de (5.18) da seguinte forma:

$$\frac{\partial r_{ik}^2}{\partial \vec{a}^i(k)} = 2r_{ik}\frac{\partial r_{ik}}{\partial \vec{a}^i(k)} = 2[y(k) - f_i(\vec{x}(k); \vec{a}^i(k))]\frac{\partial r_{ik}}{\partial \vec{a}^i(k)} \tag{5.24}$$

Substituindo (5.24) em (5.21), tem-se:

$$\sum_{k=1}^{N} (u_{ik})^2 (d_{ik})^2 \frac{\partial r_{ik}}{\partial \vec{a}^i(k)} y(k) - \sum_{k=1}^{N} (u_{ik})^2 (d_{ik})^2 \frac{\partial r_{ik}}{\partial \vec{a}^i(k)} f_i(\vec{x}(k); \vec{a}^i(k)) = 0 \tag{5.25}$$

$$\frac{\partial r_{ik}}{\partial \vec{a}^i(k)} = \vec{x}(k) \tag{5.26}$$

Seja $X \in \mathbb{R}^{N\times(n+1)}$ uma matriz onde seus elementos são os valores de $\vec{x}(k)$ em sua j-ésima coluna $j = 1, \ldots, n+1$ (a primeira coluna de X é toda composta por 1), $Y \in \mathbb{R}^N$ um vetor onde o k-ésimo elemento é o valor de $y(k)$ e $Q_i \in \mathbb{R}^{N\times N}$ uma matriz diagonal onde a k-ésima diagonal é dada pelo termo $q(k) = u_{ik}^2 d_{ik}^2$. Assim, pode-se reescrever a equação (5.25) na seguinte forma matricial:

$$X^T Q_i(k) Y - (X^T Q_i(k) X)\vec{a}^i = 0 \tag{5.27}$$

Resolvendo a equação (5.27) e usando a seguinte notação para a matriz de covariância $P_i(k)$:

$$P_i(k) = (X^T Q_i(k) X)^{-1}, \quad i = 1, 2, ..., C, \tag{5.28}$$

resulta que o vetor de parâmetros $\vec{a}^i(k)$ das partes conseqüentes da i-ésima regra para o instante k é obtido pela seguinte equação:

$$\vec{a}^i(k) = P_i(k)X^T(k)Q_i(k)Y(k) \tag{5.29}$$

Tem-se então um problema de inversão matricial na equação (5.28) para o cálculo do vetor de parâmetros das partes conseqüentes em um instante de tempo k. Como proposta, tomou-se como solução para este problema a aplicação de um algoritmo recursivo de estimação dos parâmetros com alta taxa de convergência inicial definido pelas seguintes equações (Young, 1984):

$$\vec{a}^i(k+1) = \vec{a}^i(k) + (P_i(k+1)x(k+1)q(k+1)) \times (y(k+1) - x(k+1)^T\vec{a}^i(k)) \tag{5.30}$$

e,

$$P(k+1) = P(k) - \frac{q(k+1)P_i(k)x(k+1)x(k+1)^T P_i(k)}{1 + q(k+1)x(k+1)^T P_i(k)x(k+1)}, \tag{5.31}$$

onde $x(k+1)$ é a $(k+1)$-ésima linha da matriz X e $q(k+1)$ é o $(k+1)$-ésimo elemento da matriz diagonal $Q_i(k+1)$.

A partir da segunda condição (equação 5.22) de minimização da função J, pode-se escrever a seguinte relação para o grau de ativação u_{ik} da i-ésima regra:

$$u_{ik} = \frac{\lambda_k}{2r_{ik}^2 d_{ik}^2} \tag{5.32}$$

A condição expressa pela equação (5.17) impõe que o parâmetro λ_k de (5.32) seja

$$\lambda_k = \frac{1}{\sum_{i=1}^{C} \frac{1}{2r_{ik}^2 d_{ik}^2}}. \tag{5.33}$$

Ao se substituir (5.33) em (5.32), obtém-se uma equação para o grau de ativação u_{ik} da i-ésima regra que não depende dos multiplicadores de Lagrange λ_k:

$$u_{ik} = \frac{1/2(r_{ik}^2 d_{ik}^2)}{\sum_{i=1}^{C} 1/2(r_{ik}^2 d_{ik}^2)} \tag{5.34}$$

Resolvendo a equação referente à última condição de minimização (equação 5.23), o centro do i-ésimo cluster (β_i) pode ser calculado pela seguinte equação:

$$\beta_i = \frac{\sum_{z=1}^{N} r_{iz}^2 d_{iz}^2 \vec{x}(z)}{\sum_{z=1}^{N} r_{iz}^2 d_{iz}^2} \tag{5.35}$$

Para completarmos a primeira fase do algoritmo de treinamento ARFA aplicado à modelagem *fuzzy*-FBO, devemos obter os parâmetros θ_{j1}^i e θ_{j2}^i das funções de pertinência $A_j^i(\theta_{j1}^i;\theta_{j2}^i)$ das partes premissas, isto é

$$A_j^i(\theta_{j1}^i;\theta_{j2}^i) = \exp\left\{ -\frac{(x_j(k)-\theta_{j1}^i)^2}{2(\theta_{j2}^i)^2} \right\}. \tag{5.36}$$

Tais parâmetros caracterizam a j-ésima função de pertinência da i-ésima regra *fuzzy*, onde $1 \leq j \leq n$ e $1 \leq i \leq C$, e podem ser obtidos a partir de u_{ik} e dos elementos $x_j(k)$ do vetor de entrada através das seguintes equações (Chen et al., 2003):

$$\theta_{j1}^i = \frac{\sum_{z=1}^N (u_{iz})^2 x_j(z)}{\sum_{z=1}^N (u_{iz})^2} \tag{5.37}$$

$$\theta_{j2}^i = \sqrt{\frac{\sum_{z=1}^N (u_{iz})^2 (x_j(z)-\theta_{j1}^i)^2}{\sum_{z=1}^N (u_{iz})^2}} \tag{5.38}$$

Na segunda parte do treinamento, os parâmetros das partes conseqüentes e os subespaços *fuzzy* das partes premissas são ajustados por um algoritmo de aprendizagem supervisionado para melhorar a precisão da modelagem. Aplicando-se o algoritmo de gradiente descendente (Chen et al., 2003) para o modelo TSK-FBO cuja saída é obtida pela equação (5.14), para o k-ésimo padrão de treinamento, a equação de atualização para os parâmetros das partes premissas é (Chuang et al., 2003):

$$\Delta\theta_{jl}^i(k) = \eta(y(k)-\hat{y}(k))(y^i(k)-\hat{y}(k))\frac{1}{\sum_{i=1}^C w^i(k)}\frac{\partial w^i(k)}{\partial \theta_{jl}^i(k)}, \qquad l = 1,2 \tag{5.39}$$

onde η é uma constante de aprendizagem, $y(k)$ é a saída desejada, $\hat{y}(k)$ é a saída do modelo TSK-FBO, $y^i(k)$ é a saída da i-ésima regra do modelo.

Para se calcular o termo $\frac{\partial w^i(k)}{\partial \theta_{jl}^i(k)}$, assume-se $w^i(k) = \min\limits_{j=1,2,\ldots,n} A_j^i(x_j(k))$. Sendo j^* igual ao índice j quando a minimização em $w^i(k)$ ocorre; isto é

$$j^* = \arg\min_j \min_{j=1,2,\ldots,n} A_j^i(x_j(k)). \tag{5.40}$$

Então, quando $j = j^*$, tem-se:

$$\frac{\partial w^i(k)}{\partial \theta_{jl}^i(k)} = \frac{1}{\theta_{j2}^i(k)}\frac{x_j(k)-\theta_{j1}^i(k)}{\theta_{j2}^i(k)}\exp\left\{\frac{(x_j(k)-\theta_{j1}^i(k))^2}{2(\theta_{j2}^i(k))^2}\right\} \tag{5.41}$$

E para $j \neq j^*$, verifica-se que (Chuang et al., 2003):

$$\frac{\partial w^i(k)}{\partial \theta^i_{j2}(k)} = \frac{\partial w^i(k)}{\partial \theta^i_{j1}(k)} = 0 \tag{5.42}$$

De forma análoga, os parâmetros do vetor das partes conseqüentes $\vec{a}^i(k)$ são atualizados com

$$\Delta a^i_j(k) = \zeta(y(k) - \hat{y}(k)) \frac{w^i(k) x_j(k)}{\sum_{i=1}^{C} w^i(k)}, \tag{5.43}$$

onde ζ é uma constante de aprendizagem que controla a taxa de atualização dos valores dos parâmetros das partes conseqüentes.

As funções de pertinência gaussianas obtidas são mostradas na Figura 5.6 após a segunda parte do treinamento. Também após o refinamento, a Figura 5.7 apresenta os agrupamentos formados para uma série de tráfego real utilizando o modelo *fuzzy*-FBO com 2 regras e 1 entrada, e tendo como saída desejada o valor da série a um instante de tempo a frente. O algoritmo ARFA se encarrega de encontrar a melhor posição para o centro dos *clusters*, assim como, o melhor formato para as gaussianas que compõem cada regra *fuzzy* do modelo.

A seguir é apresentado o algoritmo de treinamento proposto para o modelo fuzzy-FBO em questão.

Algoritmo 5.5.1 *Algoritmo de Treinamento ARFA para o Modelo Fuzzy-FBO*

1) Inicialização: No instante de tempo $k = 1$, inicializa-se o algoritmo com valores aleatórios para as variáveis e vetores β_i, ω_i, u_{ik}, $\vec{a}^i(k)$ e d_{ik} para cada regra $i = 1, \ldots, C$. Estabelece inicialmente $P(k) = I_{C,C}$ onde $I_{C,C}$ é a matriz identidade de tamanho CxC.

2) Estime o parâmetro multifractal $\alpha(k)$ no instante de tempo k usando o Algoritmo 3.2.6 do Capítulo 3;

3) Usando o parâmetro $\alpha(k)$ obtido no passo anterior, calcule o pólo p pela equação (5.9):

$$p = -\frac{1}{2^{\log_2(\frac{\alpha(k)+1}{\alpha(k)+1/2})}};$$

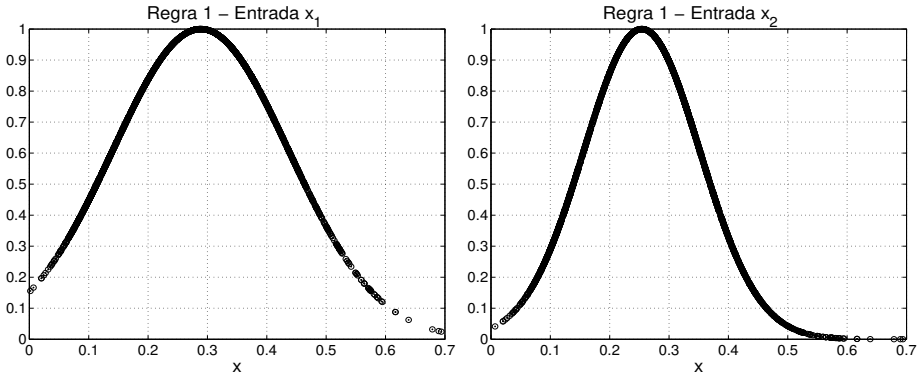
4) Calcule a matriz A e o vetor b usando as equações (5.7) e (5.8), respectivamente;

5) Dado um vetor de entrada $\vec{x}(k)$ (tráfego), calcule seu vetor ortogonal correspondente dado pela equação (5.13):

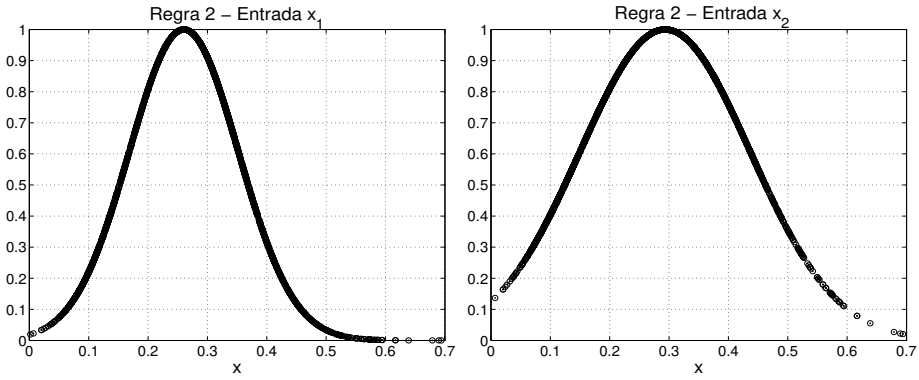
$$\mathbf{l}_i(k+1) = A\mathbf{l}_i(k) + b\vec{x}(k)$$

6) Compute a saída $\hat{y}(k)$ do algoritmo que é uma estimativa do valor desejado $y(k)$ via equação (5.14):

$$\hat{y}(k) = \frac{\sum_{i=1}^{C} y_i(k) w_i(\mathbf{l}_i(k))}{\sum_{i=1}^{C} w_i(\mathbf{l}_i(k))}$$



(a) (b)



(c) (d)

Fig. 5.6: Funções de pertinência obtidas: série de tráfego dec-pkt-2

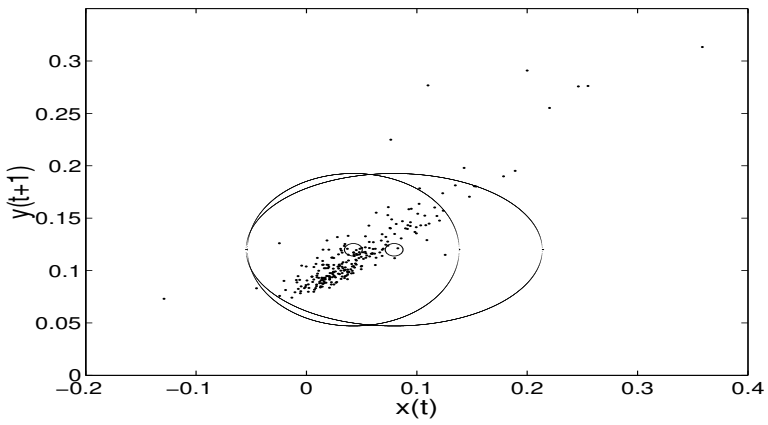


Fig. 5.7: *Clusters* formados (2 regras e 1 entrada): traço de tráfego bc-octext

onde $y_i(k) = \mathbf{l}_i(k)\vec{a}^i(k)$*;*

7) Calcule r_{ik}*, o erro entre a* k*-ésima saída desejada* $y(k)$ *do sistema modelado e a saída da* i*-ésima regra* $f_i(\vec{x}(k); \vec{a}^i(k))$ *com a* k*-ésima entrada através da equação (5.18):*

$$r_{ik} = y(k) - \mathbf{l}_i(k)\vec{a}^i(k),$$

onde $i = 1, 2, ..., C$ *e* $k = 1, 2, ..., N$*;*

8) Calcule a distância d_{ik} *entre a* k*-ésima entrada* $\mathbf{l}_i(k)$ *e o centro do* i*-ésimo cluster* β_i *por meio da equação (5.19), isto é:*

$$d_{ik} = \mathbf{l}_i(k) - \beta_i$$

9) Calcule o grau de ativação u_{ik} *para cada regra* i *(equação (5.34)):*

$$u_{ik} = \frac{1/2(r_{ik}^2 d_{ik}^2)}{\sum_{i=1}^{C} 1/2(r_{ik}^2 d_{ik}^2)}$$

10) O centro do i*-ésimo cluster (*$\vec{\beta}_i$*) pode ser calculado pela equação (5.35):*

$$\vec{\beta}_i(k) = \frac{\sum_{z=1}^{k} r_{iz}^2 d_{iz}^2 \mathbf{l}_i(z)}{\sum_{z=1}^{k} r_{iz}^2 d_{iz}^2}$$

11) Atualiza-se o vetor $\vec{a}^i(k+1)$ *usando as equações (5.30) e (5.31)*

$$\vec{a}^i(k+1) = \vec{a}^i(k) + (P_i(k+1)l_n(k+1)q(k+1)) \times (y(k+1) - l_n(k+1)^T \vec{a}^i(k))$$

e

$$P(k+1) = P(k) - \frac{q(k+1)P_i(k)l_n(k+1)l_n(k+1)^T P_i(k)}{1 + q(k+1)l_n(k+1)^T P_i(k)l_n(k+1)}$$

12) Calcule os parâmetros θ_{j1}^i *e* θ_{j2}^i *das funções de pertinência* $A_j^i(\theta_{j1}^i, \theta_{j2}^i)$ *das partes premissas pelas equações (5.37) e (5.38), respectivamente:*

$$\theta_{j1}^i(k+1) = \frac{\sum_{z=1}^{k}(u_{iz})^2 l_n(z)}{\sum_{z=1}^{k}(u_{iz})^2}$$

e

$$\theta_{j2}^i(k+1) = \sqrt{\frac{\sum_{z=1}^{k}(u_{iz})^2 (l_n(z) - \theta_{j1}^i)^2}{\sum_{z=1}^{k}(u_{iz})^2}}$$

13) Utilizando os parâmetros θ_{j1}^i *e* θ_{j2}^i *das funções de pertinência das partes premissas, calcule a*

saída do modelo pela equação (5.14):

$$\hat{y}(k+1) = \frac{\sum_{i=1}^{C} y_i(k) w_i(\mathbf{l}_i(k))}{\sum_{i=1}^{C} w_i(\mathbf{l}_i(k))}$$

onde $w^i(k) = \min_{j=1,2,\ldots,n} A_j^i(\theta_{j1}^i, \theta_{j2}^i)$*;*

14) Aplica-se o refinamento do algoritmo representado pelas equações (5.39) e (5.43);

15) Faça $k = k + 1$ *e volte ao passo 2. O algoritmo se encerra ao atingir o instante de tempo* k *desejado.*

5.5.1 Avaliação de Desempenho de Predição do Modelo Fuzzy-FBO

Foram utilizadas nas simulações desta seção, traços de tráfego TCP/IP obtidos da Digital Equipment Corporation[1], traços de tráfego Ethernet obtidos da Bellcore [2] e traços capturados entre os anos de 2000 e 2003 na rede da Petrobrás (LRPRC, 2002), já relatados na seção 3.3.1 do Capítulo 3.

Na presente seção são realizadas avaliações comparativas entre o desempenho do preditor proposto e o desempenho de outros três diferentes preditores, quando aplicados a traços de tráfego TCP/IP e Ethernet. Os outros preditores levados em consideração foram: o LMS (*Least Mean Square*) (Haykin, 1989), o RLS (*Recursive Least Square*) (Haykin, 1989) e o Fuzzy FCRM (Kim et al., 1997)(Vieira & Lee, 2004a) apresentados no Capítulo 4. Dentre estes preditores, o LMS e o RLS são preditores lineares adaptativos de destaque na literatura. Estes preditores são considerados neste trabalho devido a simplicidade do LMS e o bom desempenho de predição adaptativa de sinais do algoritmo RLS. Já o preditor *fuzzy* FCRM não é adaptativo, mas foi escolhido justamente para se comparar o treinamento adaptativo e em batelada ('on batch') entre os modelos *fuzzy* considerados. Também se considerou a avaliação do preditor *fuzzy* adaptativo proposto sem a inclusão das funções de base ortonormais, o que foi denominado de 'preditor *fuzzy* adaptativo' nas tabelas e gráficos apresentados a seguir. O preditor proposto foi avaliado utilizando-se duas medidas relativas de erro. Conhecidos como erros quadráticos médios normalizados (EQMN), a primeira medida consiste em normalizar o EQM em relação à variância da série predita, denominada de EQMN1, cuja definição foi dada na seção 4.6 do Capítulo 4, enquanto a segunda medida consiste em normalizar o EQM em relação ao erro quadrático médio do preditor ótimo para o processo passeio aleatório (*random walk*). A definição do segundo tipo de EQMN é dada a seguir.

Definição 5.5.1 *- Seja* $\hat{x}_{pa}$ *o valor predito da amostra* x *do processo* X *cuja média é* μ*, estimado como sendo igual ao valor da amostra imediatamente anterior. Define-se o erro quadrático médio*

[1] http://ita.ee.lbl.gov/html/contrib/DEC-PKT.html
[2] http://ita.ee.lbl.gov/html/contrib/BC.html

normalizado do tipo 2 como

$$EQMN2 = \frac{E\left[(x-\mu)^2\right]}{E\left[(\hat{x}_{pa}-x)^2\right]}. \tag{5.44}$$

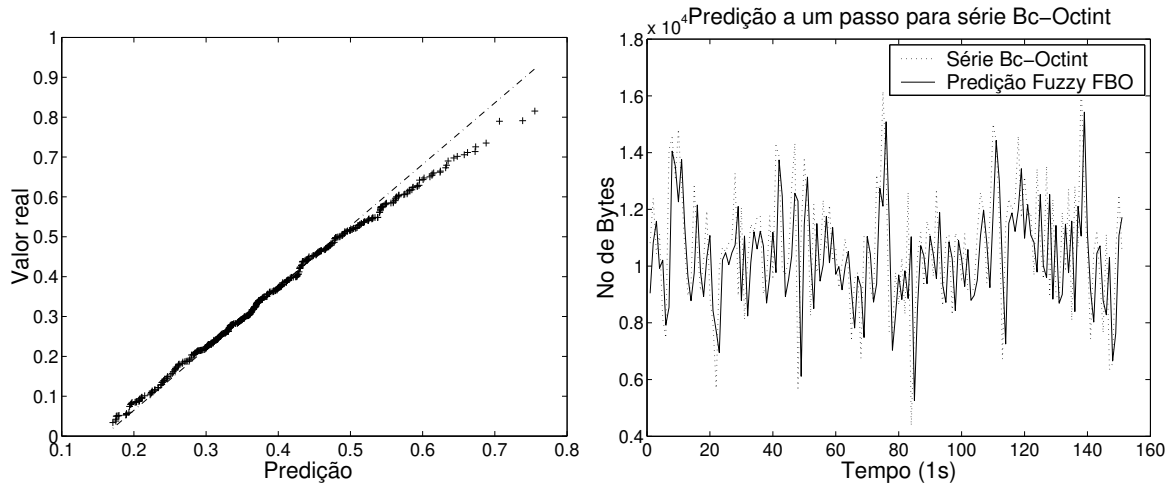
De acordo com as definições dos EQMN1 e EQMN2, um preditor que apresente o valor de EQMN1 igual ou inferior a 1 possuirá desempenho igual ou superior a um preditor que apenas estime o valor futuro como sendo igual à média do processo. Para um EQMN2 próximo a 1, o preditor analisado apresentará desempenho próximo ao de um preditor que estime o valor futuro como sendo igual ao valor imediatamente anterior.

Antes de se iniciar as comparações com outros preditores, são apresentados alguns resultados de predição do modelo preditor *fuzzy* proposto para a série da Bellcore Bc-octint. A Fig. 5.8(a) compara os valores preditos e os valores reais através de um gráfico quantil-quantil conhecido como QQplot (Rolls et al., 2005). Pode-se notar uma relação linear entre os valores preditos e os valores reais, o que denota desempenho de predição adequado. Isso pode ser constatado pela proximidade entre os valores preditos a um passo e valores reais mostrados pelo gráfico da Figura 5.8(b) e em termos mais gerais (média, intervalo de confiança para a média, desvio padrão, etc) pela Figura 5.8(c).

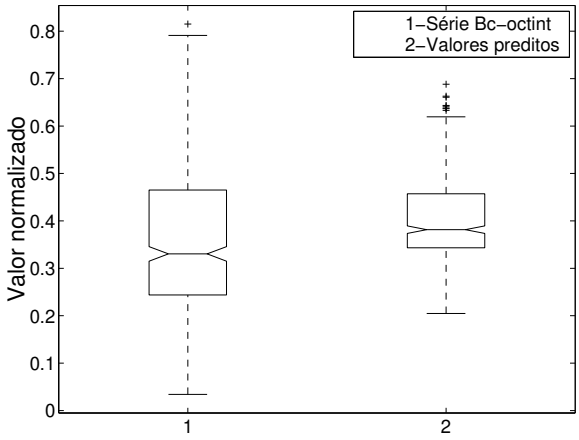
A segunda fase do treinamento ARFA proporciona uma redução do EQMN1 e 2 como pode ser visto pela Figura 5.9(a), uma vez que na primeira fase do treinamento ARFA se obtém um modelo *fuzzy* aproximado. Para realizar a comparação entre os preditores analisados, encontrou-se a configuração para cada um que minimizava os EQMN1 e 2, escolhendo assim adequadamente por exemplo, a taxa de aprendizagem, os valores iniciais para os centros, etc. Os resultados apresentados na Tabela 5.1 enfatizam que o preditor adaptativo obtido supera em termos de erro de predição os algoritmos adaptativos lineares LMS e RLS. O algoritmo RLS já é conhecido por sua rápida convergência, sendo esta característica também constatada nas simulações para o preditor *fuzzy*-FBO. Isso quer dizer que o a saída do preditor fuzzy proposto se ajusta rapidamente com poucas amostras iniciais de tráfego. Note que a tabela 5.1 apresenta os resultados de desempenho obtidos para estes algoritmos na melhor configuração encontrada, ou seja, para os parâmetros de aprendizagem e número de coeficientes mais adequados. Essa consideração é realizada para os demais testes, a não ser quando se fixa o número de coeficientes do modelo em questão.

Outra capacidade do preditor *fuzzy* FBO proposto que pode ser verificada é a de predição a um passo de valores desconhecidos sem adaptação dos pesos. Pela Figura 5.10, pode-se verificar menores valores de EQMN1 versus tempo para o preditor *fuzzy* FBO com relação ao mesmo preditor que não usa funções de base ortonormais e ao preditor RLS, aplicados à série dec-pkt-2. Isso se deve a estrutura mais complexa do modelo *fuzzy*-FBO do que a dos modelos lineares, até mesmo com relação ao modelo *fuzzy* adaptativo sem as funções de base ortonormais, que perde o ganho de modelagem propiciado pela transformação da entrada aplicada pelas funções de base ortonormais.

Um dos objetivos do treinamento adaptativo é o ajuste do algoritmo ao ambiente dinâmico do



(a) (b)



(c)

Fig. 5.8: Desempenho de predição para a série Bc-octint usando o algoritmo *Fuzzy* FBO proposto.

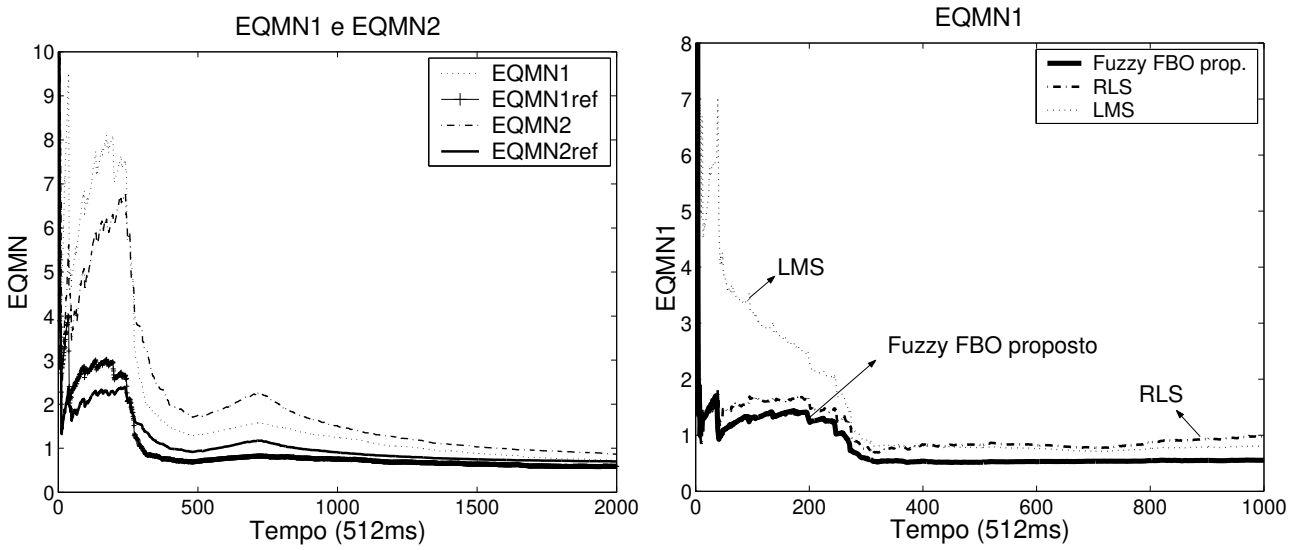
Tab. 5.1: Comparação de EQMN1

Traço de tráfego	Intervalo	Fuzzy-FBO Adapt	Fuzzy Adapt	RLS	LMS	Fuzzy FCRM
Dec-pkt-1	1-2048	0.6564	0.7366	0.8513	0.9304	0.6987
Dec-pkt-2	1-2048	0.5758	0.6704	0.7022	0.7614	0.5836
Bc-Octint	801-1701	0.4102	0.4654	0.4114	0.4817	0.3107
Bc-Octext	1000-2000	0.4144	0.4355	0.4298	0.5010	0.4654

tráfego de redes. Com objetivo de comparação, foram inseridos também na Tabela 5.1 os valores de EQMN1 obtidos com o preditor *fuzzy* FCRM não-adaptativo, cujos resultados se encontram em (Vieira & Lee, 2004a). O preditor *fuzzy* FCRM se utiliza de todas as amostras da série de tráfego para o cálculo de seus parâmetros. Em seguida, este modelo preditor é aplicado na predição a um passo da série em questão. Pode-se observar que entre os preditores adaptativos, o preditor *fuzzy*-FBO obteve menor EQMN1 para as séries analisadas, e em geral menor do que para o preditor *fuzzy* FCRM estático. Portanto, os resultados comprovam que se pode ter com o conhecimento de poucas amostras do passado (neste teste, utilizou-se 2 coeficientes e 2 regras), um erro tão baixo quanto ao processamento com toda série. Devido a este fato, é possível a implementação de um algoritmo mais rápido, que necessite de pouca capacidade de armazenamento e com convergência acelerada, como é o caso do preditor *fuzzy*-FBO proposto. Além disso, este preditor *fuzzy*-FBO adaptativo captura mais adequadamente as características do processo de tráfego a ser modelado por não supor de antemão que a 'estrutura' do processo seja invariante, como é feito por alguns modelos com treinamento 'offline'.

Em teoria à medida que se aumenta o número de regras *fuzzy* (e por conseqüência o número de funções de base) para o preditor *fuzzy*-FBO, se obtém um EQMN de predição menor para determinada série de tráfego. Além disso, é importante se estimar precisamente os valores para os pólos para o modelo *fuzzy*-FBO. No entanto, o que se observou para todos os preditores testados é que os EQMN1 e EQMN2 diminuem até um certo número de regras e coeficientes, depois disso, nem sempre é possível obter um eficiente treinamento para os modelos preditores. As Tabelas 5.2 e 5.3 corroboram esta afirmação, onde para efeito de simplificação, faz-se o número de regras igual ao número de coeficientes (igual ao número de amostras passadas) dos modelos para predição da série dec-pkt-1. Com 5 regras, o preditor *fuzzy* FBO começa a apresentar uma deterioração dos EQMN1 e 2. Para os algoritmos RLS e LMS, o mesmo ocorre com o número de coeficientes igual a 7. Note entretanto que, mesmo esses dois algoritmos estando em sua melhor configuração não propiciaram EQMN menor do que o modelo *fuzzy*-FBO com 2 regras e 2 coeficientes.

Ainda com relação ao preditor *fuzzy* proposto é importante avaliar o desempenho de predição dos



(a) Comparação entre EQMN1 e EQMN2 antes e depois do refinamento (curvas de aprendizagem)

(b) Comparação de EQMN1

Fig. 5.9: Comparação entre erros quadráticos médios normalizados para o traço de tráfego dec-pkt-2.

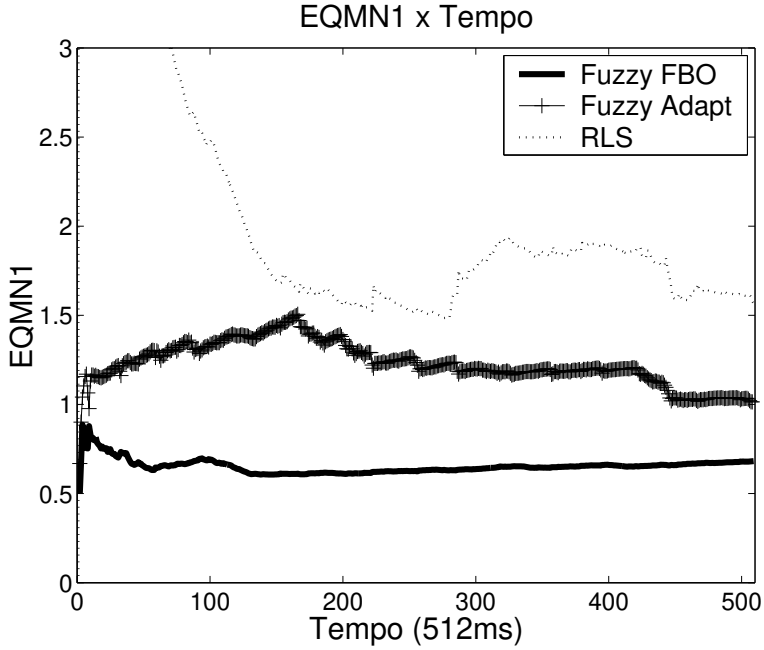


Fig. 5.10: Teste com pesos fixos.

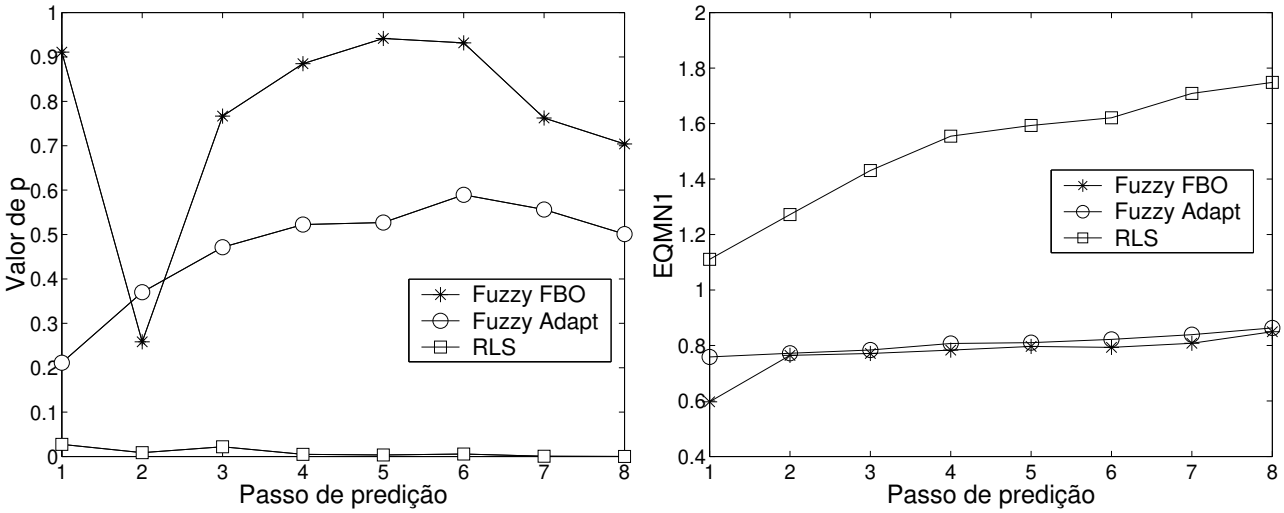
Tab. 5.2: Relação entre EQMN1 e Número de Regras

Número de regras	Fuzzy FBO	Fuzzy Adapt.	RLS	LMS
2	0.6470	0.7973	1.0387	1.1977
3	0.6203	0.7879	0.9587	1.0908
4	0.6050	0.7852	0.9492	1.0718
5	0.6088	0.7945	0.9460	1.0539

Tab. 5.3: Relação entre EQMN2 e Número de Regras

Número de regras	Fuzzy FBO	Fuzzy Adapt.	RLS	LMS
2	0.6121	0.7208	0.9553	1.0779
3	0.5762	0.7114	0.8820	0.9779
4	0.5633	0.7043	0.8739	0.9531
5	0.5846	0.7282	0.8731	0.9378

algoritmos para horizontes maiores de predição. Com esse fim, analisou-se os erros de predição com o teste-T (Qiu, 1999). O teste T é um teste de hipótese que pode ser usado para se determinar se uma afirmação sobre uma característica de uma série é verdadeira. Este teste provê a probabilidade do grau de veracidade desta afirmação através da variável conhecida na literatura como p (Qiu, 1999). O valor de p corresponde a probabilidade de se observar determinado resultado dado que a hipótese nula é verdadeira. Quanto menor o valor de p, menos crédito pode ser dado à hipótese nula. O teste T foi conduzido nas seqüências de erro produzidas pelos algoritmos de predição adaptativos, observando o valor de p a medida que o passo de predição é aumentado (Figura 5.11(a)). A hipótese nula H_0 é de que a média do erro μ seja igual a zero ($\mu = 0$) e a hipótese alternativa H_1: $\mu \neq 0$. Pode-se utilizar diferentes níveis de significância para se testar uma hipótese com o teste T. O nível de significância de um teste estatístico é a probabilidade de rejeição de uma hipótese verdadeira. Fixou-se o nível de significância do teste realizado em 0.05, que corresponde a um intervalo de confiança de 0.95. A Figura 5.11(a) apresenta os valores de p com diferentes passos de predição para a série dec-pkt-2. Em todos os passos analisados, foram obtidos $h = 0$ para o modelo *fuzzy*-FBO, ou seja, não se deve rejeitar a hipótese nula com um nível de significância 0.05. Note que os menores valores de p para o preditor *fuzzy*-FBO indicam maiores graus de segurança em afirmar que as médias dos erros de predição sejam zero ao se usar esse preditor. De fato, como pode ser visto pela Figura 5.11(b), também o EQMN1 do preditor *fuzzy*-FBO se manteve com valores abaixo dos demais preditores.



(a) p: dec-pkt-2

(b) EQMN1: dec-pkt-2

Fig. 5.11: Probabilidade de ocorrência de hipótese nula (p) versus passo de predição

5.6 Estimação Adaptativa de Banda baseada no Preditor Fuzzy Proposto

Os esquemas de alocação de banda podem se beneficiar de algoritmos de predição da taxa de tráfego de modo a antecipar as ações para alocação de recursos e controle de congestionamento (Tran & Ziegler, 2005). Neste sentido, o passo de predição pode ser ajustado de forma a possibilitar que a rede tenha tempo para obtenção e alocação dos recursos necessários, assim como liberação dos mesmos. Precisamente nesta seção é proposto um esquema de alocação dinâmica de banda que considera como entrada as predições realizadas pelo modelo *fuzzy* multifractal proposto. Neste esquema de alocação de banda, considera-se que a banda para um fluxo é alocada apenas em quantidades finitas expressas como frações da máxima banda disponível.

Seja $A(\tau, t)$ um processo a tempo discreto correspondente ao tráfego acumulado (neste caso, número de *bytes*) no intervalo de tempo (τ, t) e que chega ao servidor para ser transmitido. Para satisfazer um limite de retardo d_{req}, qualquer pacote (ou dados) deve ser transmitido até o instante de tempo '$t + d_{req}$'. A banda ε, ou seja, a taxa necessária portanto para atender a esse critério e para se ter perda zero de tráfego, deve obedecer a seguinte relação:

$$\varepsilon * (t - \tau + d_{req}) \geq A(\tau, t) + b_{t-1} \quad \forall \quad t \geq \tau, \tag{5.45}$$

onde b_{t-1} corresponde ao número de *bytes* não escoados pela rede no instante anterior.

Note que a relação acima deve ser satisfeita para todo $t \geq 0$ e $\tau \geq 0$ onde $t \geq \tau$. Um meio de se obter um algoritmo eficiente de estimação de banda é estimar o tráfego $A(\tau, t)$ através de seu valor predito $A_p(\tau, t)$, ou seja, prever a taxa com a qual o fluxo será injetado no enlace de modo que este respeite o limite de retardo estipulado.

O procedimento pelo qual se propõe mapear o valor predito da intensidade de tráfego com a banda requerida é o seguinte: Seja ΔBW a quantia de banda finita e C a máxima banda disponível. Determina-se o intervalo de banda $[(k)\Delta BW, (k+1)\Delta BW] \quad (k \geq 0)$ no qual o valor da taxa predito P_{t+1} se encontra e usa-se o valor superior do intervalo $(k+1)\Delta BW$ ou a máxima banda disponível C no caso de $C < (k+1)\Delta BW$ como banda requerida no instante $t+1$. Ou seja, a banda BW_{t+1} no instante $t+1$ de acordo com a equação (5.45) é dada por

$$BW_{t+1} = min\left\{\left\langle \frac{(P_{t+1} + b_t)/(t + d_{req})}{\Delta B} \right\rangle \Delta B, C\right\}, \tag{5.46}$$

onde o operador $\langle x \rangle$ representa o maior número inteiro mais próximo de x.

Como a banda requerida (5.46) é o valor superior do intervalo onde se encontra o valor da taxa predito, tem-se uma provisão de banda menos sensível às pequenas variações das predições, dentro dos intervalos de tamanho ΔBW. Isso porque todos os valores de taxa predito contidos em um mesmo intervalo $[(k)\Delta BW, (k+1)\Delta BW]$ estão relacionados a um mesmo valor de banda. Efetua-se assim, uma suavização implícita das predições pelo esquema de provisão de banda que evita a freqüente realocação de banda, reduzindo o custo de sinalização envolvido. Portanto, conclui-se que ΔBW está diretamente ligado ao custo de sinalização do esquema de alocação de banda. Neste trabalho, considerou-se o caso $d_{req} = 0$, deixando para trabalhos futuros a análise de garantias de diferentes retardos.

5.6.1 Medidas de Desempenho e Resultados Experimentais

Esta seção apresenta as medidas de desempenho utilizadas para avaliar o esquema de provisão de banda proposto, assim como os resultados de desempenho obtidos para o esquema de alocação de banda apresentado na seção anterior. Seja z_t denotando o valor observado de taxa de tráfego e BW_t a banda fornecida no instante de tempo t. As seguintes medidas de desempenho podem ser definidas:

1. **Utilização média (u)**: A utilização média mede a fração de banda usada para servir o fluxo de dados observados no período de tempo T, calculada por

$$u = \frac{1}{T}\sum_{t=1}^{T} min\left\{\frac{z_t}{BW_t}, 1\right\}. \tag{5.47}$$

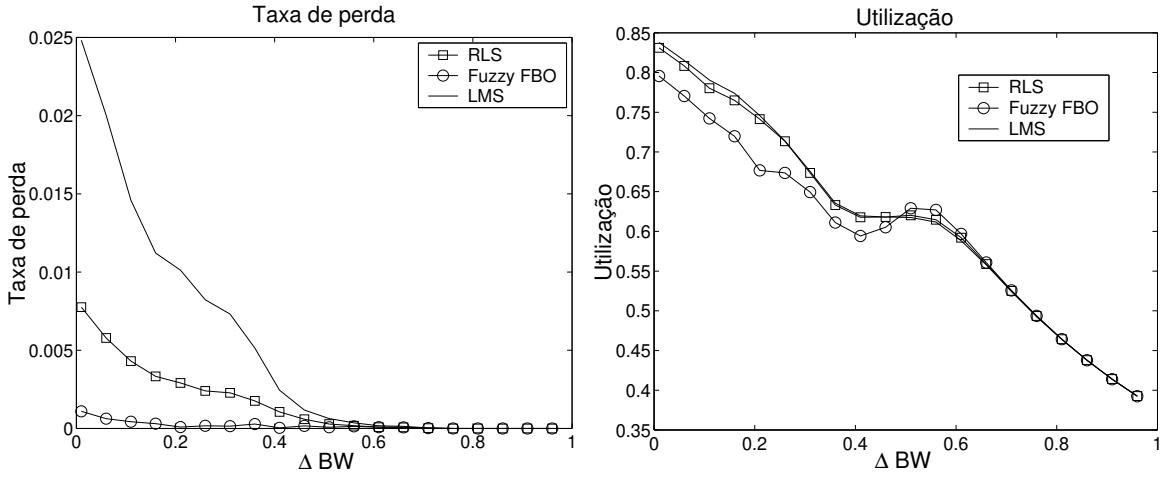
2. **Taxa de perda** (TP): A taxa de perda mede a quantidade de *bytes* perdidos devido à alocação de banda menor do que a necessária, dada por

$$TP = \frac{1}{T}\sum_{t=1}^{T} max\left\{\frac{(z_t - BW_t) - B}{z_t}, 0\right\}. \tag{5.48}$$

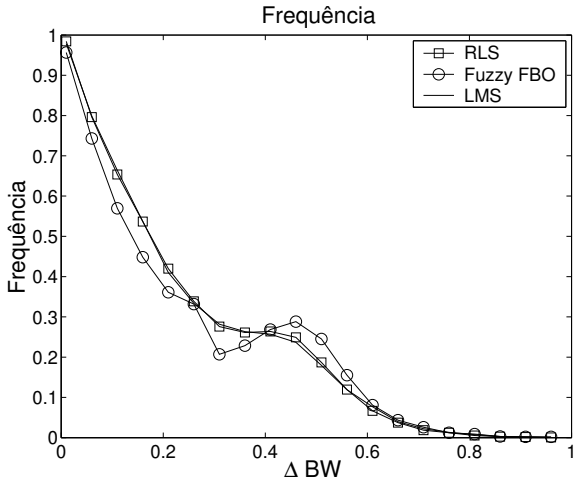
   Inicialmente avaliaremos o esquema de alocação de banda para o caso em que o *buffer* $B = 0$.

3. **Freqüência de sinalização**: A freqüência de sinalização nos ajuda a avaliar quão freqüente o esquema proposto realoca banda. A realocação de banda envolve um custo de sinalização para as redes. Assim, em projeto de esquema eficiente de alocação de banda deve-se, além de garantir simultaneamente alta utilização e baixa taxa de perda, levar em conta sua freqüência de sinalização.

Comparou-se o esquema de provisão de banda proposto com dois outros esquemas, entre eles, o esquema apresentado por Adas et al. (Adas, 1998), que usa o algoritmo LMS para a predição de tráfego de redes e estas predições são usadas como taxas exigidas pelos fluxos. Outro trabalho comparado (Chong et al., 1995), utiliza o algoritmo RLS que possui uma convergência mais rápida, sendo como o anteriormente citado, um esquema adaptativo. A fim de se estudar os efeitos do intervalo de faixas de banda sobre todos os esquemas em questão, a equação (5.46) é aplicada nos 3 esquemas em comparação, onde a diferença entre eles consiste nos valores de P_{t+1} dados pelos algoritmos de predição *fuzzy*-FBO, RLS e LMS. Os resultados de desempenho: taxa de perda, utilização e frequência dos 3 esquemas de alocação de banda aplicado a série de tráfego dec-pkt-2 são mostrados na Figura 5.12. Pode-se observar que o esquema de alocação de banda proposto atinge uma taxa de perda menor a uma freqüência de sinalização semelhante aos demais, porém a custo de uma utilização do enlace ligeiramente menor. Espera-se que isso ocorra devido a maior precisão do algoritmo de predição *fuzzy*-FBO. Observa-se que a aplicação do algoritmo de predição proposto à alocação de banda leva a menores taxas de perda mesmo com a variação dos valores de *buffers* (Figura 5.13(a)) e uma ocupação do *buffer* mais bem comportada indicada pelo seu tamanho médio da fila (Figura 5.13(b)). Devido a esses resultados, pode-se verificar que a probabilidade de perda para determinado tamanho de *buffer* é menor para o esquema de alocação de banda proposto. Ao se plotar a função de distribuição acumulada complementar (Figura 5.13(c)) que corresponde a probabilidade do tamanho da fila Q ser maior do que um certo valor de *buffer* x do processo de tráfego no *buffer*, visualiza-se claramente a afirmação anterior.

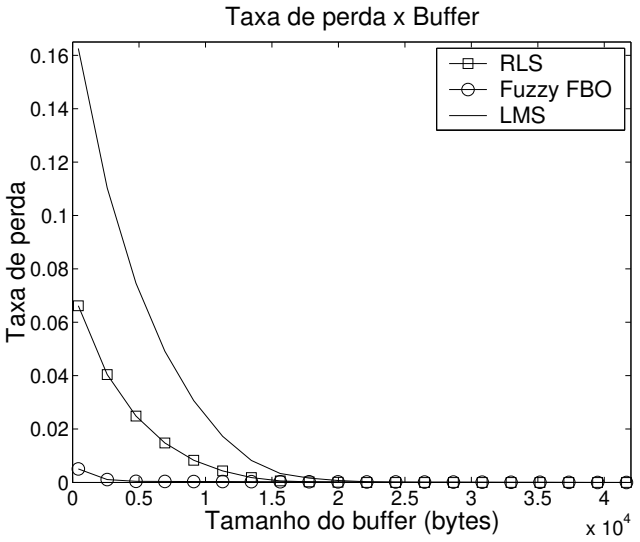


(a) (b)

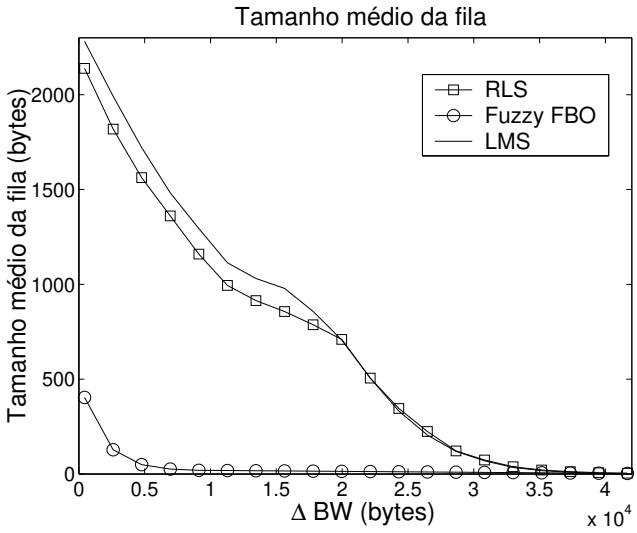


(c)

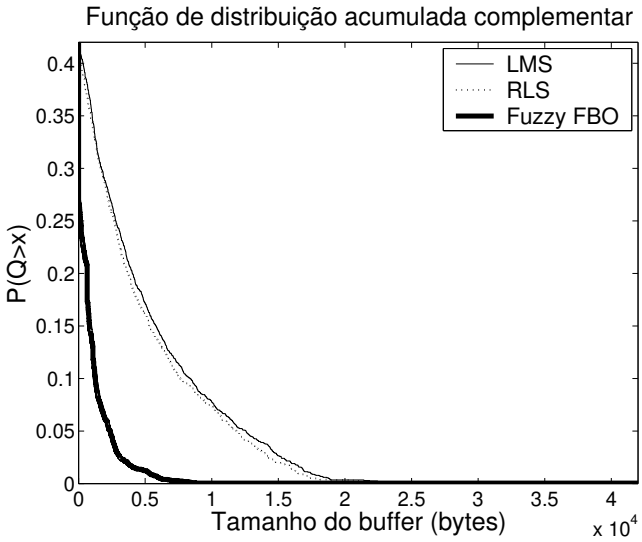
Fig. 5.12: Comparação de desempenho entre esquemas de alocação de banda



(a) Taxa de perda x Tamanho do buffer



(b) Ocupação média da fila x ΔBW



(c) $P(Q > x)$ x Tamanho do *buffer*

Fig. 5.13: Análise do comportamento de fila dos esquemas de alocação de banda para o traço de tráfego 10-7-S-1

5.7 Considerações Finais

As características dos fluxos de tráfego nas rede atuais como dependência de longo prazo e rajadas em múltiplas escalas tornam a modelagem e predição de tráfego tarefas difíceis e desafiadoras. Neste capítulo foi proposto um modelo *fuzzy*-FBO cujo algoritmo de treinamento adaptativo ARFA permite que a predição adaptativa e em tempo real do tráfego de redes seja realizada com um número reduzido de regras nebulosas. O algoritmo de treinamento ARFA desenvolvido consiste de 2 estágios, ambos adaptativos, onde no primeiro cria-se agrupamentos *fuzzy* e no segundo se faz um ajuste fino dos parâmetros obtidos no primeiro estágio, como posicionamento dos centros e forma geométrica das funções de pertinência. Mostrou-se através de simulações que o segundo estágio proporciona uma diminuição considerável do erro de predição.

A fim de se obter as funções de base ortonormais para o modelo *fuzzy* em questão através do cálculo do pólo do modelo, utilizou-se a função de autocorrelação para o MMW, assim como o algoritmo de estimação adaptativa de parâmetros multifractais, ambos apresentados no Capítulo 3. Em seguida, um procedimento para o cálculo do pólo dominante foi apresentado na Proposição 5.2.1, utilizado como pólo de Laguerre para o modelo *fuzzy*-FBO. Comprovou-se que há uma melhoria de desempenho do modelo preditor, ou seja, predições mais precisas e robustas são obtidas com o acréscimo das funções de base ortonormais. As comparações realizadas com outros preditores, demonstraram um desempenho superior de predição do modelo *fuzzy*-FBO proposto para diferentes horizontes de predição e número de regras consideradas. Constatou-se esse fato por meio do erro quadrático médio normalizado e do teste de hipótese T aplicado a série de erros obtidos com os preditores comparados.

Quanto à provisão de banda utilizando algoritmos de predição, apresentou-se um novo esquema de provisão de banda baseado em predição. Este esquema relaciona as predições de tráfego realizadas pelo modelo *fuzzy*-FBO com a banda a ser alocada no próximo instante de tempo de modo a obter informação atualizada da taxa necessária para o tráfego de dados em um enlace. Uma das vantagens do esquema adaptativo de provisão de banda é que a alocação de banda pode ser realizada baseada nas amostras de tráfego disponíveis no instante de tempo atual. Como este esquema de provisão de banda se adapta segundo as mudanças do tráfego, um melhor aproveitamento dos recursos é obtido comparado a uma alocação estática de banda. O método de alocação de banda proposto consegue manter um bom equilíbrio entre perda de dados, freqüência de sinalização e utilização. Verificou-se que o objetivo de conseguir uma menor taxa de perda com a aplicação do nosso esquema preditivo de provisão de banda foi atingido. Aliado a isso, observou-se uma taxa média de ocupação do *buffer* mais baixa e uma probabilidade de perda de *bytes* menor do que os outros esquemas. Tal resultado se deve ao melhor desempenho de predição proporcionado pelo modelo *fuzzy*-FBO. A partir do método de predição proposto, esquemas que considerem também minimização do custo de sinalização podem

ser projetados. Questão essa, que será deixada para trabalhos futuros.

Capítulo 6

Cálculo de Banda Efetiva para Tráfego Multifractal

6.1 Introdução

Em engenharia, o termo banda é muitas vezes relacionado com a largura espectral dos sinais eletromagnéticos. Mais especificamente no contexto de redes, a banda quantifica a taxa na qual o enlace de rede ou caminho de rede pode transferir dados. A banda fornecida às aplicações tem impacto direto no desempenho da rede. Aplicações sensíveis a retardo se beneficiam com menores retardos providos por maiores bandas. Várias aplicações e tecnologias de rede podem se beneficiar com o conhecimento das características dos fluxos de tráfego e da banda disponível nos caminhos de rede. Por exemplo, aplicações *peer-to-peer* formam suas redes de nível de usuário baseados na banda disponível entre 'peers' (Tran & Ziegler, 2005). Redes *overlay* podem configurar suas tabelas de roteamento baseado nas bandas de seus enlaces (Duan et al., 2002). Provedores de rede podem estabelecer preço pela banda usada. Os contratos de níveis de serviço (SLA- *Service Level Agreements*) estabelecidos entre provedores e usuários definem serviços em termos da banda disponível em pontos de interconexão limites (Blake, 1998).

O conceito de banda efetiva provê um modo de caracterizar os requisitos de recursos de uma conexão. A banda efetiva é uma ferramenta útil para análise e descrição do tráfego em redes. Tem como limite inferior a taxa média e limite superior a taxa de pico do fluxo de tráfego. Pode-se dizer que a banda efetiva, dado um tamanho de *buffer*, representa a taxa de serviço que é efetivamente necessária para servir um fluxo de tráfego respeitando uma determinada probabilidade de perda, ou seja, ela corresponde à capacidade que pode ser usada para atender parâmetros de QoS exigidos por um fluxo. Além disso, se vários fluxos de tráfego forem simultaneamente servidos a taxas equivalentes às suas bandas efetivas, então as demandas de QoS não serão violadas (Duffield & O'Connell,

1993a). Com isso, a banda efetiva simplifica consideravelmente os algoritmos de controle de admissão (CAC-*Connection Admission Control*) para tráfego de redes, podendo ser empregada também em outras aplicações de dimensionamento e controle de rede.

Os modelos multifractais são considerados mais adequados para modelar o tráfego de redes por incorporar várias características do tráfego, generalizando assim vários modelos existentes na literatura. É de se esperar que uma descrição mais fiel dos fluxos de tráfego baseada em modelos multifractais contribua com a melhoria nas ferramentas de estimação de propriedades desses fluxos. Essa afirmação será verificada com relação à banda efetiva neste capítulo e para a probabilidade de perda e controle de admissão, nos próximos capítulos.

Este capítulo inicialmente introduz alguns princípios e teoremas da Teoria dos Grandes Desvios, objetivando dar suporte teórico para outras seções e capítulos. Como contribuição original, na seção 6.4, deriva-se uma expressão para a banda efetiva do MMW dada em função de seus multiplicadores. Na subseção 6.4.1, verifica-se o desempenho da equação de banda proposta em atender aos requisitos de perda dos fluxos e comparações com outras abordagens de estimação de banda são realizadas. A capacidade de atualização do MMW em tempo real é usada na elaboração de um algoritmo de estimação adaptativo de banda efetiva, onde o valor da banda é atualizado à medida que dados dos processos de tráfego são disponibilizados. Este algoritmo permite que na subseção 6.5.1 seja desenvolvido um esquema de provisão adaptativo de banda, no qual a alocação de banda ao fluxo pode ocorrer em um intervalo de tempo maior do que o cálculo da banda efetiva para este fluxo, que é realizado em intervalos regulares de tempo menores. Acredita-se que muitas aplicações possam se beneficiar com o algoritmo proposto. Suas possíveis aplicações em contextos reais de rede são comentadas na subseção 6.5.2. Deve-se ressaltar que no Capítulo 5 foi desenvolvido um esquema de alocação de banda cuja meta era de se obter perda zero de dados de tráfego em uma conexão. Neste capítulo, esta condição é relaxada, ou seja, objetiva-se estimar a taxa necessária que dever ser fornecida ao fluxo de forma a atender aos parâmetros de perda requisitados, os quais podem ser diferentes de zero.

6.2 Teoria dos Grandes Desvios e Banda Efetiva

A Teoria dos Grandes Desvios se refere a um conjunto de técnicas para estimar propriedades de eventos raros tais como suas freqüências e qual a forma mais provável deles ocorrerem. Eventos raros são eventos causados por um conjunto de coisas não-prováveis acontecendo ao mesmo tempo; como se algo determinasse e conspirasse para que tal evento ocorresse. Não deve ser confundido com um evento único de pequena probabilidade. Um exemplo de evento raro associado a telecomunicações acontece quando temos chuvas fortes ou então alguma data comemorativa (Ano Novo, Natal, etc.) e várias pessoas resolvem ligar ao mesmo tempo, levando a um congestionamento nas vias de

comunicação.

Eventos que acontecem longe da média e fora da principal distribuição de freqüência são eventos raros (Courcoubetis & Weber, 1994). De modo simplificado pode-se dizer que estes eventos são causados por um grande número de fatores improváveis acontecendo juntos ao invés de um único evento de pequena probabilidade. Outro exemplo de evento raro em redes é o caso da perda de células em redes ATM, na ordem de 10^{-6} a 10^{-12} (Onvural, 1995) (de Prycker, 1995).

A Teoria dos Grandes Desvios (*Large Deviation Theory* $- LDT$), avalia pequenas probabilidades em uma escala exponencial (Dembo & Zeitouni, 1998). É um método poderoso para se obter estimativas da probabilidade de certos eventos raros, principalmente para aqueles causados por diferentes motivos acontecendo ao mesmo tempo. A Teoria dos Grandes Desvios pode também prover, e isto é aonde se pretende chegar, uma descrição assintótica da probabilidade de perda que pode ser usada em estimação de banda.

Considerando a variável independente e identicamente distribuída X_i, $i = 1, 2, \ldots, n$ onde $X_i \in \mathbb{R}$ e seja $S_n = X_1 + X_2 + ... + X_n$, então usando a Lei dos Grandes Números, é possível demonstrar que $\frac{S_n}{n}$ converge para a média $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ com probabilidade 1 (Feller, 1971). O limitante superior da Lei dos Grandes Números é dado por $P\{|S_n - M_n| \geq b\}$, para um valor grande de b (Feller, 1971). Um dos principais objetivos da Teoria dos Grandes Desvios é estimar a probabilidade de desvio da média mencionada usando funções geradoras acumuladas e considerando uma certa taxa de convergência. Além do mais, esta teoria estabelece um método sistemático de estimar a função taxa (*rate function*), ou seja, a função que representa o decaimento da cauda da distribuição de M_n. Nas próximas seções serão tratados alguns conceitos da Teoria dos Grandes Desvios, principalmente aqueles que se relacionam com a teoria de banda efetiva.

6.2.1 Princípio dos Grandes Desvios

Os teoremas da Teoria dos Grandes Desvios são geralmente estabelecidos em termos do Princípio dos Grandes Desvios. Seja uma seqüência de variáveis aleatórias $x_1, x_2, \ldots, x_n$ em um espaço de estados Ω e $P(x_n \in C)$, a probabilidade da variável x_n pertencer ao conjunto fechado $C \subset \Omega$. Denotaremos a probabilidade da variável x_n pertencer a um conjunto aberto ou fechado $Z \subset \Omega$ por μ_n, ou seja, $\mu_n = P(x_n \in Z)$. O Princípio dos Grandes Desvios é um conceito que caracteriza o comportamento limite de μ_n, quando $n \to \infty$, em termos de funções de taxa. Esta caracterização ocorre via limitantes exponenciais superiores e inferiores para os valores assumidos por μ_n em subconjuntos mensuráveis de Ω. Aqui Ω é considerado um espaço topológico de tal forma que subconjuntos abertos e fechados de Ω são bem definidos (Dembo & Zeitouni, 1998).

Definição 6.2.1 *Seja uma seqüência de variáveis aleatórias* $y_1, y_2, \ldots, y_n$ *tal que* $y_n \to y$ *em* Ω

quando $n \to \infty$. *Uma função* I *é semicontínua inferior se* $\liminf_n I(y_n) \geq I(y)$.

Definição 6.2.2 *A função de taxa* I *é um mapeamento semicontínuo inferior* $I : \Omega \to [0, \infty)$, *tal que para todo* $\alpha \in [0, \infty)$, *o conjunto de níveis* $\Psi_I(\alpha) \triangleq \{x : I(x) \leq \alpha\}$ *é um subconjunto fechado de* Ω. *Uma função de taxa de interesse conhecida como 'good rate function' é aquela que para todos os conjuntos de níveis* $\Psi_I(\alpha)$ *são subconjuntos fechados de* Ω.

Definição 6.2.3 *A variável aleatória* X_i, $i = 1, \ldots, n$ *satisfaz o Princípio dos Grandes Desvios com função de taxa* I *se para qualquer conjunto fechado* $C \subset \Omega$, *tem-se que*

$$\lim_{n\to\infty} \sup \frac{1}{n} \log \mu_n(C) \leq -\inf_{x\in C} I(x), \tag{6.1}$$

e para um conjunto aberto $G \subset \Omega$, *pode-se escrever:*

$$\lim_{n\to\infty} \inf \frac{1}{n} \log \mu_n(G) \geq -\inf_{x\in G} I(x) \tag{6.2}$$

Pode-se dizer que o processo X obedece ao Princípio dos Grandes Desvios com função de taxa I e velocidade $\frac{1}{n}$ se as equações (6.1) e (6.2) são válidas para os subconjuntos $C \subseteq \Omega$ e $G \subseteq \Omega$. O Princípio dos Grandes Desvios é usado na obtenção de vários resultados como o Teorema Gärtner-Ellis e na formulação da banda efetiva e da probabilidade de perda assintótica.

6.2.2 Função Geradora de Momento

A função geradora de momento é muito importante para a Teoria dos Grandes Desvios. Em si, ela é uma ferramenta útil para o cálculo dos momentos de uma variável aleatória X, estimação da função de densidade de probabilidade e para resolver problemas que envolvam cálculo de soma de variáveis aleatórias (Stark & Woods, 1994).

Definição 6.2.4 *Seja* X *uma variável aleatória assumindo os valores* $x_1, x_2, ..., x_n$, *com probabilidades* $p_1, p_2, ..., p_n$, *respectivamente. A função geradora de momento é dada pela seguinte equação:*

$$G(\theta) = E\{e^{\theta X}\} = \sum_{i=1}^{n} p_i e^{\theta x_i} \tag{6.3}$$

Esta função é definida para todo $\theta \in \mathbb{R}$, e pode ser associada à variável aleatória X ou a sua distribuição, determinando de forma única a distribuição de uma variável aleatória (Ross, 1989).

A função geradora de momentos logarítmica $M_n(\theta)$ também chamada de função geradora de momentos acumulada de X, é dada por

$$M_n(\theta) = \log G(\theta) = \log Ee^{\theta S_n}, \tag{6.4}$$

onde $S_n = (X_1 + \ldots + X_n)$.

Levando em consideração que $M_n'(\theta) = \frac{G(\theta)\prime}{G(\theta)}$, $M_n''(\theta) = \frac{[G(\theta)G(\theta)'' - (G(\theta)')^2]}{G^2(\theta)}$ (aqui $M_n'(\theta)$ e $M_n''(\theta)$ são as derivadas de primeira e segunda ordem de $M_n(\theta)$, respectivamente) e $G(0) = 1$, tem-se que $M_n(0) = 0$, $M_n'(0) = E\{X\}$ e $M_n''(0) = Var\{X\}$. Assim, várias parâmetros estatísticos de um processo pode ser obtido através de sua função geradora de momentos logarítmica $M_n(\theta)$.

Ambas função geradora de momento e função geradora de momento logarítmica são funções convexas. Seja $M_e(\theta)$, a função geradora de momento acumulada escalada definida como (Lewis & Russell, 1998):

$$M_e(\theta) := \lim_{n\to\infty} \frac{1}{n} \log Ee^{n\theta M_n}. \tag{6.5}$$

onde $M_n = \frac{1}{n}(X_1 + \ldots + X_n)$.

Com base no Teorema de Varadhan da Teoria dos Grandes Desvios pode-se demonstrar o seguinte resultado para a função geradora acumulada escalada $M_e(\theta)$ (Lewis & Russell, 1998):

$$M_e(\theta) = \sup_x \{\theta x - I(x)\}. \tag{6.6}$$

A expressão (6.6) é conhecida como Transformada de Legendre de $I(x)$. Quando $X_i, i \in \mathbb{N}$ é uma seqüência i.i.d, $M_e(\theta)$ se reduz à função geradora acumulada $M(\theta)$. A transformada de Legendre é inversível para funções convexas, então se $I(x)$ for uma função convexa, temos a seguinte relação (Lewis & Russell, 1998):

$$I(x) = M_e^*(x) = \sup_\theta \{\theta x - M_e(\theta)\}, \tag{6.7}$$

onde $M_e^*(x)$ representa a transformada de Legendre de $M_e(\theta)$.

6.2.3 O Teorema de Gärtner-Ellis

O Teorema de Gärtner-Ellis assume a existência da função geradora acumulada em escala $M_e(\theta)$ e estabelece limitantes de grandes desvios em $\mathbb{R}^n$ usando as propriedades de $M(\theta)$. Estes limitantes são aplicáveis às variáveis aleatórias que não são independentes e identicamente distribuídas (i.i.d) (Weiss, 1995). Portanto, o Teorema de Gärtner-Ellis estende os resultados da Teoria dos Grandes Desvios para processos com amostras correlacionadas como processos markovianos, poissonianos, etc.

Considere um processo aleatório $X_i \in \mathbb{R}$, $i = 1, \ldots, n$, possuindo função geradora de momento logarítmica dada pela seguinte equação:

$$M_n(\theta) \stackrel{\Delta}{=} \log E\{e^{\langle \theta, X_n \rangle}\}. \tag{6.8}$$

O Teorema de Gärtner-Ellis faz a seguinte suposição quanto a função geradora de momento logarítmica do processo X (Dembo & Zeitouni, 1998):

Suposição 6.2.1- *Para cada $\theta \in \mathbb{R}^d$, a função geradora de momento logarítmica definida como o limite*

$$M(\theta) \stackrel{\Delta}{=} \lim_{n\to\infty} \frac{1}{n} M_n(n\theta) \tag{6.9}$$

existe como um número real, onde $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Definição 6.2.5 *O domínio efetivo de uma função qualquer f é $D_f \stackrel{\Delta}{=} \{x : f(x) < \infty\}$*

Definição 6.2.6 *Uma função convexa $M : \mathbb{R}^n \to (-\infty, \infty)$ é dita ser essencialmente suave ('smooth') se (Dembo & Zeitouni, 1998):*

- $D_M \neq \emptyset$;
- $M(\cdot)$ é diferenciável em D_M;
- $M(\cdot)$ é íngreme, ou seja, $\lim_{n\to\infty} |\nabla M(\theta_n)| = \infty$ sempre que θ_n for uma seqüência em D_M que converge para um ponto de fronteira interior de D_M (Dembo & Zeitouni, 1998). Denotamos por $\nabla M(.)$, o gradiente da função $M(.)$.

Com base no exposto, o Teorema de Gärtner-Ellis é enunciado a seguir.

Teorema 6.2.1 (Teorema de Gärtner-Ellis) *Se a Suposição 6.2.1 é válida, pode-se afirmar que (Dembo & Zeitouni, 1998) (Weiss, 1995) :*

1. *Limitante superior: Seja o intervalo $[a, b]$ tal que $[a, b] \cap D_I \neq \emptyset$ e $-\infty < a < b < \infty$, tem-se que:*

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{S_n}{n} \in [a, b]\right) \leq -\inf_{x\in[a,b]} M^*(x) \tag{6.10}$$

2. *Limite inferior: Supondo que a função $M(\theta)$ seja diferenciável no seu domínio efetivo D_M e que para qualquer $v \in (a, b)$, $-\infty < a < b < \infty$ exista θ_v tal que $M'(\theta_v) = v$, tem-se que:*

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{S_n}{n} \in (a, b)\right) \geq -\inf_{x\in(a,b)} M^*(x), \tag{6.11}$$

onde $M^*(x)$ é a transformada de Legendre de $M(x)$.

Em resumo, pode-se afirmar que se $M(.)$ é uma função essencialmente suave e semicontínua (definição 6.2.1), então o Princípio dos Grandes Desvios se aplica com função de taxa $M^*(\cdot)$ definida pela equação (6.7) (Duffield & O'Connell, 1993b) (Lewis & Russell, 1998). O Teorema de Gärtner-Ellis é importante com relação às formulações da banda efetiva e da probabilidade de perda. Além disso, se um processo de tráfego satisfaz suas condições, pode-se dizer que sua função geradora de momento logarítmica existe, e como conseqüência, sua banda efetiva.

6.2.4 Teoria dos Grandes Desvios e Sistemas de Fila

Consideremos um sistema de fila com um único servidor. Seja Q_i, a quantidade de pacotes em espera na fila no instante de tempo discreto $i \in \mathbb{Z}^+$, V_i o número de pacotes (ou *bytes*) chegando no instante de tempo i, Y_i a quantidade de pacotes que podem ser atendidos e $\{Z_i, i \in \mathbb{Z}^+\}$, um processo ergódico estacionário com $EZ_0 < 0$ representando a diferença $Z_i = V_{i-1}$- Y_{i-1}. Para um instante de tempo i, o tamanho da fila Q_i pode ser encontrado aplicando a equação recursiva de Lindley (Rolls et al., 2005):

$$Q_i = \max\{0, Q_{i-1} + Z_i\}. \tag{6.12}$$

Denotemos por S_n o processo de carga de tráfego representado pela seguinte equação:

$$S_n = \sum_{i=1}^{n} Z_i, \tag{6.13}$$

Se este processo é estacionário e a distribuição de probabilidade de V_i e de Y_i são independentes do tempo, e $EV_i < EY_i$, o tamanho de fila de equilíbrio (em regime) é dado por (O'Connell, 1999):

$$Q = \max_{n \geq 0}\{S_n\} \tag{6.14}$$

É de interesse deste estudo analisar a cauda da distribuição de probabilidade de perda de pacotes (ou *bytes*), ou seja $P\{Q > b\}$ para *buffer* b grande. Tradicionalmente, uma fila é dita estável se a taxa média de chegada for menor do que a taxa média de serviço (Duffy, 2000). Se a fila é estável, então a probabilidade de perda $P\{Q > b\}$ decai muito rapidamente com aumento de b, conforme demonstrado em (Glynn & Whitt, 1994). De acordo com esses autores, ao se considerar um sistema de fila em regime permanente com *buffer* ilimitado sob disciplina FIFO (*First in, First out*) sem prioridades, a seguinte condição para a distribuição assintótica do processo Q pode ser estabelecida

(Glynn & Whitt, 1994):

$$\frac{1}{b}\log P\{Q > b\} \to -\delta \quad \text{para} \quad b \to \infty, \ \text{e} \ \ \delta > 0, \tag{6.15}$$

onde δ é denominado de taxa de decaimento assintótica. Em alguns casos a equação (6.15) é estendida da seguinte forma:

$$P\{Q > b\} \to \alpha^* e^{-b\delta} \ \text{ para } b \to \infty, \tag{6.16}$$

onde α^* é denominado de constante assintótica. Em geral, a equação (6.16) é válida e é uma aproximação natural para a probabilidade de cauda para b não muito pequeno. Para alguns propósitos $\alpha^* \approx 1$ é satisfatório (Glynn & Whitt, 1994). A função de taxa de decaimento assintótica δ representa a taxa de convergência da probabilidade de cauda e também é objeto de estudo da Teoria dos Grandes Desvios.

Nesta parte da seção, serão utilizados alguns resultados da Teoria dos Grandes Desvios para se verificar a equação (6.15). Se o processo de chegada de tráfego e de serviço são estacionários e satisfazem a condição de estabilidade e o processo de carga de tráfego S_n satisfaz o Princípio dos Grandes Desvios com uma função de taxa I dada pela equação (6.7), ou seja, para um valor de x qualquer, tem-se que (Lewis & Russell, 1998):

$$P\left\{\frac{S_n}{n} \approx x\right\} \asymp e^{-nI(x)}. \tag{6.17}$$

Neste mesmo trabalho, Lewis apresentou os seguintes resultados com relação à probabilidade de perda $P\{Q > b\}$ (Lewis & Russell, 1998):

$$P\{Q > b\} = P\left\{\sup_{n\geq 0} S_n > b\right\}; \tag{6.18}$$

$$P\{Q > b\} \leq \sum_{n\geq 0} P\{S_n > b\}. \tag{6.19}$$

Uma vez que $P\{\frac{S_n}{n} > x\} \asymp e^{-nI(x)}$ segundo a equação (6.17), aplicando as equações (6.18) e (6.19), tem-se:

$$P\{S_n > b\} = P\left\{\frac{S_n}{n} > \frac{b}{n}\right\} \asymp e^{-nI(\frac{b}{n})} = e^{-b[\frac{I(b/n)}{b/n}]}, \tag{6.20}$$

e então

$$P\{Q > b\} \asymp e^{-b\frac{I(b)}{b}} + e^{-b\frac{I(b)/2}{b/2}} + ... + e^{-b\frac{I(b)/n}{b/n}} + ... \tag{6.21}$$

O termo que domina na equação (6.21) quando b é grande é aquele para o qual o termo $\frac{I(b/n)}{b/n}$ é o menor, assim esta equação é reescrita da seguinte forma (Lewis & Russell, 1998):

$$P\{Q > b\} \asymp e^{-b\min_x \frac{I(x)}{x}} = e^{-b\delta}, \tag{6.22}$$

o que comprova a equação (6.15).

Sabe-se que a constante de decaimento δ pode ser calculada pela função taxa para o processo de carga S_n (Lewis & Russell, 1998):

$$\delta = \inf_x \frac{I(x)}{x}, \tag{6.23}$$

ou em termos da função geradora de momento logarítmica escalada $M_e(\theta)$:

$$\delta = \sup\{\theta : M_e(\theta) \leq 0\}. \tag{6.24}$$

6.2.5 O Conceito de Banda Efetiva

Considerando um sistema de fila com servidor único, pode-se dizer que a capacidade mínima do servidor que é capaz de atender a condição representada pela equação (6.16) é denominada de Banda Efetiva; este conceito foi introduzido por Frank Kelly (Kelly, 1996).

Seja um enlace com capacidade igual a c e ΔN_k, a quantidade de *bytes* que chega no servidor no instante de tempo k ($N_k = \sum_{r=1}^{k} \Delta N_r$), o tamanho da fila no *buffer* é dado por

$$Q_n \stackrel{d}{=} \sup_{1 \leq k \leq n} E_k, \tag{6.25}$$

onde

$$E_k = N_k - kc. \tag{6.26}$$

Supondo que ΔN_k seja estacionário, a perda de *bytes* medida pela probabilidade de transbordo do *buffer* b é descrita segundo o Teorema de Gartner-Ellis pela seguinte equação (Kelly, 1996):

$$P(Q_n > b) \approx \sup_{1 \leq k \leq n} e^{-kI_n(c+b/k)} = e^{-\inf_k kI_n(c+b/k)} \tag{6.27}$$

onde $I_n(x)$ é a função de taxa e $K_n(\delta)$ é a função pseudo-acumulada, respectivamente dadas por

$$I_n(x) = \sup_{\delta}(x\delta - K_n(\delta)) \tag{6.28}$$

e

$$K_n(\delta) = \lim_{n\to\infty} \frac{1}{n} \log E(e^{\delta \sum_{k=1}^{n} \Delta N_k}). \tag{6.29}$$

Assim, a taxa de serviço c que atende aos requisitos de QoS especificados por $P(Q_n > b) \leq e^{-b\delta}$ para *buffer*s grandes é por definição, a banda efetiva $\alpha(\delta)$ dada por

$$\alpha(\delta) = K_n(\delta)/\delta. \tag{6.30}$$

A partir da equação (6.30) foram determinadas as bandas efetivas para vários modelos de tráfego como fluidos Markovianos, processo de Poisson, etc (Kesidis et al., 1993)(Kelly, 1996). Em (Kesidis et al., 1993), Kesidis et al. demonstram que se o processo de chegada de tráfego N_k satisfaz as condições do Teorema de Gärtner-Ellis e se sua função de taxa $I_n(x)$ for estritamente convexa, a banda efetiva para este este processo existe, ou seja, ela é finita.

A constante de decaimento δ é intensamente referida na literatura como parâmetro de espaço, denotado por s, e assim será denominada daqui em diante. A banda efetiva de um processo $A_i(t)$ normalmente depende do parâmetro s e do tempo t, e é denotada por $\alpha_i(s,t)$, exibindo algumas propriedades interessantes, tais como:

1) Se $A_i(t)$ tem incrementos independentes, então $\alpha_i(s,t)$ não depende de t;

2) Seja $X[0,t] = \sum_i A_i[0,t]$, onde os processos $(A_i[0,t])$ são independentes, então tem-se:

$$\alpha(s,t) = \sum_i \alpha_i(s,t). \tag{6.31}$$

3) Para um valor fixo de t, a banda efetiva $\alpha(s,t)$ é uma função que aumenta com s e seu valor se encontra entre a média e o valor de pico medido no intervalo de tamanho t.

Os parâmetros s e t da banda efetiva não dependem apenas da fonte de tráfego mas também do contexto da conexão, tais como, a capacidade do servidor, tamanho do *buffer*, esquema de escalonamento, parâmetros de QoS, etc. O parâmetro t corresponde à duração mais provável do período de ocupação do *buffer* antes da ocorrência de transbordo e também representa a granularidade mínima necessária ao traço de tráfego para capturar suas propriedades estatísticas que afetam o transbordo do *buffer*. O parâmetro de espaço s representa o potencial de multiplexação estatística de um enlace (Siris, 2000). Um valor grande de s indica que um baixo nível de multiplexação estatística pode ser efetuado, enquanto um valor pequeno para o mesmo corresponde a uma alta capacidade de multiplexação estatística. Basicamente, pode-se dizer que a banda efetiva indica o número de pacotes (ou *bytes*) mais prováveis de ocorrerem no intervalo $[0,t]$ dividido pelo grau de multiplexação que pode ser efetuada neste mesmo período representado pelo parâmetro s.

No caso de processos auto-similares, a distribuição do tamanho da fila tem a forma de uma dis-

tribuição de Weibull, ou seja, $e^{-ax^{\beta}}$, onde $\beta \leq 1$. Para processos auto-similares com $H > \frac{1}{2}$, tem-se que:

$$P\{Q > b\} \approx e^{-bs^{2(1-H)}}. \tag{6.32}$$

A equação (6.32) obtida por Duffield et al. é uma generalização do resultado de Glynn et al. representado pela equação (6.16), e quando apropriadamente escalada satisfaz o Princípio dos Grandes Desvios, permitindo se escrever (O'Connell, 1999) (Rananand, 1996):

$$\lim_{b\to\infty} b^{-2(1-H)} \log P\{Q > b\} = -a^{-2(1-H)} \frac{(a+C)^2}{2}, \tag{6.33}$$

onde $a = C/(H - C)$, com $0.5 < H < 1$, e C é a taxa de serviço.

Por este resultado nota-se que a Teoria dos Grandes Desvios pode ser então aplicada para prover banda efetiva para processos com longa-dependência. Sua aplicação só é limitada a processos com função geradora de momento finita.

6.3 Métodos de Determinação da Banda Efetiva

A banda efetiva pode ser modelada de forma paramétrica, ou seja, assumindo um modelo para a série de tráfego analisada. Algumas equações de banda efetiva analítica são conhecidas, por exemplo, para processos de Poisson, On-Off e ruído gaussiano fracionário (fGn) (Kelly, 1996). O conceito de banda efetiva para redes de alta velocidade foi introduzido independentemente em (Kelly, 1991)(Gibbens & Hunt, 1991), onde foi testado para fontes i.i.d e On-Off. O cálculo de banda efetiva para processos Markovianos e outros processos com curta-dependência são descritos também nos trabalhos (Chang, 1994)(Elwalid & Mitra, 1993). Desenvolvimentos posteriores da teoria de banda efetiva em aplicações como controle de admissão e em reguladores de tráfego podem ser encontrados em (de Veciana et al., 1995) e em muitos outros. Por outro lado, há métodos baseados na 'banda efetiva medida', ou seja, onde não se assume nenhum modelo, mas sim, se obtém uma estimativa de banda efetiva diretamente pela medição da fonte. Entre estes métodos de banda efetiva medida podem ser citados: o estimador direto, o estimador em bloco, o estimador baseado na distância de Kullback-Leibler, o baseado em regressão linear e a banda efetiva empírica (Tartarelli et al., 2000).

Seja $X[0,t]$ o tráfego acumulado durante o intervalo de tempo $[0,t]$ para um fluxo de tráfego. Conforme apresentado anteriormente, a banda efetiva do fluxo de tráfego é definida pela seguinte equação (Kelly, 1996):

$$\alpha(s,t) = \lim_{t\to\infty} \frac{1}{st} \log E(e^{sX[0,t]}), \tag{6.34}$$

que é uma função de $s > 0$. Na teoria de banda efetiva, s representa a taxa de decaimento exponencial

assintótico da distribuição do tamanho da fila conforme contatado na subseção 6.2.5. Ou seja, quando a taxa de serviço é $\alpha(s,t)$, a cauda da distribuição de probabilidade do tamanho da fila Q é dada por

$$P(Q > B) \approx \exp(-sB). \tag{6.35}$$

Note que no cálculo da banda efetiva de um fluxo de tráfego suas propriedades estatísticas assim como os parâmetros do sistema (tamanho do *buffer*, disciplina de serviço) devem ser levados em conta. A banda efetiva é afetada pela estrutura de correlação da fonte e pelo parâmetro de perda que é escolhido tal que os requisitos de QoS das fontes sejam atendidos. É sabido que a banda efetiva designada a diferentes fluxos de tráfego multiplexados tem comportamento aditivo (Chang & Thomas, 1995). Na prática, a estimativa de banda usando a equação (6.34) não é uma tarefa trivial. Esta equação requer uma completa caracterização do processo. Daí surgem métodos para estimá-la tanto parametricamente (a partir de modelos de tráfego), ou através de aproximações que usam os próprios dados de tráfego medidos. As próximas subseções descrevem os principais métodos de estimação de banda efetiva e na seção 6.4 se propõe uma expressão de banda efetiva para o MMW.

6.3.1 Estimador Direto e em Bloco

A banda efetiva dada pela equação (6.34) pode ser estimada, como mostra R.J. Gibbens em (Gibbens, 1996), encontrando a média temporal do processo da seguinte forma:

$$\hat{\alpha}(s,t) = \frac{1}{st} \log \frac{1}{N-t} \int_0^{N-t} e^{s \sum_{i=1}^{N} x_i I(\tau \leq t_i \leq \tau + t)} d\tau \tag{6.36}$$

para $0 \leq \tau \leq N-t$, onde N é o tamanho da série de tráfego e o termo $x_i I(\tau \leq t_i \leq \tau + t)$ representa a quantidade de dados que chegaram durante o intervalo $[\tau, \tau + t]$. Porém, a integral na equação (6.36) pode ainda ser de difícil obtenção numérica. Em (Duffield et al., 1995), o estimador em bloco que facilita o cálculo da equação anterior é proposto, possuindo a seguinte forma:

$$\hat{\alpha}(s,t) = \frac{1}{st} \log \frac{1}{N/t} \sum_{i=1}^{N/t} e^{s \sum_{k=(i-1)i+1}^{it} X_k} \tag{6.37}$$

para um dado t. Este estimador também se baseia na equação (6.34). Em contraste ao estimador direto, o estimador em bloco considera blocos não-sobrepostos de chegadas de dados em um intervalo de tempo *t*. O estimador em bloco atinge uma precisão de estimação comparável à do estimador direto, mas a um menor tempo de processamento (Tartarelli et al., 2000).

6.3.2 Banda Efetiva Empírica

Outra abordagem de estimação de banda efetiva encontrada na literatura que não assume especificamente um modelo para os fluxos de tráfego é a banda efetiva empírica, definida como (Tartarelli et al., 2000):

$$\alpha_{emp}(s,t,N) = \frac{1}{st}\log \hat{E}_{N_t}[e^{sX(0,t)}] \qquad 0 < s; 0 < t < N_t \tag{6.38}$$

onde $X(0,t)$ indica o número agregado de chegadas de dados (*bytes*) dentro de um intervalo de tempo t e $\hat{E}_{N_t}[e^s X(0,t)]$ é a função geradora de momento medida para a série de tráfego com N_t amostras. O termo $\hat{E}_N[e^{\theta X(0.t)}]$ ao ser estimado através da equação (6.37), faz com que a banda efetiva empírica seja equivalente ao estimador em bloco. Para processos de Poisson e On-Off, as bandas efetivas empíricas são muito próximas de suas respectivas bandas efetivas analíticas (Tartarelli et al., 2000).

6.3.3 Banda Efetiva de Courcoubetis

O método descrito por Courcoubetis em (Courcoubetis & Weber, 1996) é baseado na Teoria dos Grandes Desvios e portanto, na suposição de *buffer* grande. A banda efetiva de um processo estacionário X_n com amostras correlacionadas é dada por

$$\alpha = m + \frac{\gamma s}{2b}, \tag{6.39}$$

em que

$$\gamma = \lim_{N\to\infty} \frac{1}{N}\text{var}\left(\sum_{n=1}^{N} X_n\right) = \pi f(0). \tag{6.40}$$

Os parâmetros *m, b, s*, e γ são a taxa média, o tamanho do *buffer*, o parâmetro de espaço e o índice de dispersão, respectivamente. O índice de dispersão γ pode ser estimado a partir dos dados de tráfego através de técnicas de estimação do espectro de potência f graças à relação (6.40). Courcoubetis e Weber ainda afirmam que esta banda efetiva não só é apropriada para fontes de tráfego gaussianas, mas também para outras fontes estacionárias, como por exemplo, para processos auto-regressivos (Courcoubetis & Weber, 1994). A equação (6.39) simplifica o cálculo de banda, expressando-a em termos de dois parâmetros: a taxa média da fonte e seu índice de dispersão. O índice de dispersão mede o quanto de rajadas possui a fonte ('burstiness') (Courcoubetis & Weber, 1994). O parâmetro de espaço s é obtido através de $P(Q_n > b) \leq e^{-\gamma} = e^{-bs}$.

6.3.4 Banda Efetiva usando a Teoria Assintótica de Muitas Fontes

Este esquema de cálculo de banda utiliza uma forma diferente de estimação dos parâmetros de espaço s e de tempo t. Ao invés de se assumir *buffer* grande como é o caso da Banda Efetiva de Courcoubetis, considera-se um número elevado de fontes de tráfego. Sejam M fontes multiplexadas em um servidor com *buffer* de tamanho B, r_j a porcentagem de fluxos do tipo j e e^{-a} a máxima probabilidade de transbordo do *buffer* permitida, então a banda mínima C requerida por estas M fontes pode ser calculada resolvendo a seguinte equação (Courcoubetis et al., 1999):

$$C = \sup_s(\inf_s(R(s,t)), \tag{6.41}$$

em que

$$R(s,t) = \frac{stM\sum_j r_j\alpha(s,t) + a}{st} - \frac{B}{t}. \tag{6.42}$$

O termo $\alpha(s,t)$ é calculado usando a definição de banda efetiva empírica (estimador em bloco). Para um dado t, $R(s,t)$ é uma função unimodal de s que tem um único mínimo. Assim a condição $R(s,t) = R_t(s)$ é resolvida por um método de busca direta descrito em (Courcoubetis et al., 1999). Este processo é repetido para vários valores de t menores do que a janela de tempo de medição e o máximo entre eles é escolhido como a capacidade de serviço a ser alocada.

6.3.5 Banda Efetiva de Norros

Norros introduziu o modelo gaussiano auto-similar fBm para modelagem de tráfego em redes reais (Norros, 1995). Além disso, derivou a banda efetiva para este modelo, que é dada por

$$\alpha = m + K(H)\sqrt{-2\ln(P_{loss})}^{1/H} a^{\frac{1}{2H}} b^{-(1-H)/H} m^{\frac{1}{2H}}, \tag{6.43}$$

onde $K(H) = H^H(1-H)^{1-H}$. Os parâmetros *m, H*, P_{loss}, b e a correspondem, respectivamente, à média, ao parâmetro de Hurst, a probabilidade de transbordo do *buffer*, ao tamanho do *buffer* e ao coeficiente de variação. Quando o tráfego possui curta-dependência, o parâmetro a pode ser aproximado pelo índice de dispersão γ. A banda efetiva de Norros leva em consideração a auto-similaridade através do parâmetro de Hurst, sendo então apropriada para tráfego com longa-dependência, produzindo melhores estimativas de banda para o caso de *buffer* grande.

6.4 Banda Efetiva para o MMW

Esta seção explora alguns conceitos da Teoria dos Grandes Desvios e certas aproximações estatísticas com a finalidade de desenvolver uma expressão para a estimação de banda efetiva de modelos multifractais baseados em cascatas multiplicativas, perfeitamente aplicável ao MMW proposto. A novidade aqui introduzida é que a banda efetiva pode ser dada em função dos multiplicadores destes modelos. Assim, uma vez determinados os multiplicadores do modelo MMW aplicado a uma determinada série de tráfego, sua banda efetiva pode ser diretamente calculada. Como os multiplicadores do MMW podem ser encontrados através de seu conjunto de parâmetros (α, γ, ρ), pode-se dizer que esta proposta de banda efetiva, de uma forma inédita, leva em consideração parâmetros multifractais. Além disso, tira-se proveito desse fato na seção 6.5, onde se propõe um algoritmo adaptativo de cálculo de banda baseado no algoritmo de estimação adaptativa dos parâmetros (α, γ, ρ) apresentado no Capítulo 3. Dito isto, a seguinte proposição estabelece a equação de banda efetiva proposta.

Proposição 6.4.1 *A banda efetiva para um processo multifractal baseado em cascata como o MMW, é dada, em termos de seus multiplicadores $A_{i,j}$, por:*

$$\alpha(s,t) = \frac{1}{st} \log \left(\frac{\sum_{k=1}^{2^N-1} |wg_{N,k}|}{2^N} \right) \tag{6.44}$$

onde

$$wg_{N,k} = 2^{-N/2}[e^{sv1} - e^{sv2}] \tag{6.45}$$

com

$$v1 = 2^{-N/2} U_{0,0} \prod_{i=0}^{j-1} [1 + (-1)^{k'_i} A_{i,2k_i}], \tag{6.46}$$

$$v2 = 2^{-N/2} U_{0,0} \prod_{i=0}^{j-1} [1 + (-1)^{k'_i} A_{i,2k_i+1}], \tag{6.47}$$

e N é o número de estágios da cascata.

Demonstração Primeiramente, consideremos uma notação matemática mais precisa para as cascatas multiplicativas descritas no Capítulo 2. Seja *T* um cubo unitário em $\mathbb{R}^d$ para $d \geq 1$, dividido repetidamente em um número b^n ($b \geq 2$) de sub-cubos ('pixels') de volume b^{-n}, $n = 1, 2, \ldots$, tais que a massa do *pixel* $\Delta_n(t_1, t_2, ..., t_n)$, $t_k \in \{0, 1, ..., b-1\}$ no estágio *n* é dada pela medida aleatória $\lambda_n(\Delta_n(t_1, t_2, ..., t_n)) = b^{-n} \prod_{k=1}^{n} W_{t_1,...,t_k}$, onde W_v's são os geradores (multiplicadores) da cascata,

uma família de variáveis i.i.d não-negativas de média 1, indexadas pelo endereço do pixel *v* em diferentes escalas mais finas b^{-n} para $n \geq 1$. A medida λ_∞ da cascata é obtida pelo limite da seqüência λ_n para $n \longrightarrow \infty$.

A função geradora de momentos acumulada modificada (também conhecida como função de estrutura) para $\log W_v$ é definida como (Ossiander & Waymire, 2002):

$$\Omega_b(h) = \log_b E[W_v^h 1[W > 0]] - (h-1) \tag{6.48}$$

que descreve a distribuição dos multiplicadores W_v's. A função $\Omega_b(h)$, por sua vez, pode ser estimada através de ξ_n^h (Ossiander & Waymire, 2002):

$$\xi_n^h = \frac{1}{n} \log_b \sum_{\Delta_n} \lambda_\infty^h(\Delta_n(t_1, t_2, ..., t_n)) \tag{6.49}$$

Ou seja, ξ_n^h converge para $\Omega_b(h)$ à medida que $n \to \infty$ para todo $h \in \mathbb{Z}^+$ e cascatas para as quais $E\lambda_\infty^h(T) > 0$ (Resnik et al., 2003).

De modo alternativo, a função de estrutura $\Omega_b(h)$ pode também ser estimada para uma série de tamanho 2^l através da função $\hat{\xi}(q,l)$ obtida pelas seguintes equações (Feldmann et al., 1998)(Resnik et al., 2003):

$$Z(q,l) = \sum_{n=0}^{2^l-1} |d_{-l,n}|^q \tag{6.50}$$

$$\hat{\xi}(q,l) = \frac{\log(Z(q,l))}{l} \tag{6.51}$$

onde $d_{-l,n}$ são os coeficientes *wavelet* de um processo estocástico *X*, por definição (Apêndice 1), dados por

$$d_{-l,n} = \int_0^1 \varphi_{-l,n}(x) X_\infty(dx). \tag{6.52}$$

Na construção da cascata, após a divisão de T em l partes, obtém-se subintervalos $I(j_1, j_2, ..., j_l)$ iguais de comprimento 2^{-l} indicados por

$$I(j_1, j_2, ..., j_l) = [\sum_{k=1}^{l} \frac{j_k}{2^k}, \sum_{k=1}^{l} \frac{j_k}{2^k} + \frac{1}{2^l}), \tag{6.53}$$

onde j_k denota o estágio k da cascata.

Sejam os primeiros *l* valores de j representados da seguinte maneira:

$$j/l = (j_1, j_2, ..., j_l) \tag{6.54}$$

Com base nas equações (6.53) e (6.54), define-se o processo g como sendo a versão exponencial do processo multifractal λ_∞ baseado em cascata multiplicado por s. Em outras palavras, pode-se expressar o processo g pela seguinte equação:

$$g_\infty(j/l) = e^{s\lambda_\infty(I(j/l))} \tag{6.55}$$

Usando *wavelets* de Haar, os coeficientes *wavelet* podem ser explicitamente expressos para o processo $X(k)$ a tempo discreto k por

$$d_{-l,k} = 2^{-l/2}[X(2^l k) - X(2^l k + 1)]. \tag{6.56}$$

Recorrendo ao Capítulo 2, verifica-se que para o modelo MMW na escala mais fina j, as seguintes expressões são válidas:

$$X(k) = 2^{-j/2} u_{j,k} \tag{6.57}$$

e

$$u_{j,k_j} = 2^{-j/2} U_{0,0} \prod_{i=0}^{j-1} [1 + (-1)^{k'_i} A_{i,k_i}] \tag{6.58}$$

onde $A_{i,k}$ são os multiplicadores da cascata e k_i é dado pela equação (2.69).

Seja $wg_{j,k}$ os coeficientes *wavelets* para o processo *g* dado pela equação (6.55). Fazendo $j = N$ e ao se substituir as equações (6.58) e (6.57) em (6.56), pode-se escrever os coeficientes $wg_{N,k}$ como

$$wg_{N,k} = 2^{-N/2}[e^{sv1} - e^{sv2}], \tag{6.59}$$

onde

$$v1 = 2^{-(N)/2} U_{0,0} \prod_{i=0}^{j-1} [1 + (-1)^{k'_i} A_{i,2k_i}] \tag{6.60}$$

e

$$v2 = 2^{-(N)/2} U_{0,0} \prod_{i=0}^{j-1} [1 + (-1)^{k'_i} A_{i,2k_i+1}]. \tag{6.61}$$

Lembrando que a banda efetiva de um processo X pode ser estimada, segundo o método de banda efetiva empírica, da seguinte forma (Tartarelli et al., 2000):

$$\alpha(s,t) = \frac{1}{st} \log(E[e^{sX(0,t)}]) = \frac{1}{st} \log \frac{1}{b_t} \sum_{i=1}^{b_t} e^{sX(0,i)} \tag{6.62}$$

ou

$$\alpha(s,t) = \frac{1}{st}\log\frac{1}{b_t} + \frac{1}{st}\log\sum_{i=1}^{b_t} e^{sX(0,i)}, \tag{6.63}$$

onde $b_t = 2^N$ é o número de amostras de X, sendo que N é o número de estágios da cascata multiplicativa. Note que o segundo termo da equação (6.63), $\log\sum_{i=1}^{b_t} e^{sX(0,i)}$, pode ser estimado usando a equação (6.49) aplicada a um processo g. Uma vez que a função de estrutura (6.49) pode ser estimada via $Z(q,l)$, o segundo termo do lado direito de $\alpha(s,t)$ é estimado usando $2^N\hat{\xi}(1,l)$ onde

$$\hat{\xi}(1,l) = \log\left(\sum_{n=0}^{2^l-1}|d_{-l,n}|\right). \tag{6.64}$$

Finalmente, inserindo a equação (6.59) em (6.64), é possível calcular a banda efetiva $\alpha(\theta,t)$ para um série de tráfego de tamanho 2^N, por meio dos multiplicadores da cascata pela seguinte equação:

$$\alpha(\theta,t) = \frac{1}{\theta t}\left(\log\frac{\sum_{k=1}^{2^N-1}|wg_{j,k}|}{2^N}\right). \tag{6.65}$$

■

A derivação da banda efetiva proposta parte de uma teoria de convergência bem definida, a qual considera que a função ξ_n^h tende à função geradora de momento acumulada modificada $\Omega_b(h)$ para $n \to \infty$. Esta função geradora de momento acumulada modificada por sua vez, após um certo desenvolvimento, pode ser estimada através dos multiplicadores da cascata.

A Proposição 6.4.1 afirma que a equação (6.65) é aplicável para modelos baseados em cascata, uma vez que o modelo MMW pode ser visto como uma cascata aleatória binomial. Seja $W_{k_n}^n$ um multiplicador de uma cascata binomial na escala j. Realmente, fazendo $W_0^0 = U_{0,0}$ e

$$W_{k_n}^n = \frac{1+(-1)^{k'_{n-1}}A_{n-1,k_n-1}}{2}, \tag{6.66}$$

pode-se demonstrar que os incrementos da cascata binomial coincidem com os incrementos do processo MMW, cujos multiplicadores são $A_{j,k}$ (Riedi et al., 1999).

Com propósito de comparação, foram escolhidos os métodos de estimação EEB (Empirical Effective Bandwidth)(Tartarelli et al., 2000) e NEB (Norros Effective Bandwidth)(Norros, 1995). O primeiro por ser um método não-paramétrico, ou seja, não se assume nenhum modelo para o cálculo de banda. Já o segundo, por levar em consideração a longa-dependência presente nos dados de tráfego. A Figura 6.1 mostra as bandas efetivas calculadas via equação (6.65), e através dos métodos EEB e NEB para o traço de tráfego dec-pkt-2. O tamanho do *buffer* foi ajustado em 64Kbytes e tomou-se

como probabilidade da perda de byte desejada, 10^{-6}. O traço de tráfego dec-pkt-2 foi usado nesta simulação por ser monofractal com $H = 0.8$, apropriado assim para o método NEB, que é baseado em processos monofractais. Note que o método proposto por Norros (NEB) tem uma característica mais conservadora (Norros, 1995), confirmada pela Figura 6.1. Os resultados do método EEB e o do nosso são similares. As bandas efetivas calculadas pelos três métodos versus o tamanho do *buffer* são apresentadas na Figura 6.2. O mesmo é apresentado na Figura 6.3 para a série de tráfego 4-7-I-9 na escala de tempo de 100ms, com a seguinte configuração: $B = 6500$bytes e taxa de perda igual a 10^{-5}. Para *buffer*s de tamanhos pequenos, o NEB apresentou um valor elevado de banda efetiva. As estimativas de banda obtidas com a equação proposta e com o método EEB se alteraram menos em relação à variação do tamanho do *buffer* do que as estimativas dadas pelo método NEB. Nota-se então que os resultados da equação de banda efetiva proposta são menos conservadores do que os de Norros e acompanham de perto os resultados da banda efetiva empírica. A principal diferença entre a banda de Norros e a equação de banda efetiva proposta é que a função geradora de momento $\log \sum_{i=1}^{b_t} e^{sX(0,i)}$, parte essencial da equação de banda efetiva, é calculada através dos parâmetros do MMW, mais especificamente, pelos seus multiplicadores.

No intuito de se comprovar que o cálculo de banda efetiva proposto atende aos requisitos de perda, simulou-se computacionalmente um enlace servindo o traço de tráfego dec-pkt-3 (2^{16} amostras) na escala de tempo de 100ms. Estabeleceu-se como capacidade do servidor, o valor de banda efetiva dado pela equação (6.65), uma taxa da perda desejada de 10^{-4}. A Figura 6.4 comprova que a banda efetiva proposta satisfaz a restrição de probabilidade de perda desejada para diferentes tamanhos de *buffer*. Uma vez que a aproximação de Norros é mais conservadora do que a nossa, ela também atinge o objetivo de perda mas com um maior desperdício de banda.

6.5 Cálculo Adaptativo de Banda Efetiva

A alocação estática de taxa ou banda em um servidor assumindo um modelo, mesmo sendo multifractal, pode ser ineficiente quanto à utilização do enlace por causa da natureza imprevisível do tráfego cujo comportamento é variante no tempo. Devido ao grande número de rajadas presentes em tráfego correlacionado, como por exemplo, o tráfego de vídeo, é difícil manter os requisitos de retardo e de perda ao mesmo tempo em que se obtém uma alta utilização de enlace pela alocação de uma taxa fixa. A alocação dinâmica de banda tem como objetivo resolver este problema, fornecendo mais taxa quando for necessário e é vista como uma solução promissora na prática para se prover QoS. Mas para que haja alocação eficiente de taxa, é preciso que a banda efetiva da fonte seja estimada, de forma a atender aos requisitos adequadamente e atingir uma maior utilização dos recursos da rede. Como exemplo de trabalhos neste sentido, poucos podem ser citados, entre eles, V.Solo et al.

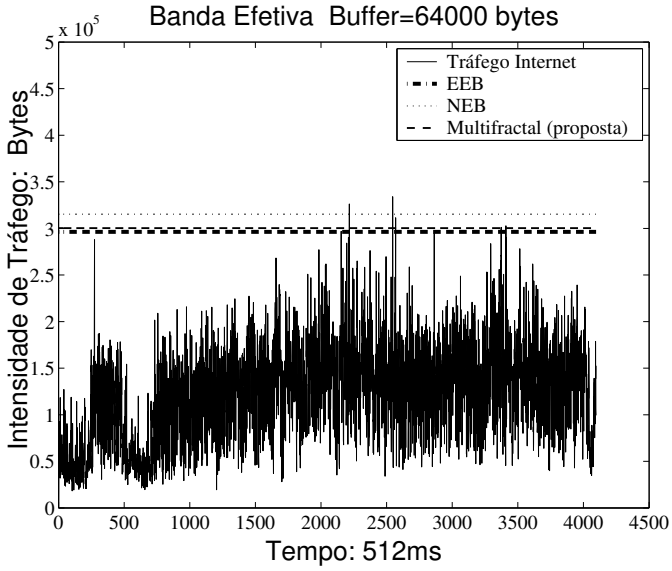


Fig. 6.1: Cálculo de banda efetiva para a série de tráfego dec-pkt-2.

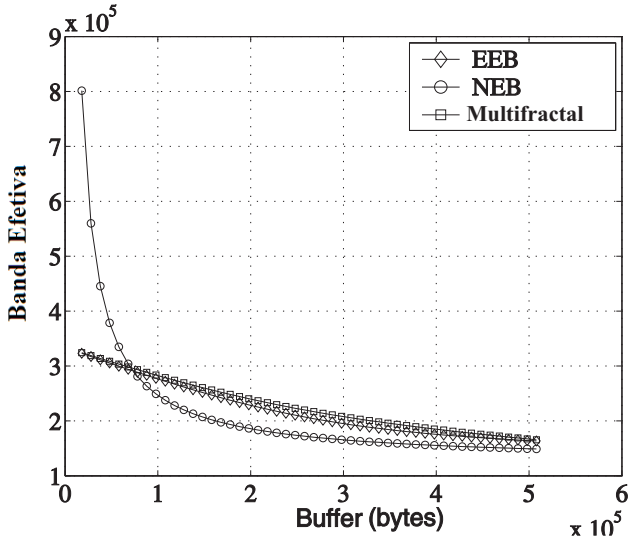


Fig. 6.2: Banda efetiva x Tamanho do *buffer* (série de tráfego dec-pkt-2)

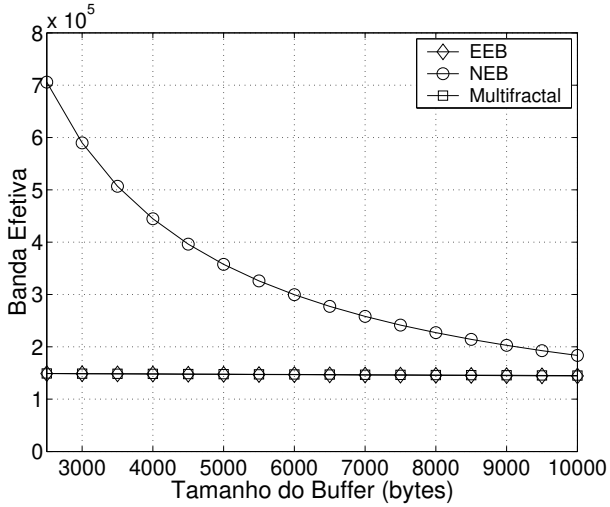


Fig. 6.3: Banda efetiva x Tamanho do *buffer* (série de tráfego Petrobrás 4-7-I-9)

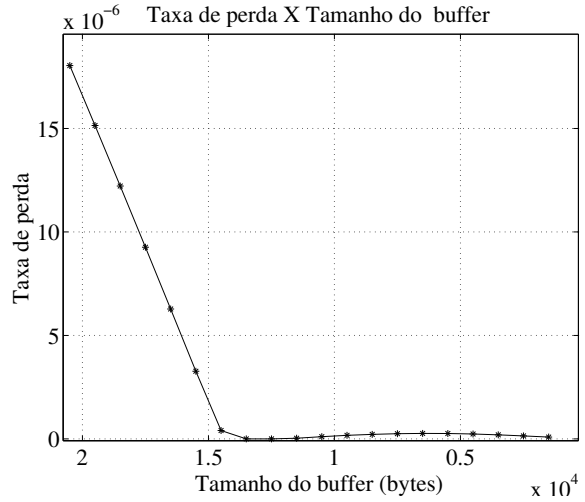


Fig. 6.4: Taxa de perda obtida com a equação de banda efetiva proposta x Tamanho do *buffer* (série de tráfego dec-pkt-3)

propuseram uma técnica de estimação adaptativa de banda efetiva aplicada a representações lineares de processos estocásticos. Porém, esta abordagem limita a aplicação deste método, o qual não foi testado para tráfego real (Solo, 1996).

O MMW possui uma estrutura analítica adequada para modelagem 'on-line'. Aproveitando este fato, nesta seção, é desenvolvido um procedimento para cálculo de banda efetiva em tempo real levando em consideração os parâmetros do MMW. Note que a equação de banda efetiva proposta (6.65) é dada em função dos multiplicadores do MMW, os quais podem ser obtidos a partir dos parâmetros de entrada deste modelo. Assim, para efetuar estimação em tempo real de banda efetiva é preciso que se atualizem periodicamente os valores dos parâmetros α, γ e ρ do MMW obtidos através da função de escala $t(q)$ e pelo fator de momento $c(q)$. Um método para estimação adaptativa desses parâmetros foi proposto no Capítulo 3. Nesta seção, este método é incorporado em um algoritmo de estimação adaptativa de banda efetiva baseado na equação proposta (6.65) e no MMW.

No cálculo adaptativo de banda efetiva, atualizam-se periodicamente os valores de α, γ, e ρ a cada instante de tempo k, fazendo $t = 1$ nas equações (3.49) e (3.50) do Capítulo 3. A seguir, é apresentado o algoritmo que obtém valores atualizados dos multiplicadores do MMW para cada instante de tempo diádico k e por fim realiza estimativas em tempo real de banda efetiva:

Algoritmo 6.5.1 *Algoritmo de Estimação Adaptativa de Banda Efetiva*

1. *Sejam* $p_{1,0} = 1$, $\underline{\hat{a}}_0^q = [0\ 0]$, $k = 0, ..., 2^N$ *e* $q = 1, \ldots, q_2$*;*

2. *Calcule* $\underline{\hat{a}}_k^q = [\hat{\tau}_0(q) \log \hat{c}(q)]$ *para cada valor de* q *usando as seguintes equações recursivas:*

$$p_{q,k} = p_{q,k-1} - p_{q,k-1}\underline{x}_k[1 + \underline{x}_k^T p_{k-1}\underline{x}_k] - \underline{x}_k^T p_{q,k-1} \tag{6.67}$$

$$\underline{\hat{a}}_k^q = \underline{\hat{a}}_{k-1}^q - p_k[x_k x_k^T \underline{\hat{a}}_{k-1}^q - \underline{x}_k y_k^q] \tag{6.68}$$

   onde $y_k^q = \log E(|\bar{X}_k|)^q$, $\bar{X}_k = [X_1 \quad X_2 \quad \ldots \quad X_k]$ *e* $\underline{x}_k = [1 \quad \log 2 \quad \ldots \quad \log k]$*;*

3. *Estime o parâmetro* α *de* τ_0 *(equação 3.5) usando o algoritmo de Levenberg-Marquardt de acordo com a seguinte regra de atualização (Marquardt, 1963):*

$$\alpha_{i+1} = \alpha_i - (H_{es} + \eta diag(H))^{-1} \nabla\phi(\alpha_i) \tag{6.69}$$

   onde H_{es} *é a matriz Hessiana* $(H_{es} = \nabla^2\phi(\alpha_i))$ *e* η *é um parâmetro de controle na iteração* i *do algoritmo de Levenberg-Marquardt;*

4. *Aplique novamente o algoritmo de Levenberg-Marquardt para estimar os parâmetros* ρ *e* γ *da função* $c(q)$*:*

$$c(q) = e^{\rho q+\gamma^2 q^2/2} 2^{N\left(q - log_2 \frac{\Gamma(\alpha)\Gamma(2\alpha+q)}{\Gamma(2\alpha)\Gamma(\alpha+q)}\right)};$$

5. *Faça $j = 1$, o que corresponde à fração de agregação igual a $2^j = 2$;*

6. *Calcule a variância $var[X^m]$ do processo agregado, onde $m = 2^j$, pela equação (3.12):*

$$var[X^{2^j}] = e^{2\rho+2\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right)^{N-j} - (e^{2\rho+\gamma^2}2^{2j-2N})$$

7. *Calcule Z_j com o auxílio da equação (3.19) e do valor de $var[X^m]$ estimado no passo anterior:*

$$Z_j = 2^{j-1}[var[X^{2^j}] - e^{2\rho+2\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right) + (e^{2\rho+\gamma^2})];$$

8. *Estime o momento de segunda ordem dos coeficientes* wavelet *$E(w_j^2)$ por meio da equação (3.18):*

$$E(w_j^2) = e^{4\rho+4\gamma^2}\left(\frac{\alpha+1}{\alpha+1/2}\right)^2 - 2Z_j$$

9. *Incremente j de 1. Se $j = \log_2(k)$, calcule a média e a variância dos coeficientes de escala nesta escala usando as equações (3.20) e (3.21), caso contrário, vá para o passo 5;*

10. *Calcule $p'_j s$ usando a equação (2.67) do Capítulo 2;*

11. *Aqui começa um procedimento recursivo usando os valores de $p'_j s$. Faça $j = 0$ e calcule os coeficientes de escala $U_{0,0}$ na maior escala;*

12. *Gere variáveis aleatórias $A_{j,k}$ e calcule os coeficientes wavelet $(W_{j,k})$ na escala j usando a equação (2.66) do Capítulo 2;*

13. *Usando os multiplicadores obtidos $A_{j,k}$, atualiza-se a função geradora de momento $M(k)$ no instante de tempo k e escala $j = \log_2(k)$:*

$$M(k) = \left(1 - \frac{1}{k}\right) M(k-1) + \frac{1}{k}\sum_{h=k-1}^{k} 2^{-\log_2(h)/2}[e^{sv1} - e^{sv2}] \tag{6.70}$$

onde

$$v1 = 2^{-(j)/2} U_{0,0} \prod_{i=0}^{j-1}[1 + (-1)^{k'_i} A_{i,2k_i}] \tag{6.71}$$

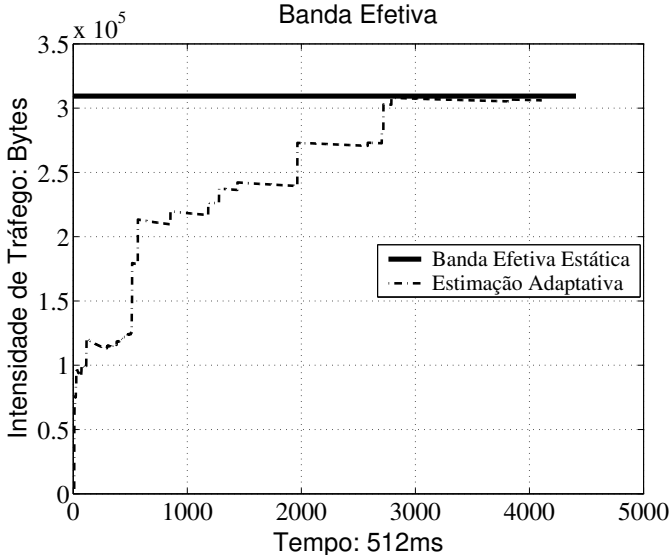


Fig. 6.5: Estimação adaptativa de banda efetiva.

$$v2 = 2^{-(j)/2} U_{0,0} \prod_{i=0}^{j-1} [1 + (-1)^{k'_i} A_{i,2k_i+1}] \tag{6.72}$$

14. A banda efetiva no instante de tempo k é dada por

$$\alpha(s,k) = \frac{1}{sk} \log(M(k)). \tag{6.73}$$

Na validação do algoritmo proposto para estimação adaptativa de banda efetiva, estabeleceu-se arbitrariamente as seguintes condições para o teste: o tamanho do *buffer* de $B = 50K$ *bytes* e probabilidade de perda de *bytes* $P(Q > B) = e^{-Bs}$ igual a 10^{-7}. A Figura 6.5 compara as estimativas de banda efetiva para um traço de tráfego TCP-IP (dec-pkt-2 com parâmetro de Hurst igual a 0.80) obtidas com a equação de banda efetiva proposta e sua versão adaptativa representada pelo algoritmo 6.5.1. A banda efetiva adaptativa é aproximadamente igual ao valor da banda efetiva calculada de forma estática, onde esta última é baseada em todas as amostras da série de tráfego. Pode-se notar portanto, que a banda efetiva adaptativa atinge o valor dado pela banda efetiva estática em um dado instante de tempo, como se fosse calculada estaticamente usando todas as amostras de tráfego até o instante de tempo em questão. Como conseqüência, o método adaptativo requer menos tempo de processamento e menor capacidade de armazenamento. A próxima subseção aplica o princípio em questão de cálculo adaptativo de banda efetiva para desenvolver um esquema de provisão adaptativa de banda.

6.5.1 Esquema de Provisão Adaptativa de Banda

É uma tarefa desafiadora prover QoS para aplicações de rede e ao mesmo tempo manter alta a utilização da rede. Devido a este motivo, a provisão de banda em redes de alta velocidade precisa ser dinâmica, adaptativa e baseada em medições de tráfego de modo a atingir um uso mais eficiente dos recursos da rede. Um algoritmo de provisão de banda pode ser classificado de acordo com algumas características como QoS e predição de tráfego. O esquema de provisão de banda proposto nesta seção focaliza em atender QoS em termos da probabilidade de perda em um enlace para os fluxos e se baseia na banda efetiva de um modelo de tráfego multifractal calculada adaptativamente. Note que neste caso, algoritmos de predição de tráfego não são usados, diferente do que foi feito no Capítulo 5. O algoritmo de cálculo *on-line* de banda efetiva descrito na seção anterior foi usado para implementar um completo esquema de provisão adaptativa de banda. Em nosso esquema de provisão de banda, o tráfego é medido a cada 'slot' de tempo t_{slot}. A alocação de banda é efetuada em uma escala de tempo maior, a cada $N_t * t_{slot}$ instante de tempo, conhecido como janela de redimensionamento (*resizing window*). Aplica-se o algoritmo adaptativo de estimação de banda efetiva da seção anterior para atualizar o valor da banda efetiva calculada para determinada série de tráfego a cada instante t_{slot} que compõe a janela de redimensionamento.

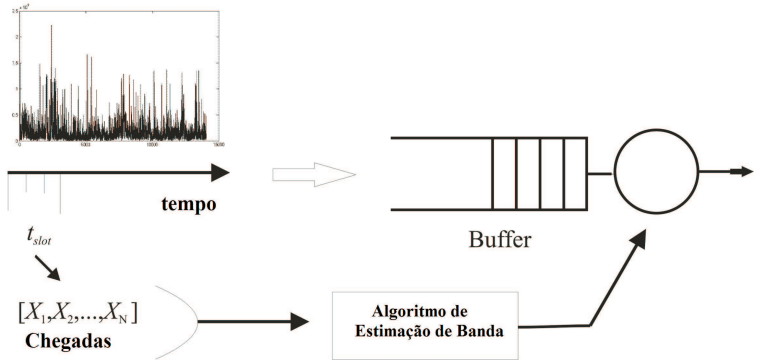


Fig. 6.6: Esquema de provisão adaptativa de banda

Na literatura, poucos trabalhos lidam com alocação adaptativa de banda com garantia de qualidade de serviço. A maioria simplesmente recorre à utilização do enlace como fator base para o provimento de banda (Tran & Ziegler, 2005). Com propósito de comparação, implementamos o esquema de provisão de banda de Duffield que considera $c_j = m_j + d\sqrt{v_j}$ como banda requerida na janela de redimensionamento j onde m_j e v_j são, respectivamente, a média e a variância da carga de tráfego na janela atual j. O parâmetro d está relacionado com a probabilidade de taxa de sobrecarga (*rate overload probability*) (Duffield et al., 1999). Deve-se ressaltar que o esquema de Duffield requer o conhecimento com antecedência da média e da variância da janela de redimensionamento, além de

também supor que o tráfego de entrada seja gaussianamente distribuído.

Para se avaliar quantitativamente os esquemas de provisão considerados nesta seção, adotou-se o Fator de Adequação Médio (Average Goodness Factor - AGF) introduzido em (Tran & Ziegler, 2005). O AGF mede quão rápido e próximo a banda provida segue a dinâmica do tráfego real e ainda assegura os parâmetros de QoS alvos. A idéia básica é que sob um esquema ideal de provisão, a utilização do enlace deveria se manter constante em um nível ótimo fixo u_{opt}. O valor de u_{opt} é escolhido como sendo a média das taxas de utilização atingidas em todas as janelas de redimensionamento utilizando a definição de banda efetiva com todos as amostras de tráfego dessas janelas; atendendo portanto aos requisitos de QoS.

Para uma dada janela de redimensionameto j, denotemos l_j a capacidade do enlace alocada e r_j a real taxa de tráfego agregado. O Fator de Adequação (Goodness Factor (GF)) é definido como (Tran & Ziegler, 2005):

$$GF_j = \begin{cases} \frac{(l_j - r_j)/r_j}{u_{opt}} & \text{if } l_j \leq r_j \\ \frac{r_j/l_j}{u_{opt}} & \text{if } l_j > r_j \; and \; r_j/l_j \\ \frac{u_{opt}}{r_j/l_j} & \text{if } l_j > r_j \text{ and } r_j/l_j > u_{opt} \end{cases} \tag{6.74}$$

O AGF é a média de todos os valores individuais dos Fatores de Adequação GF em todas as janelas de redimensionamento consideradas. Quanto mais alto o grau de sobreprovisão (banda alocada acima da realmente necessária) ou subprovisão de banda (banda alocada abaixo da realmente necessária), menor o valor de AGF. O valor de AGF mínimo é $-1/u_{opt}$, e o máximo é 1. Além do mais, quanto mais o AGF é próximo de 1, melhor o esquema de provisão de banda.

O esquema adaptativo de provisão de banda efetiva proposto foi testado experimentalmente com várias séries de tráfego em comparação ao esquema de Duffield e a alocação de banda efetiva estática. No caso do traço de tráfego dec-pkt-1 com 2^{18} amostras, estabeleceu-se $t_{slot} = 10ms$ e $N_t = 256$. A Tabela 6.1 mostra os valores de AGF obtidos nas simulações para este traço de tráfego, onde fixou-se como probabilidade de perda desejada, 10^{-3} e tamanho de *buffer* igual a $B = 55K$bytes. O método de provisão de Duffield considera um modelo sem *buffer* onde a média e a variância devem ser conhecidas em cada janela de redimensionamento. A Tabela 6.1 mostra que o esquema de provisão em tempo real de banda proposto consegue satisfazer a probabilidade de perda alvo, de fato, uma menor taxa de perda é obtida do que a requisitada de 10^{-3} para a série inteira de tráfego. O esquema adaptativo e em tempo real de alocação de banda pode ser empregado para atender aos parâmetros de QoS requisitados, provendo um valor de AGF maior para a rede do que a alocação de banda estática, a qual requer a disponibilidade de toda série de tráfego. Além disso, nota-se um valor de AGF ligeiramente maior para o esquema de provisão proposto do que o esquema de Duffield.

Para um completo estudo dos algoritmos adaptativos de provisão de banda efetiva ainda são

Tab. 6.1: AGF e Probabilidade de Perda para a série de tráfego dec-pkt-1.

Esquema de provisão de banda	AGF	Taxa de perda
Estático	0.3374	0
Duffield	0.3680	$8.5101\text{x}10^{-4}$
Proposto	0.3734	$7.8082\text{x}10^{-4}$

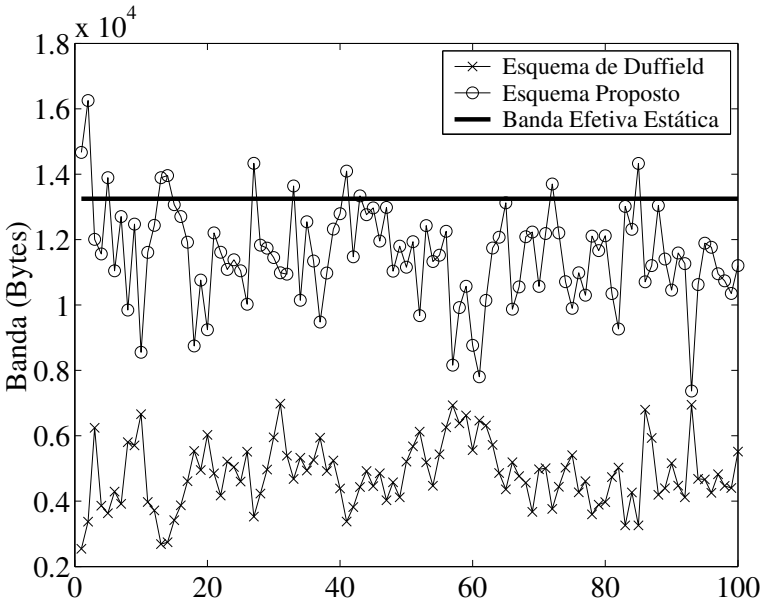


Fig. 6.7: Provisão de banda para a série de tráfego dec-pkt-1

necessárias investigações sobre os efeitos do tamanho do *slot* de tempo usado; o que serão deixadas para trabalhos futuros. Em (Drummond, 2005), pode-se encontrar uma comparação entre alguns métodos de cálculo de banda efetiva para diferentes tamanhos de *slots* de tempo.

6.5.2 Emprego do Esquema de Provisão Adaptativo de Banda

Nesta subseção são listados alguns cenários típicos das redes atuais para os quais o emprego do esquema de provisão proposto é conveniente.

Redimensionamento de LSPs em redes MPLS

No contexto de redes MPLS (*Multiprotocol Label Swicthing*), os LSPs (*Label Switching Paths*) podem ser vistos como túneis que acomodam o tráfego sobre o domínio MPLS. Aplicando o esquema

de provisão de banda proposto, para um dado túnel, o operador pode redimensionar adaptativamente a banda alocada de modo a atender aos requisitos de QoS do tráfego de entrada. Isso irá envolver o emprego de monitoramento periódico de tráfego no roteador de ingresso do LSP e a execução do nosso esquema de designação de banda.

Note que sinalização é requerida para a alocação de banda. Para tal, pode-se explorar as mensagens de 'refresh' do RSVP-TE (Resource Reservation Protocol: Trafic Extension) (Awduche et al., 2001), o protocolo atual de sinalização para o MPLS que trata deste propósito. Assim, não são necessários cabeçalhos de sinalização a mais em relação a arquitetura de rede MPLS atual com RSVP-TE para se atingir provisão adaptativa de banda com QoS.

Reconfiguração de contrato de serviço entre domínios Diffserv

Um modo de prover parâmetros de QoS sobre múltiplos domínios de rede é utilizar a arquitetura Diffserv em domínios de rede individuais e empregar contrato de serviço SLA (*Service Level Agreement*) para controlar o tráfego entre domínios vizinhos (Blake, 1998). Tais contratos especificam o volume de tráfego que pode ser admitido no enlace entre os domínios. Para assegurar qualidade de serviço ponto a ponto, VLLs (*Virtual Leased Lines*) podem ser estabelecidos, onde o esquema de provisão adaptativo se aplica no redimensionamento de tais VLLs. A sinalização necessária aos ajustes de banda novamente é resolvido por requisições envolvendo um roteador de borda e um entidade de gerenciamento como um BB (*Bandwidth Broker*). O roteador de borda de um domínio realiza monitoração de tráfego e inicializa o pedido de banda ao BB. O *Bandwidth Broker* processa este pedido, atualiza sua base se necessário e envia uma confirmação de reconhecimento ao roteador.

Redimensionamento de túneis em redes VPN

Redes virtuais privadas (VPN-*Virtual Private Networks*) têm como conceito servir um grupo especial de usuários. Os provedores de serviço reservam bandas entre pontos de borda para atender a demanda de tráfego. A banda é alocada usando o modelo de tubo (*pipe*) (Lim et al., 2001). O esquema de provisão adaptativo de banda pode ser uma solução eficiente para o gerenciamento dos tubos. De maneira similar aos LSPs, o procedimento de sinalização para adaptação de banda depende da tecnologia de uma dada VPN. O protocolo de sinalização pode ser o Beagle (Lim et al., 2001), ou o RSVP, com os quais a capacidade de monitoração de tráfego é requerida apenas no roteador de ingresso do enlace virtual.

Redimensionamento de enlaces lógicos em redes SON

A rede SON (*Service Overlay Network*) é uma alternativa a provisão de parâmetros de QoS fim a fim (Duan et al., 2002). Nesta rede, *gateways* de serviço são empregados para interconectar os domínios por enlaces lógicos. Cada enlace lógico é na verdade um caminho que inclui uma série de enlaces IP. Ajustando a banda dos enlaces lógicos da rede SON, se atinge um gerenciamento eficiente da banda o que torna o emprego e a operação do SON mais adequados. As medidas de tráfego devem ser feitas nos *gateways* de serviço. Em adição, mensagens de sinalização devem ser trocadas entre os *gateways* da SON e os roteadores dos domínios que estão sendo sobrepostos para declarar e processar alocação de banda.

6.6 Considerações Finais

Este capítulo inova ao propor uma equação para a banda efetiva do MMW. Esta equação relaciona a banda efetiva do fluxo de tráfego com os parâmetros deste modelo. Os resultados experimentais mostraram que a estimativa de banda efetiva para o modelo MMW é bem mais realista do que a do caso monofractal de Norros, não só garantindo a probabilidade de perda exigida para processos multifractais, mas também atingindo mais alta utilização do enlace.

Muitos problemas de filas em redes envolvem distribuições de probabilidade que são de cauda pesada. A estimação de banda efetiva pode ser difícil para processos com tais distribuições por não possuírem uma forma fechada para a função geradora de momentos e por nem mesmo todos os seus momentos existirem. Esta tese soluciona estes problemas ao se desenvolver uma expressão para a estimação da função geradora de momentos em termos dos multiplicadores do MMW, o que garante que ela exista e seja finita. Quanto à existência dos momentos, esta é garantida pela estrutura do modelo MMW conforme descrita no Capítulo 3, que captura momentos estatísticos de ordem q do tráfego. Outras vantagens da banda efetiva proposta além desta, podem ser citadas, como por exemplo, sua capacidade de atualização adaptativa e obtenção de um valor estimado de banda efetiva mesmo na ausência de dados, uma vez estabelecidos os parâmetros do modelo.

Por ser vista como uma solução promissora em vários contextos de redes atuais, desenvolveu-se também um algoritmo para o cálculo adaptativo de banda efetiva. Com base neste algoritmo, incorporou-se a capacidade de atualização em tempo real do MMW em um esquema adaptativo de provisão de banda. Resultados experimentais validaram o esquema de provisão de banda efetiva proposto, mostrando que este apresentou maior Fator de Adequação Médio (*Average Goodness Factor* - AGF) do que o esquema estático, além de garantir a probabilidade de perda exigida.

Capítulo 7

Análise de Fila para Tráfego Multifractal: Probabilidade de Perda

7.1 Introdução

Probabilidade de perda e atraso de pacotes são medidas de desempenho fundamentais associadas à qualidade de serviço (QoS) em redes de computadores, como as redes TCP-IP e ATM. Vários estudos têm sido realizados com o intuito de caracterizar o tamanho médio da fila (*backlog*) e a distribuição do número de pacotes no *buffer*. Para que, dessa forma, se consiga estabelecer limitantes para essas medidas de desempenho, tais como perda e atraso. O conhecimento destes limitantes permite garantir a qualidade de serviço requerida pelos fluxos de tráfego.

R. L. Cruz obteve limitantes determinísticos para medidas de desempenho de rede utilizando o conceito de processo envelope linear (PEL), oriundo do Cálculo de Rede (Cruz, 1991b)(Cruz, 1991a). Enquanto, C. S. Chang derivou limitantes de desempenho tanto determinísticos quanto estatísticos usando o conceito de mínimo PEL e relacionando o mesmo com a banda efetiva do tráfego (Chang, 1994). Com base no trabalho de Chang, L. Dai obteve limitantes de desempenho mais precisos (Dai, 1997). Nestes dois últimos trabalhos citados, os limitantes de desempenho podem ser calculados através das expressões analíticas existentes para a banda efetiva de modelos de tráfego, ao invés de somente por meio do cálculo do processo envelope. A vantagem dessa incorporação é que estimativas de parâmetros de desempenho como perda e atraso podem ser realizadas de forma analítica ao se assumir um modelo para o tráfego. Outros trabalhos sobre limitantes de desempenho de redes surgiram, principalmente usando Cálculo de Rede, mas sem o uso de banda efetiva (Boudec & Thiran, 2001)(Liebeherr et al., 2003)(Boorstyn et al., 2000). Devem ser também mencionados os trabalhos de Vieira et.al (Vieira & Ling, 2006a) (Vieira & Ling, 2006c) por relacionar a união do cálculo de rede e banda efetiva com a modelagem multifractal.

No Capítulo 3, derivou-se um fator de escala global para tráfego multifractal, que é equivalente ao parâmetro de Hurst. Será mostrado neste capítulo que, através deste parâmetro, os pontos de operação s e t podem ser calculados. Como contribuição original, inicialmente se propõe uma equação para probabilidade de perda baseada na teoria de muitas fontes e em propriedades do MMW, envolvendo o parâmetro de escala global no cálculo do ponto de operação. Como extensão a esta análise, são apresentados limitantes para o desempenho de fila derivados a partir dos conceitos do cálculo de rede (Chang, 1994) em conexão à banda efetiva para o modelo MMW proposta no Capítulo 6. Os objetivos adotados foram o de se obter limitantes para probabilidade de perda e tamanho médio da fila e mostrar que estes limitantes são mais precisos do que a teoria dos grandes desvios prega para tráfego monofractal.

O foco do presente capítulo é a análise de fila de um servidor alimentado por tráfego multifractal, principalmente sob a luz do conceito de probabilidade de perda. O capítulo está organizado da seguinte forma: na seção 7.2, discute-se várias abordagens para a estimação da probabilidade de perda. A seção 7.3 propõe uma equação de probabilidade de perda que leva em conta as propriedades do MMW e já apresenta os resultados obtidos nas simulações. A seção 7.4 aborda como a banda efetiva do MMW se relaciona com o cálculo de rede estatístico a fim de se derivar limitantes para a probabilidade de perda e para o tamanho de fila. São apresentados os testes realizados para validar a proposta de cálculo de limitantes de desempenho de fila. Finalmente, a seção 7.7 enfatiza as conclusões obtidas.

7.2 Probabilidade de Perda

Nesta seção são apresentados alguns pontos básicos da teoria de fila que servem como introdução às seções subseqüentes. Sejam $A(t)$ o processo de chegada de pacotes acumulados em um servidor no intervalo de tempo contínuo $[0, t)$ com incrementos estacionários, C a capacidade do servidor e $I(t)$, o processo de chegada de dados à entrada da rede para chegadas no intervalo de tempo $[0, t)$, dado por

$$I(t) = A(t) - Ct. \tag{7.1}$$

O processo correspondente ao tamanho de fila $W(t)$ em tempo contínuo é representado pelas seguintes equações:

$$W(0) = 0 \qquad e \qquad W(t) = \sup_{s \leq t} \{I(t) - I(s)\} \tag{7.2}$$

Para que a fila seja estável é preciso que $E[A(t) - Ct] < 0$. Dessa forma, a distribuição em regime permanente do processo de tamanho de fila Q pode ser expressa como uma generalização da equação

de Lindley da seguinte forma (Rolls et al., 2005):

$$Q \stackrel{d}{=} \sup_{t \geq 0} \{A(t) - Ct\}, \tag{7.3}$$

onde $\stackrel{d}{=}$ denota igualdade em distribuição.

As próximas seções tratam de alguns resultados de estimação de probabilidade de perda citados na literatura baseados em diferentes teorias e em modelos de tráfego que assumem distribuição com variância finita. Outros resultados de fila para modelos com variância infinita também foram reportados na literatura, porém não serão abordados nesta tese (Karasaridis & Hatzinakos, 2001).

7.2.1 Probabilidade de Perda pela Teoria dos Grandes Desvios

A análise de fila baseada na Teoria dos Grandes Desvios revela que para uma gama de tipos diferentes de tráfego de entrada, a probabilidade de perda decai exponencialmente com o aumento do tamanho do *buffer* (Chang, 1994) (Duffield, 1994). A equação abaixo resume este resultado, já apresentado no Capítulo 6, onde a probabilidade de perda em regime de *buffer* grande b é caracterizada por 2 parâmetros, a constante assintótica β e a taxa de decaimento assintótico η:

$$P(q(t) \geq b) \leq \beta e^{-b\eta}, \tag{7.4}$$

onde $q(t)$ é o número de *bytes* na fila no instante de tempo t.

Nesta seção, apresenta-se alguns resultados alternativos à equação (7.4) obtidos através da Teoria dos Grandes Desvios para a estimação da probabilidade de perda.

Seja $a(t)$, o número de *bytes* que chegam a um servidor no instante de tempo discreto t. Supõe-se que o tamanho do *buffer* seja infinito e a capacidade do servidor igual a c. Considere também uma disciplina de serviço conservativa, ou seja, o servidor torna-se inativo apenas na ausência de serviço. Assim, a fila é regida pela versão discreta da equação recursiva de Lindley (equação (7.3)):

$$q(t+1) = \max(q(t) + a(t+1) - c, 0). \tag{7.5}$$

Ao se considerar o processo de chegada como sendo um fluido contínuo com taxa *a(t)*, a equação a tempo discreto acima pode ser reescrita da seguinte forma (Chang, 1994) (Chang & Thomas, 1995):

$$\dot{q}(t) = \begin{cases} a(t) - c & if \quad q(t) > 0 \\ \max(a(t) - c, 0) & if \quad q(t) = 0 \end{cases} \tag{7.6}$$

O Teorema de Gartner-Ellis apresentado na seção 6.2.3 do Capítulo 6, pode ser usado para caracterizar a função densidade de probabilidade $f(\alpha, t)$ do processo de taxa variável $a(t)$, de tal modo

que este processo possa ser visto como uma fonte de taxa constante α no instante de tempo *t*. A partir disso, a distribuição de cauda do tamanho da fila pode ser representado pela integral de $f(\alpha, t)$ em um certo intervalo. Com a ajuda da Teoria dos Grandes Desvios, demonstra-se que $f(\alpha, t)$ é uma distribuição de Gibb, ou seja, tem a forma $\exp(-t\Lambda^*(\alpha))$, onde $\Lambda^*(\alpha)$ é a transformada de Legendre da função de energia $\Lambda(\theta)$ (Bucklew, 1990). A função $\Lambda(\theta)$ pode ser derivada através do teorema limite de Gartner-Ellis também apresentado no Capítulo 6, ou seja (Ellis, 1984):

$$\Lambda(\theta) = \lim_{t\to\infty} \frac{1}{t} \log \mathrm{E}\mathrm{e}^{\theta \mathrm{A(0,t)}}, \tag{7.7}$$

para todo $\theta \in \mathbb{R}$.

Segundo as afirmações acima, a função $f(\alpha, t)$ pode ser escrita como

$$f(\alpha, t) \cong \exp(-t\Lambda^*(\alpha)), \tag{7.8}$$

onde $\Lambda^*(\alpha)$ é a função de entropia do processo $a(t)$ obtida através da transformada de Legendre de $\Lambda(\theta)$, ou seja

$$\Lambda^*(\alpha) = \sup_{\theta}[\theta\alpha - \Lambda(\theta)]. \tag{7.9}$$

A aproximação $f(\alpha, t) \cong \exp(-t\Lambda^*(\alpha))$ implica dizer que:

$$\lim_{t\to\infty} \frac{\log P(A(0,t) \approx \alpha t)}{t} = \Lambda^*(\theta). \tag{7.10}$$

Tendo em mãos uma estimativa para a função $f(\alpha, t)$, a probabilidade de que haja pelo menos x *bytes* no *buffer* usando a equação (7.6) é dada por (Chang & Thomas, 1995):

$$P(q(t) \geq x) = \int_{(\alpha-c)^+ t \geq x} f(\alpha, t) d\alpha \approx \int_{(\alpha-c)^+ t \geq x} \exp(-t\Lambda^*(\alpha)) d\alpha \tag{7.11}$$

Supondo que em regime permanente se tem $\sup_t P(q(t) \geq x) = P(q(\infty) \geq x)$, então as seguintes equações são válidas (Bucklew, 1990):

$$\sup_t P(q(t) \geq x) \approx \int_\alpha \exp\left(-x \inf_\alpha \frac{\Lambda^*(\alpha)}{(\alpha - c)^+}\right) d\alpha$$

$$P(q(\infty) \geq x) \approx \exp\left(-x \inf_\alpha \frac{\Lambda^*(\alpha)}{(\alpha - c)^+}\right). \tag{7.12}$$

Portanto, para se estimar a probabilidade de perda em regime permanente, como pode ser observado pela equação (7.12), é necessário resolver um problema de minimização envolvendo a função de entropia da fonte de tráfego.

7.2.2 Probabilidade de Perda para Processos com Longa-dependência

Várias questões de engenharia de tráfego, como dimensionamento de *buffer* e controle de fluxo, estão relacionadas ao comportamento de fila do tráfego nas redes. A característica de longa-dependência do tráfego tem um impacto significativo em seu comportamento de fila, principalmente com relação à probabilidade de perda (Grossglauser & Bolot, 1999).

Norros (Norros, 1994) e Duffield e O'Connell (Duffield & O'Connell, 1993a) apresentaram limitantes inferiores para a probabilidade de perda $P(Q > b)$ (probabilidade do tamanho de fila Q exceder o valor do *buffer* b) para processos auto-similares. Entretanto, em muitos casos, esta aproximação subestima o valor real de $P(Q > b)$. O limitante inferior para $P(Q > b)$ decai assintoticamente (para *buffer* muito grande) de acordo com uma função de Weibull (Norros, 1994)(Duffield & O'Connell, 1993a). A distribuição de cauda da ocupação do *buffer* (a densidade de probabilidade do tamanho da fila para b grande) é mais 'pesada', ou seja, possui decaimento mais lento do que a distribuição exponencial predita por modelos de tráfego tradicionais de curta-dependência. A distribuição do tamanho da fila ou a probabilidade de perda para processos que têm parâmetro de Hurst $H \in (0.5, 1)$ pode ser dada, segundo (Duffield & O'Connell, 1993a), por

$$\lim_{b\to\infty} b^{-2(1-H)} \ln P(Q > b) = -a^{-2(1-H)}(a + c)^2/2, \tag{7.13}$$

onde $a = c/H - c$.

A fórmula acima pode ser usada para estimar $P(Q > b)$ da seguinte forma:

$$P(Q > b) \approx e^{-\gamma b^{2(1-H)}}, \tag{7.14}$$

onde $\gamma = a^{-2(1-H)(a+C)^2/2}$. Esta aproximação é precisa para *buffer b* muito grande, mas pode não valer para outros valores de *b*, o que de fato resulta em uma subestimação de $P(Q > b)$ (Kim & Shroff, 2001).

Pode-se demonstrar que a distribuição de probabilidade de transbordo do *buffer* (o mesmo que probabilidade de cauda) para tráfego com curta-dependência como processos Markovianos e de Poisson apresentam um decaimento exponencial relativo ao tamanho do *buffer*; em contraste, tem-se uma distribuição de Weibull para processos com longa-dependência como o processo fBm (Movimento Browniano fracionário) e um decaimento hiperbólico para processos alfa-estáveis (Karasaridis & Hatzinakos, 2001). As discrepâncias nas distribuições de probabilidade de perda se tornam ainda mais acentuadas para tráfego não-gaussiano (Norros, 1994). Vários autores afirmam que o tempo entre chegadas de pacotes age como um fenômeno multiplicativo, ao invés de aditivo. Isto implica que a distribuição dos tempos de chegada para um fluxo de tráfego tende a ser lognormal, que é uma distribuição de cauda pesada (Riedi et al., 1999). Estas constatações tanto para a distribuição dos

dados de tráfego como para a probabilidade de perda incentivam a combinação de métodos para a estimação da probabilidade de perda. Assim, este estudo faz a união da teoria de regime limite de muitas fontes com as propriedades do MMW proposto, obtendo uma formulação mais precisa para a probabilidade de perda, que leva em consideração características multifractais do tráfego.

7.2.3 Probabilidade de Perda Assintótica pela Teoria das Muitas Fontes

Em geral, a quantidade disponível de recursos da rede depende não apenas das propriedades estatísticas e dos requisitos de qualidade de serviço de uma fonte, mas também das propriedades estatísticas de outros fluxos de tráfego que estão sendo multiplexados em conjunto e das características do enlace (capacidade do enlace e tamanho do *buffer*). Em (Kelly, 1996)(Courcoubetis et al., 1997), é mostrado que a banda efetiva de uma fonte depende do ponto de operação ('operating point') do enlace através dos parâmetros de espaço s e de tempo t, que dependem dos recursos do enlace e das propriedades estatísticas do tráfego multiplexado. A banda efetiva pode ser estimada usando a teoria assintótica de muitas fontes ao invés de se considerar o limite assintótico de *buffers* grandes, obtendo assim uma estimativa menos conservadora (banda menor mas que atende à probabilidade de perda estipulada) para *buffers* menores (Choudhury et al., 1994). Este resultado ocorre porque a teoria assintótica de *buffer* grande não leva em consideração o possível ganho de multiplexação quando fontes de tráfego independentes são multiplexadas. Na teoria assintótica limite de muitas fontes, estuda-se o decaimento da probabilidade de perda $P(Q > b)$ com o aumento do número de entradas independentes para cada roteador, enquanto o tamanho do *buffer* por entrada e a taxa de serviço por entrada permanecem fixos. Nesta seção, é derivado o ponto de operação do enlace considerando o MMW e a teoria assintótica de muitas fontes.

Seja $\alpha_n(s,t)$ uma estimativa da banda efetiva $\alpha(s,t)$ para uma série de tráfego de tamanho $T = nt$. Deseja-se determinar o ponto de operação, ou seja, os valores de tempo t e do parâmetro de espaço s com os quais a banda efetiva se relaciona com a probabilidade de transbordo assintótica (Courcoubetis et al., 1999). Esta banda efetiva é relacionada com a probabilidade de transbordo do *buffer* $\Psi \approx \log P(Q_n > b)$ sob o regime assintótico de 'muitas fontes' através da fórmula (Courcoubetis & Weber, 1996):

$$\Psi = \inf_{t \geq 0} \sup_{s \geq 0}((b + ct)s - N_s st\alpha(s,t)), \tag{7.15}$$

onde *c* é a capacidade do enlace, *b* é o tamanho do *buffer* e N_s é o número de fontes de entrada com banda efetiva igual a $\alpha(s,t)$. O par de variáveis (s^*, t^*) com o qual a otimização *inf sup* na equação (7.15) é atendida, corresponde ao 'ponto de operação' do enlace.

Seja a função $\Lambda_n(s,t)$ uma estimativa de $\Lambda(s,t) = E(e^{sX})$, a função geradora de momento de um processo $X(t)$. Pode-se afirmar que o par de variáveis (s_n^*, t_n^*) é um estimador consistente de (s^*, t^*),

dado pelas seguintes equações (Aspirot et al., 2005):

$$b + ct_n^* = \frac{(\partial/\partial s)\Lambda_n(s_n^*, t_n^*)}{\Lambda_n(s_n^*, t_n^*)} \tag{7.16}$$

$$cs_n^* = \frac{(\partial/\partial t)\Lambda_n(s_n^*, t_n^*)}{\Lambda_n(s_n^*, t_n^*)} \tag{7.17}$$

Quando um modelo de tráfego é adotado para o tráfego de chegada, uma aproximação paramétrica pode ser empregada para o cálculo da função geradora de momento. Faremos uma aproximação para o cálculo da função geradora de momento de forma a obter uma expressão analítica para o 'ponto de operação'. Sabe-se que com a agregação de várias fontes, o tráfego no enlace tende a ser gaussiano (Zhang et al., 2003). Como a análise apresentada aqui envolve muitas fontes, assumiremos que $\Lambda_n(s,t)$ é a função geradora de momento de um processo fBm, sendo este gaussiano. Assim, tem-se o seguinte para função geradora de momento (Norros, 1994):

$$\Lambda_n(s,t) = \exp\left(\mu ts + \frac{t^{2H}\sigma^2 s^2}{2}\right), \tag{7.18}$$

onde μ é a média, σ^2 é a variância do processo em questão. Neste capítulo, o parâmetro de Hurst será sempre calculado utilizando-se o parâmetro de escala global proposto H_g aplicado às séries de tráfego consideradas. Resolvendo o sistema de equações (7.16) e (7.17), encontram-se as seguintes equações para o 'ponto de operação' do enlace:

$$t^* = \frac{H_g b}{(c-\mu)(1-H_g)} \tag{7.19}$$

e

$$s^* = \frac{b(t^*)^{-2H_g}}{\sigma^2(1-H_g)} \tag{7.20}$$

Dessa forma, uma expressão analítica para a estimação da probabilidade de perda de J fluxos de tráfego em um servidor pode ser obtida utilizando a equação (7.15):

$$\log P(Q_n > b) \approx ((b + ct^*)s^* - s^*t^* \sum_{j=1}^{J} \alpha_j(s,t)), \tag{7.21}$$

onde o ponto de operação (s^*, t^*) é dado pelas equações (7.19) e (7.20).

Uma vez que H_g está relacionado com o parâmetro α da função $\tau_0(q) = \log_2 \frac{\Gamma(\alpha)\Gamma(2\alpha+q)}{\Gamma(2\alpha)\Gamma(\alpha+q)}$, pela expressão $H_g = 1 - \frac{\log_2(\frac{\alpha+1}{\alpha+1/2})}{2}$, a equação (7.21) provê uma conexão entre a probabilidade de perda e a função de escala $\tau(q)$. Note que é demonstrado na literatura que a aproximação através da análise

assintótica de 'muitas fontes' pode ser aplicada igualmente a qualquer modelo de tráfego tradicional (por exemplo, modelos Markovianos) e modelos com longa dependência (Courcoubetis et al., 1999).

7.2.4 Probabilidade de Perda Assintótica para Tráfego Multifractal

Poucos trabalhos tratam de problemas de fila cujo tráfego de entrada tem um comportamento em escala mais complexo, como é o caso de enlaces com tráfego de entrada multifractal (Gao & Rubin, 1999a)(Molnár et al., 2002). Para muitos modelos de tráfego, a análise assintótica é a única forma disponível. A aproximação assintótica geralmente fornece uma visão bastante simplificada do tráfego de redes e geralmente não se aplica a *buffer* finito. Para o caso de tráfego multifractal de entrada, as análises de fila tradicionais e clássicas como M/M/1, M/G/1 e G/G/1 não ajudam muito. Neste caso, os comportamentos e leis de potência de diferentes ordens das estatísticas dos fluxos de tráfego em várias escalas são aspectos mais importantes e relevantes.

Dentre os estudos de estimação assintótica de probabilidade de perda para tráfego multifractal, Gao et al. simularam filas com processos multiplicativos multifractais de entrada, mas não apresentando resultados analíticos (Gao & Rubin, 1999a)(Gao & Rubin, 2000). Enquanto, S. Molnár et al. propuseram uma expressão analítica aproximada para o comportamento assintótico de cauda de fila, ou seja, eles derivaram a probabilidade de cauda de perda em um servidor tendo como entrada um processo multifractal. Em particular, quando o tráfego de entrada é monofractal, esta probabilidade de perda assintótica tem um decaimento de Weibull, o que é consistente com outros resultados (Norros, 1994)(Molnár et al., 2002). Explicitamente, para um servidor único com tráfego multifractal de entrada e para *buffer* b grande, a probabilidade de perda assintótica pode ser estimada como (Molnár et al., 2002):

$$\log P(Q > b) \approx \min_{q>0} \log\{c(q) \frac{[\frac{b\tau_0(q)}{s(q-\tau_0(q))}]^{\tau_0(q)}}{[\frac{bq}{(q-\tau_0(q))}]^{q}}\}, \tag{7.22}$$

onde o fator de momento $c(q)$ e a função de escala $\tau_0(q) = \tau(q) + 1$ são funções que definem o processo multifractal de entrada.

7.3 Probabilidade de Perda para o Modelo Multifractal Baseado em Wavelets

A cauda da distribuição do tamanho da fila, ou seja, a estimação assintótica da probabilidade de perda em um sistema de fila com *buffer* infinito, tem sido bastante estudada (Chang, 1994)(Glynn & Whitt, 1994)(Addie & Zukerman, 1994); enquanto, há um número reduzido de trabalhos que tratam diretamente da probabilidade de perda em sistemas com *buffer* finito (Likhanov & Mazumdar,

1998)(Baiocchi et al., 1991)(Shroff & Schwartz, 1998). Na literatura, a taxa de perda de pacote (probabilidade de perda) $P_l(x)$ é freqüentemente aproximada pela probabilidade de cauda (probabilidade de transbordo) $P(Q > x)$, que de fato, fornece um limitante superior para a probabilidade de perda, mas muitas vezes sem muita precisão (Krunz & Ramasamy, 2000)(Gyorgy & Borsos, 2001). Estudos realizados comprovam que tanto para vários modelos de tráfego, como para séries de tráfego reais, esta estimativa é realmente um limitante superior bastante aproximado (Duffield et al., 1995)(Kelly, 1991).

Seja um servidor com taxa constante *c*, um fluído de entrada λ_n, e os processos Q_n e $\hat{Q}_n$ denotando o tamanho da fila para *buffer* finito e infinito no instante de tempo *n*, respectivamente. Supõe-se que λ_n seja estacionário e ergódico e que $E(\lambda_n) < c$. No cálculo de probabilidade de perda, considera-se que Q_n e $\hat{Q}_n$ sejam estacionários e ergódicos. A probabilidade de perda $P_l(x)$ para um *buffer* de tamanho x é definida como a razão entre a quantia de fluido de tráfego perdido pela quantia de fluido de entrada como (Kim & Shroff, 2001):

$$P_l(x) = \lim_{N\to\infty} \frac{\sum_{k=1}^{N} max(Q_{k-1} + \lambda_k - c - x, 0)}{\sum_{k=1}^{N} \lambda_k}. \tag{7.23}$$

Já a probabilidade de cauda é definida como o tempo gasto pelo fluido de tráfego no *buffer* de tamanho infinito, acima de um nível x, dividido pelo tempo total de observação e pode ser expressa como (Kim & Shroff, 2001):

$$P(Q > x) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} I(\hat{Q}_k > x), \tag{7.24}$$

onde se A é uma hipótese verdadeira, $I(A) = 1$; do contrário, $I(A) = 0$.

Para processos de chegada de tráfego sem memória em um sistema de fila, Kelly (Kelly, 1996) obteve uma expressão assintótica para a probabilidade de perda $P_l(x)$ para *buffer* grande, e em (Likhanov & Mazumdar, 1998), Likhanov e Mazumdar obtiveram uma fórmula assintótica para a probabilidade de perda através da análise assintótica de muitas fontes. Este último resultado é válido para a maioria dos processos de tráfego e é mais preciso e robusto do que os obtidos através da Teoria de Grandes Desvios.

Com relação a tráfego multifractal de entrada, podemos citar o trabalho de Ribeiro et al. (Ribeiro et al., 2000), que desenvolveram uma análise de fila multiescala para modelos multifractais baseados em cascata via um método não-assintótico, válida para qualquer tamanho de *buffer*. Esta aproximação nomeada de Análise de Fila Multiescala (*Multiscale Queueing*) incorpora as distribuições dos dados de tráfego em múltiplas resoluções temporais (não apenas as estatísticas de segunda ordem) (Ribeiro et al., 2000). Considere o processo aleatório discreto L_i representando a carga (volume) de tráfego

que entra um servidor com *buffer* infinito e capacidade de serviço constante c. Supondo também que Q_i represente o tamanho da fila no instante de tempo i. Denotemos por K_r, o tráfego agregado que chega entre os instantes 0 a r e 0, ou seja

$$K_r = \sum_{i=0}^{r} L_i. \tag{7.25}$$

O processo K_r refere-se aos dados de tráfego na escala de tempo r. Modelos baseados em cascata provêem fórmulas explícitas e simples para K_r em escalas de tempo diádicas, ou seja, $r = 2^n$ ($n = 1, 2, \ldots, \infty$). A Análise de Fila Multiescala mostra que a probabilidade de perda $P_l(x)$ pode ser estimada pela seguinte equação (Ribeiro et al., 2000):

$$P[Q > b] \approx 1 - \prod_{i=0}^{n} P[K_{2^{n-i}} < b + c2^{n-i}] \tag{7.26}$$

O modelo multifractal baseado em *wavelets* proposto (MMW) considera momentos de todas as ordens e leva em conta a distribuição do tráfego, portanto o comportamento do tráfego é completamente especificado por este modelo. Esta seção faz uso de propriedades do MMW (média, variância, parâmetro de escala global) na derivação de uma expressão para a probabilidade de perda para tráfego de entrada em um servidor. Pode-se estimar a taxa de perda de *bytes* (probabilidade de perda de *bytes*) $P_l(x)$ em um *buffer* finito através da cauda da distribuição do tamanho de fila (probabilidade de cauda ou probabilidade de transbordo) $P(Q > X)$ para processos MMW de entrada. Devido à estrutura multiplicativa dos processos multifractais, a distribuição não-gaussiana destes processos é aproximadamente do tipo lognormal. Este resultado pode ser constatado na prática principalmente para tráfego real em pequenas escalas de tempo (Erramilli et al., 1996) (Feldmann et al., 1998). Aproveitando este fato, a análise de fila pode ser simplificada, gerando o seguinte teorema.

Teorema 7.3.1 *Seja um processo $X(t)$ descrito pelo modelo MMW caracterizado pelo conjunto de parâmetros (α, γ, ρ) e parâmetro de escala global H_g. A probabilidade de perda em um servidor com capacidade c e* buffer *finito b para o processo $X(t)$ pode ser expressa como*

$$P_l(x) = e^{(xs^*)-(\rho+\gamma^2/2)} \int_c^{\infty} (r-c) \frac{1}{r\theta\sqrt{2\pi}} e^{\frac{-(ln(r)-\varpi)^2}{2\theta^2}} dr, \tag{7.27}$$

onde

$$s^* = \frac{b(t^*)^{-2H_g}}{\sigma^2(1-H_g)} \tag{7.28}$$

e

$$t^* = \frac{H_g b}{(c-\mu)(1-H_g)}. \tag{7.29}$$

Demonstração Tendo estimativas da probabilidade assintótica de cauda $P(Q > x)$ e da probabilidade de perda $P_l(a)$ para um *buffer* de tamanho a qualquer, a probabilidade de perda $P_l(x)$ pode ser calculada por (Kim & Shroff, 2001):

$$P_l(x) = \frac{P_l(a)}{P(Q > a)} P(Q > x). \tag{7.30}$$

Convém utilizar a equação (7.30) para $a = 0$ porque a probabilidade de perda $P_l(0)$ é mais fácil de ser calculada. Para o cálculo de $P(Q > x)$ e $P(Q > 0)$, pode-se usar a equação (7.21) que é um resultado assintótico derivado da teoria de muitas fontes.

Seja λ_n a taxa do tráfego de entrada, assim a probabilidade de perda $P_l(x = 0)$ pode ser calculada como (Kim & Shroff, 2001):

$$P_l(0) = \frac{E\{max\{(\lambda_n - c), 0\}\}}{E\{\lambda_n\}}. \tag{7.31}$$

Supondo agora que o processo λ_n seja um processo MMW com média μ e variância σ^2 dadas pelas equações (3.8) e (3.9), respectivamente. Sabe-se que o MMW, sendo um modelo baseado em cascata multiplicativa, possui distribuição de probabilidade lognormal conforme discutido no Capítulo 3. Dessa forma, a equação (7.31) pode ser reescrita como

$$P_l(0) = \frac{1}{e^{\rho+\gamma^2/2}} \int_c^{\infty} (r - c) \frac{1}{r\theta\sqrt{2\pi}} e^{\frac{-(ln(r)-\varpi)^2}{2\theta^2}} dr, \tag{7.32}$$

onde os parâmetros ϖ e θ são relacionados com a média μ e a variância σ^2 do MMW pelas seguintes equações:

$$\varpi = \ln \mu - \frac{1}{2} \ln \left(\frac{\sigma^2}{\mu^2} + 1 \right), \tag{7.33}$$

$$\theta = \sqrt{\ln \left(\frac{\sigma^2}{\mu^2} + 1 \right)}. \tag{7.34}$$

A teoria assintótica de muitas fontes provê o resultado necessário para os cálculos de $P(Q > x)$ e $P(Q > 0)$ (equação (7.21)) e a partir da aplicação da equação (7.32) em (7.30), obtém-se a probabilidade de perda enunciada por este teorema.

$$P_l(x) = e^{(xs^*)-(\rho+\gamma^2/2)} \int_c^{\infty} (r - c) \frac{1}{r\theta\sqrt{2\pi}} e^{\frac{-(ln(r)-\varpi)^2}{2\theta^2}} dr. \tag{7.35}$$

Note que s^* é um dos parâmetros do ponto de operação (s^*, t^*) apresentado na seção 7.2.3, que é calculado utilizado o parâmetro de escala global H_g.

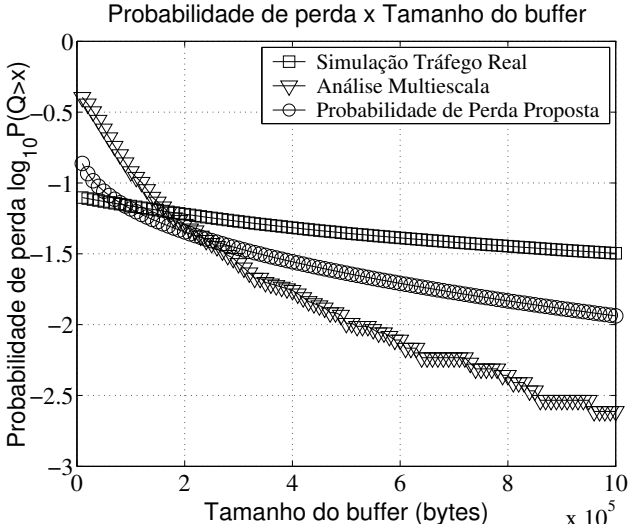


Fig. 7.1: Probabilidade de perda para a série de tráfego dec-pkt-2

■

Para validar a equação proposta para a probabilidade de perda de bytes em um enlace, simulações com vários traços de tráfego foram realizadas. A Figura 7.1 compara as probabilidades de perda em termos do tamanho do *buffer* obtidas com a equação (7.35), pela Análise de Fila Multiescala (Ribeiro et al., 2000) e através da simulação de um enlace com um servidor alimentado pela série de tráfego dec-pkt-2 na escala de tempo de 512ms. A capacidade do servidor foi ajustada para 120 porcento da taxa média. O mesmo é apresentado para a série de tráfego dec-pkt-3 na Figura 7.2. A análise das Figuras 7.1 e 7.2 revela que com o conhecimento do conjunto de parâmetros (α, ρ, γ), estimativas mais precisas para a probabilidade de perda podem ser realizadas para processos de tráfego de redes. Ou seja, a expressão para a probabilidade de perda proposta descreve melhor o comportamento de fila no *buffer*. Note que, assim como a equação de probabilidade de perda proposta, a Análise de Fila Multiescala estima a probabilidade de perda $P_l(x)$ e não a probabilidade de cauda; este é o principal motivo de sua escolha para comparação.

Prosseguindo a análise de probabilidade de perda, com o intuito de verificar o decaimento da probabilidade de perda em termos da capacidade do servidor, fixou-se o tamanho do *buffer* em $1.4\text{x}10^5$. Este tamanho do *buffer* foi escolhido pois com este valor os dois métodos baseados em processo multifractais e a simulação apresentaram desempenho próximos em relação à perda para a série de tráfego dec-pkt-3, como pode ser visto pela Figura 7.2. A Figura 7.3 mostra o comportamento do nosso método de cálculo de probabilidade de perda comparado à Análise de Fila Multiescala e a real probabilidade de perda estimada através de simulação com a série de tráfego dec-pkt-3, em função da capacidade do servidor. Observa-se que a equação (7.35) provê resultados mais realistas, apresentando uma taxa de decaimento mais próxima da probabilidade de perda real do que o método de fila

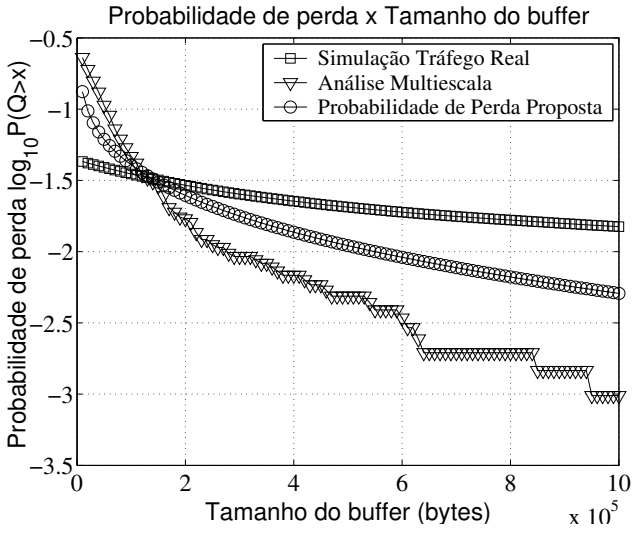


Fig. 7.2: Probabilidade de perda para a série de tráfego dec-pkt-3

multiescala.

7.4 Limitantes de Desempenho utilizando Cálculo de Rede e Banda Efetiva

As pesquisas sobre limitantes de desempenho de redes têm aberto novos rumos à análise e dimensionamento de redes de alta velocidade (Kurose, 1992). Limitantes de desempenho para atraso e tamanho de fila (*backlog*) podem ser obtidos em termos da banda efetiva dos fluxos de tráfego em conjunto com o Cálculo de Rede (Chang, 1994). O Cálculo de Rede consiste de uma série de resultados matemáticos fundamentados na álgebra Min-Plus capaz de fornecer soluções para vários problemas de fluxos de tráfego encontrados nas redes (Boudec & Thiran, 2001). Inicialmente o Cálculo de Rede foi introduzido em sua versão determinística, sendo capaz de prover limitantes superiores para o retardo e o tamanho da fila (Cruz, 1991b). Uma vez que os resultados de pior caso do Cálculo de Rede Determinístico geralmente superestimavam os recursos da rede, vários pesquisadores saíram em busca de estender o Cálculo de Rede a uma formulação probabilística. Estas extensões probabilísticas ao Cálculo de Rede são comumente denominadas de Cálculo de Rede Estatístico. Nesta seção serão explorados alguns conceitos do Cálculo de Rede para a obtenção de limitantes de desempenho em função da banda efetiva derivada para o MMW. Antes de descrevermos o Cálculo de Rede usando a formulação estatística descrita em (Chang, 1994), serão mencionados alguns elementos do Cálculo de Rede Determinístico (Boudec & Thiran, 2001).

Um dos conceitos envolvidos com o Cálculo de Rede é o de processo envelope. O Cálculo de Rede

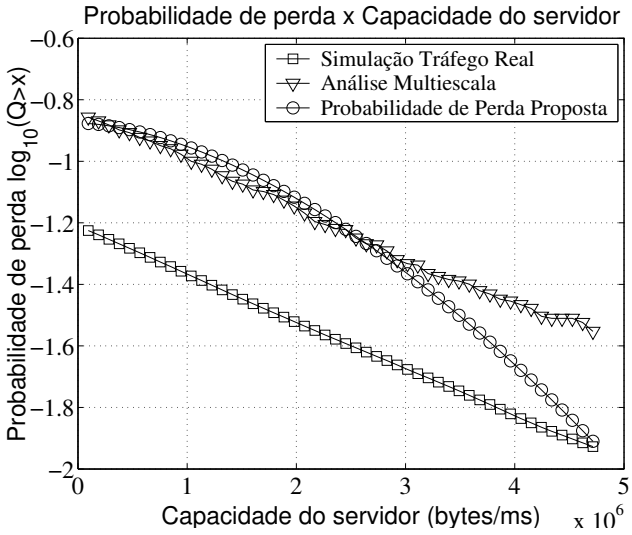


Fig. 7.3: Probabilidade de perda x Capacidade do servidor para a série de tráfego dec-pkt-3

usa processos envelopes para descrever o tráfego de chegadas e os serviços em uma rede. Considere uma seqüência aleatória não-negativa $\{a(t), t = 0, 1, 2, \ldots\}$ correspondente ao processo de chegada de tráfego (por exemplo, quantidade de *bytes*) e seja $A(t_1, t_2) = \sum_{t=t_1}^{t_2-1} a(t)$. Um processo $\hat{A}(t)$ é dito ser um processo envelope de $a(t)$ se:

$$A(t_1, t_2) \leq \hat{A}(t_2 - t_1) \qquad \forall \quad t_1 \leq t_2 \tag{7.36}$$

Note que $\hat{A}(t)$ é 'estacionário' no sentido de que depende apenas da diferença entre os instantes de tempo t_1 e t_2. A noção de processo envelope é similar ao de estacionaridade pois o processo envelope limita o processo original mesmo sob um deslocamento de tempo arbitrário (Cruz, 1991b) (Boudec & Thiran, 2001).

Um processo envelope $\hat{A}(t)$ é subaditivo se $\hat{A}(t_1, t_2) \leq \hat{A}(t_1) + \hat{A}(t_2)$ para todo t_1 e t_2. Supondo que $\hat{A}(t)$ seja um processo crescente e subaditivo, então tem-se que (Chang, 1994):

$$\lim_{t\to\infty} \frac{\hat{A}(t)}{t} = \inf_{t\geq 1} \frac{\hat{A}(t)}{t} \stackrel{d}{=} \hat{a}, \tag{7.37}$$

onde $\hat{a}$ é a taxa do processo envelope $\hat{A}(t)$.

Como o processo envelope para o processo $a(t)$ não é único, é natural perguntar se há um processo envelope $A^*(t)$ que satisfaz $A^*(t) \leq \hat{A}(t)$ para qualquer t e para todo processo envelope $\hat{A}(t)$. O processo envelope $A^*(t)$ é denominado de PEM (Processo Envelope Mínimo) e é dado por

$$A^*(t) = \sup_{s\geq 0} A(s, s + t). \tag{7.38}$$

O processo envelope mínimo (PEM) $A^*(t)$ é crescente e subaditivo e sua média é denominada de taxa de envoltória mínima (TEM). O PEM representado pela equação (7.38) pode ser alternativamente obtido através dos processos envelopes lineares prospostos por Cruz (Cruz, 1991b)(Cruz, 1991a): $A^*(t) \leq \hat{a}t + \hat{\sigma}$ para algum valor de $\hat{\sigma}$ constante e não-negativo. Através do processo envelope mínimo A^* pode-se obter a taxa de envelope mínima (TEM) a^* de $a(t)$ que é

$$\lim_{t\to\infty} \frac{A^*(t)}{t} = a^*. \tag{7.39}$$

Como exemplo de aplicação da taxa de envelope mínima (TEM), sabe-se que considerando um sistema com um servidor com capacidade c, *buffer* infinito e uma disciplina de serviço conservativa, se $a^* < c$, existe uma constante $d < \infty$, tal que o retardo máximo não é maior do que d (Dai, 1997).

Agora estenderemos os resultados apresentados nesta seção a um contexto estatístico. Ao invés de limitantes determinísticos para variáveis aleatórias, serão considerados limitantes para funções geradoras de momento. O desenvolvimento a seguir explorará o conceito de processo envelope com relação a um parâmetro θ. Pode-se dizer que uma variável aleatória X é limitada exponencialmente em relação a um parâmetro θ ($0 < \theta < \infty$) se existe uma constante $g < \infty$, tal que:

$$(Ee^{\theta}X)^{\frac{1}{\theta}} \leq g \tag{7.40}$$

O limitante de Chernoff afirma que (Papoulis, 1991):

$$P(X \geq x) \leq g^{\theta}e^{-\theta x} \quad \forall x, \tag{7.41}$$

que provê um limitante para a distribuição de cauda de X, ou seja, para $P(X \geq x)$.

De forma análoga ao processo envelope dado pela equação (7.36), consideremos $\hat{A}(\theta, t)$ um processo 'limitante' de $a(t)$ tal que (Chang, 1994):

$$\frac{1}{\theta}\log Ee^{\theta A(t_1,t_2)} \leq \hat{A}(\theta, t_2 - t_1) \qquad \forall \quad t_1 \leq t_2 \tag{7.42}$$

O processo $\hat{A}(\theta, t)$ é denominado processo envelope de $a(t)$ com relação a θ (Chang, 1994). Dessa forma, o processo envelope mínimo (PEM) em relação a θ é

$$A^*(\theta, t) = \sup_{s\geq 0} \frac{1}{\theta}\log Ee^{\theta A(s,s+t)}. \tag{7.43}$$

Ao contrário do PEM na formulação determinística do Cálculo de Rede inicialmente apresentada, este PEM estatístico em geral não é subaditivo. Devido a isso, a taxa mínima de $a(t)$ (TEM) em

relação a θ é definida como

$$a^*(\theta) = \limsup_{t\to\infty} \frac{A^*(\theta,t)}{t}. \tag{7.44}$$

O processo de chegada $a(t)$ é dito limitado por um processo envelope linear (PEL) $\alpha(\theta)t + \sigma(\theta)$ onde $\alpha(\theta) \geq 0$, $\sigma(\theta) \geq 0$ com relação a $\theta \in \mathbb{R}^+$, se

$$\frac{1}{\theta}\log Ee^{\theta A(t_1,t_2)} \leq \alpha(\theta)(t_2 - t_1) + \sigma(\theta), \tag{7.45}$$

para todo $t_2 \geq t_1 \geq 0$.

A equação (7.45) permite que a seguinte interpretação seja feita: $\alpha(\theta)$ é um limitante para a taxa de chegada estacionária, enquanto $\sigma(\theta)$ pode ser interpretado como um limitante do grau de variação de rajada (*burstiness*) presente no processo de chegada.

Deve-se destacar que a definição de taxa de envelope mínima (TEM) está vinculada à Teoria dos Grandes Desvios através do Teorema de Gärtner-Ellis apresentado no Capítulo 6 (Bucklew, 1990). Para estabelecer esta conexão, assume-se as seguintes condições para o processo de chagada $a(t), t \geq 0$: i) $\{a(t), t \geq 0\}$ é estacionário e ergódico; ii) $a^*(\theta) = \lim_{t\to\infty} \frac{A^*(\theta,t)}{t}$ para todo $0 < \theta < \infty$; iii) $\theta a^*(\theta)$ é estritamente convexo e diferenciável para todo $0 < \theta < \infty$. Sob estas condições, a seqüência $\{A(0,t), t \geq 1\}$ obedece ao Princípio dos Grandes Desvios também apresentado no Capítulo 6, com a função de taxa $I(v)$ dada por (Bucklew, 1990)

$$I(v) = \sup_{\theta}\{\theta v - \theta a^*(\theta)\}. \tag{7.46}$$

Para o caso em que o processo $a(t)$ é uma seqüência de variáveis aleatórias i.i.d, a TEM $a^*(\theta)$ é citada como sendo a banda efetiva por Kelly (Kelly, 1996). Analogamente ao caso de limitantes determinísticos, o par $(\alpha^*(\theta), \sigma^*(\theta))$ define um processo envelope linear (PEL) mínimo $\alpha^*(\theta)t+\sigma^*(\theta)$ com relação a θ pelas equações:

$$\alpha^*(\theta) = \lim_{t\to\infty} \sup \frac{1}{t} \sup_{\theta\geq 0} \frac{1}{\theta}\log Ee^{\theta A(t_1,t_2)} \tag{7.47}$$

e

$$\sigma^*(\theta) = \inf\left\{\sigma(\theta)|\frac{1}{\theta}\log Ee^{\theta A(t_1,t_2)} \leq \alpha^*(\theta)(t_2 - t_1) + \sigma(\theta)\right\} \quad \forall t_2 \geq t_1 \geq 0. \tag{7.48}$$

De fato, o PEL mínimo é definido para qualquer tipo de processo $a(t)$ sendo este i.i.d ou não. Caso o seguinte limite exista:

$$h(\theta) = \lim_{t\to\infty} \frac{1}{t}\log Ee^{\theta A(0,t)}, \tag{7.49}$$

então $h(\theta)/\theta$ é a banda efetiva de $a(t)$ com relação a θ. Assim sendo, $\alpha^*(\theta)$ (equação 7.47) representa

a banda efetiva procurada.

Um servidor possui disciplina de serviço conservativa se este se torna inativo apenas na ausência de demanda de serviço. Para um servidor com capacidade c operando sob uma disciplina de serviço conservativa, se a TEM do processo de entrada for menor do que c, então pode-se afirmar que (Chang, 1994): i) O comprimento de fila é limitado exponencialmente com relação a θ; ii) O retardo virtual é limitado exponencialmente com relação a θc se a política de escalonamento for FIFO (First-In First-Out). Usando estes resultados, limitantes para a distribuição de cauda do comprimento de fila podem ser estimados através do processo envelope linear do tráfego de entrada. Seja $\{W(t)\}$ o processo correspondente ao tamanho da fila no *buffer* (*backlog*). Considerando-se que $a(t)$ é independente de $W(0)$ e $\alpha^*(\theta) < c$, então a função geradora de momento para o processo de *backlog* é limitada por (Dai, 1997):

$$E[e^{\theta W(t)}] \leq e^{\theta(\alpha^*(\theta)-c)t} e^{\theta\sigma^*(\theta)} E[e^{\theta W(0)}] + B(\theta), \tag{7.50}$$

onde

$$B(\theta) = \frac{(1 - e^{-c\theta}) e^{\theta\sigma^*(\theta)}}{1 - e^{\theta(\alpha^*(\theta)-c)}}. \tag{7.51}$$

Quando $t \to \infty$, verifica-se que o termo $B(\theta)$ é mais preciso do que o obtido em (Chang, 1994) por um fator igual a $(1 - e^{-c\theta})$. O valor limite para a função geradora de momento pode ser usado para se derivar limitantes para a probabilidade de perda e para o *backlog* em um nó com um único servidor, como segue (Dai, 1997):

i) A probabilidade de perda de *bytes* é limitada exponencialmente por

$$P[W(t) \geq w] \leq e^{-\theta w} \{e^{\theta(\alpha^*(\theta)-c)t} e^{\theta\sigma^*(\theta)} E[e^{\theta W(0)}] + B(\theta)\}. \tag{7.52}$$

ii) O tamanho médio da fila é limitado em

$$E[W(t)] \leq \frac{\{e^{\theta(\alpha^*(\theta)-c)t} e^{\theta\sigma^*(\theta)} E[e^{\theta W(0)}] + B(\theta)\}}{(1 - e^{-\theta})}. \tag{7.53}$$

No caso de modelagem de tráfego pelo MMW, substituindo a variável $\alpha^*(\theta)$ pelo valor da banda efetiva para o MMW dado pela equação (6.44) e $\sigma^*(\theta)$ por (7.48), é possível obter limitantes de desempenho para uma fila alimentada com processo caracterizado por parâmetros multifractais. O parâmetro θ que aparece nas equações (7.52) e (7.53), pode ser determinado através do ponto de operação (t^*, θ^*) introduzido na seção 7.2.3, ou seja,

$$\theta^* = \frac{b(t^*)^{-2H_g}}{\sigma^2(1 - H_g)}, \tag{7.54}$$

onde

$$t^* = \frac{H_g b}{(c-\mu)(1-H_g)} \tag{7.55}$$

e H_g é o parâmetro de escala global do processo MMW. Segundo as simulações realizadas, ao se determinar θ como sendo igual ao parâmetro do ponto de operação θ^* garante que limitantes mais rígidos sejam obtidos. A próxima seção verifica a acurácia desses limitantes para a probabilidade de perda de *bytes* e para o tamanho médio da fila.

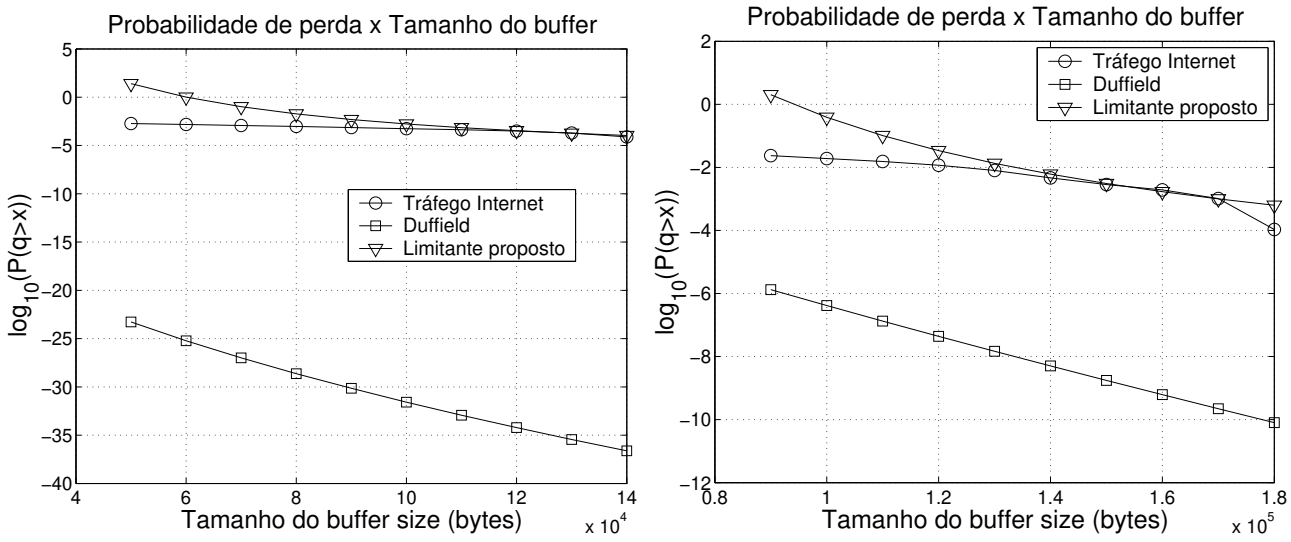
7.4.1 Resultados Experimentais para os Limitantes Obtidos

Utilizou-se nas simulações diferentes tipos de traços de tráfego como TCP/IP (lbl-pkt-5), Ethernet (Bc-Aug89) e o capturado na rede da Petrobrás (10-7-S-1) (Perlingeiro & Ling, 2005). Considerou-se amostras de tráfego em uma escala de agregação de 100ms, devido ao fato de os traços apresentarem características multifractais nesta escala (Erramilli et al., 2000) (Perlingeiro & Ling, 2005). Esta seção apresenta os resultados obtidos com as séries lbl-pkt-5, Bc-Aug89 e 10-7-S-1, com número de amostras N_t iguais a 2^{15}, 2^{14} e 2^{12}, respectivamente.

Deve ser ressaltado antes, que a Internet por exemplo, usa o controle de congestionamento com realimentação do TCP/IP, formando sistemas de malha fechada enquanto as simulações de fila aqui realizadas são de malha aberta (Low et al., 2002) (Postel, 1981) (Joo et al., 2001). Este capítulo se concentra em prover um estudo sobre o tráfego e seu comportamento de fila em um servidor, sem levar em consideração protocolos mais específicos de rede. Na verdade, como conseqüência, propomos métodos alternativos ao controle de malha fechada, como o esquema adaptativo de provisão de banda efetiva e os limitantes obtidos neste capítulo.

Na avaliação da proposta de cálculo de probabilidade de perda utilizando a equação (7.52), considerou-se um servidor com *buffer* finito com capacidade arbitrariamente escolhida igual a c=(w)x média da série de tráfego, onde w é igual a 5.6, 1.7 e 3.1, para as séries de tráfego lbl-tcp-5, 10-7-S-1 e Bc-Aug89, respectivamente. A probabilidade de perda em regime permanente calculada através da equação (7.52) para $t \rightarrow \infty$ e a probabilidade de perda obtida na simulação usando a série real de tráfego são apresentadas na Figura 7.4. A abordagem de Duffield e O'Connell subestima a taxa de perda para o tráfego real, como pode ser visto pela Figura 7.4. O método proposto utilizando Cálculo de Rede em conjunto com a banda efetiva para o MMW provê melhores resultados, se tornando mais precisos quanto maior o tamanho do *buffer*.

A Figura 7.5 mostra o limitante superior para o tamanho médio da fila utilizando a equação (7.53). Pode-se notar que à medida que se aumenta o tamanho do *buffer*, o limitante para a ocupação média do *buffer* se torna mais próximo da ocupação obtida via simulação com o traço de tráfego real.



(a) Traço de tráfego lbl-tcp-5

(b) Traço de tráfego 10-7-S-1

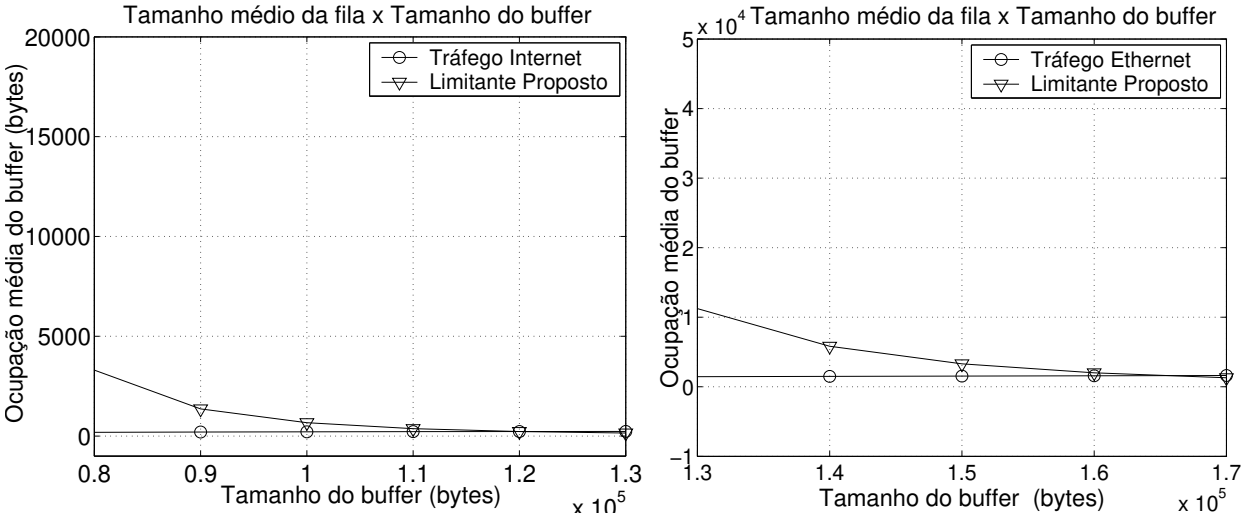
Fig. 7.4: Probabilidade de perda de *bytes* versus tamanho do buffer

7.5 Considerações Finais

Este capítulo inicialmente fez uma descrição de algumas abordagens existentes na literatura para a estimação de probabilidade de perda incluindo as teorias assintóticas de muitas fontes e dos grandes desvios e versões não-assintóticas aplicadas à processos multifractais. Em seqüência, derivou-se uma expressão para a probabilidade de perda em um servidor de fila com *buffer* finito cujo tráfego de entrada é dado pelo MMW. Mostrou-se através de simulações que os resultados da equação proposta são continuamente mais precisos com relação à variação do tamanho do *buffer* e ou a capacidade do servidor, do que os da Análise de Fila Multiescala.

O parâmetro de escala global derivado no Capítulo 3 foi usado neste capítulo no cálculo do ponto de operação do enlace. Demonstrou-se que o comportamento de fila pode ser melhor caracterizado utilizando-se Cálculo de Rede Estatístico associado à banda efetiva e ao parâmetro de escala global do MMW. O Cálculo de Rede permitiu se obter limitantes para a probabilidade de perda de *bytes* e para o tamanho médio da fila. As simulações comprovaram que o uso do Cálculo de Rede tornou mais preciso o limitante para a probabilidade de perda de *bytes* em comparação, por exemplo, ao obtido pela Teoria dos Grandes Desvios para processos com longa-dependência.

Os limitantes para a probabilidade de perda e para o tamanho médio da fila se mostraram adequados tanto para tráfego Internet quanto Ethernet. Mesmo o tráfego de 'backbone' Internet sendo considerado monofractal em escalas de tempo pequenas (1 a 100ms) como afirmam alguns autores



(a) Traço de tráfego lbl-tcp-5

(b) Traço de tráfego Bc-Aug89

Fig. 7.5: Ocupação média do buffer

(Zhang et al., 2003), os limitantes propostos também são aplicáveis a esse caso, pois os mesmos por tratarem de processos multifractais, englobam o caso monofractal. Estas características possibilitam a inclusão desses limitantes assim como a equação de probabilidade de perda proposta em esquemas de controle de admissão e outros tipos de controle de tráfego para garantir QoS. Além do mais, com a possibilidade de análise de vários fatores para as redes como perda e tamanho da fila, diferentes parâmetros de QoS (perda, atraso, *jitter*) podem ser simultaneamente analisados. O que de fato provê uma descrição mais completa do cenário de rede em questão.

Capítulo 8

Cálculo de Rede Estatístico para o MMW: Controle de Admissão

8.1 Introdução

Aplicações que exigem garantias de qualidade de serviço (QoS) têm sido cada vez mais encontradas na Internet, tais como voz sobre IP e vídeo conferência. Acredita-se que vídeo codificado MPEG-4 tomará grande parte das aplicações das redes sob a exigência de qualidade de serviço para serviços de vídeo. No entanto, hoje a Internet ainda provê em sua maioria serviço de melhor esforço, se tornando precária quando ocorre, por exemplo, congestionamento na rede ou surtos mais intensos de tráfego. Esse fato tem motivado muitas pesquisas e propostas de implementação de mecanismos de QoS na Internet. As propostas mais populares são as arquiteturas Intserv (*Integrated Services*) e Diffserv (*Differentiated Services*) (Blake, 1998)(Braden et al., 1994). Infelizmente o emprego de algoritmos de escalonamento baseados nas especificações dessas arquiteturas, resulta em uma diminuição na utilização da rede, principalmente quando o tráfego é composto de rajadas em diversas escalas de tempo (Chang, 1994).

Com o objetivo de prover uma eficiente multiplexação de fluxos de tráfego, o conceito de curva de serviço foi introduzido em (Cruz, 1991b), que especifica garantias de serviço por fluxo. Vários trabalhos que lidam com aplicações de curvas de serviço se baseiam no Cálculo de Rede para derivar limitantes de retardo e de perda (Cruz, 1991b)(Boudec & Thiran, 2001). No Cálculo de Rede em sua versão determinística envolvendo curvas de serviço e processos envelopes, as garantias de QoS são modeladas de forma determinística. Um serviço determinístico garante que todos os pacotes de um fluxo satisfaçam os limitantes de pior caso de retardo fim-a-fim e que não haja perda de pacote (Chang, 1994). Devido às curvas de serviço determinístico serem conservadoras, neste caso, freqüentemente os recursos da rede são subutilizados (baixa utilização da rede). Isto pode reduzir a eficiência da

multiplexação estatística. Dessa forma, a abordagem determinística não pode ser efetivamente usada na Internet futura. Para contornar esse problema, vários esquemas para prover garantias estatísticas de QoS foram propostos (Ferrari & Verma, 1990) (Reisslein et al., 2002) (Boorstyn et al., 2000). Nestes esquemas, ao se permitir que uma fração do tráfego viole as especificações de QoS se obtém um aumento significativo da utilização do enlace. Isso se torna possível visto que algumas aplicações típicas podem tolerar uma pequena taxa de perda, principalmente aquelas que aplicam técnicas de correção de erro (Wang & Zhu, 1998).

Para garantir QoS determinístico ou estatístico, os fluxos devem estabelecer contratos com a rede de maneira a limitar, de certo modo, a quantidade de tráfego a ser inserida na rede em um determinado intervalo de tempo. Apenas com o estabelecimento e a manutenção desses contratos espera-se que a rede possa prover garantias de QoS. Como exemplo, reguladores do tipo balde furado são mais freqüentemente usados para garantir o cumprimento dos contratos de tráfego. Tendo início com o trabalho de (Elwalid et al., 1995), várias pesquisas têm sido realizadas sobre ganhos de multiplexação estatística quando se assume que os fluxos são estatísticamente independentes e regulados, por exemplo, pelo algoritmo de balde furado (Knightly, 1998)(Rajagopal et al., 1998)(Reisslein et al., 2002). Estas propostas empregam envelopes determinísticos para caracterizar o grau de rajada do tráfego de entrada. Entretanto, para se atingir uma maior utilização da rede, e por fim, um maior número de fluxos admitidos, envelopes estatísticos são melhores opções (Boorstyn et al., 1999). Com o tráfego de entrada sendo caracterizado estatisticamente, a análise se torna mais complexa devido à existência de correlação entre os fluxos na saída do multiplexador. Pode-se obter limitantes estatísticos de QoS fim a fim evitando a correlação entre os fluxos (Reisslein et al., 2002) e ou através de controle da variação do retardo em cada nó (Ferrari, 1992)(Liebeherr, 2000). Kurose derivou limitantes estatísticos de QoS para sessões supondo que o tráfego de entrada é estocasticamente limitado, contudo o uso de suas estimativas leva a um controle de admissão bastante conservador. Em (Burchard et al., 2001), os autores definem uma curva estatística de serviço. Embora estes autores considerem o tráfego de entrada como sendo limitado por envelopes determinísticos, garantias probabilísticas para retardo e tamanho da fila no *buffer* são estabelecidas. Já Boorstyn et al. propõem um envelope estatístico de tráfego denominado de envelope efetivo, aplicando-o no controle de admissão para serviços estatísticos sob vários esquemas de escalonamento (Boorstyn et al., 1999), o que generaliza alguns resultados de trabalhos anteriores.

Além do emprego do Cálculo de Rede, um modo alternativo de se determinar os requisitos de recurso dos fluxos de tráfego é por meio da banda efetiva (Kelly, 1996) (Knightly & Shroff, 1999). Expressões de banda efetiva para vários modelos de tráfego foram propostas, até mesmo para tráfego auto-similar, mas não diretamente para algum modelo de tráfego multifractal.

No Capítulo 6, derivamos uma expressão de banda efetiva para o MMW, conseqüentemente tam-

bém aplicável às cascatas multiplicativas multifractais. O presente capítulo propõe a integração do conceito de banda efetiva de um processo multifractal com o formalismo do Cálculo Estatístico de Rede, onde ambos os processos de chegada de pacotes e as curvas de serviço são expressos em termos de limitantes probabilísticos, respectivamente, pelo envelope efetivo e pela curva de serviço efetiva. Como resultado, vários esquemas de escalonamento de fluxos de tráfego não facilmente tratáveis de forma analítica usando apenas a banda efetiva, agora podem ser analisados. Além disso, aborda-se pela primeira vez tráfego multifractal de entrada em um servidor em um contexto de Cálculo de Rede Estatístico. Através de curvas de serviços efetivas, analisamos os serviços dos seguintes esquemas de escalonamento: SP (*Static Priority*)(Boorstyn et al., 2000), EDF (*Earliest-Deadline-First*)(Georgiadis et al., 1996) e GPS (*Generalized Processor Sharing*) (Parekh & Gallager, 1993). Por fim, aplicamos o envelope efetivo obtido a partir da banda efetiva do MMW em um esquema de controle de admissão, comparando os resultados entre os diferentes escalonadores acima mencionados e para 3 modelos distintos de tráfego: fBm (fractional Brownian motion), tráfego regulado e tráfego multifractal (modelo MMW). As contribuições deste capítulo são: a) Obtenção de envelope efetivo, e assim, limitantes de desempenho fim-a-fim sem o uso de reguladores de tráfego. Verifica-se portanto, uma situação mais geral do que a encontrada na literatura onde se utilizam reguladores de tráfego para facilitar a obtenção das equações necessárias; b) Controle de admissão usando o envelope efetivo proposto para tráfego multifractal.

Este capítulo tem a seguinte organização: na seção 8.2, discute-se alguns métodos de controle de admissão. Na seção 8.3, são revistos alguns conceitos do Cálculo de Rede Determinístico, e a seção 8.4 trata do Cálculo de Rede Estatístico. Na seção 8.5, mostra-se como a teoria de banda efetiva pode se relacionar com o Cálculo de Rede Estatístico. Assim, deriva-se uma expressão para o cálculo do envelope efetivo para um processo multifractal baseado no MMW. Os envelopes efetivos tráfego fBm e tráfego regulado também são apresentados. Na seção 8.6, são apresentadas as expressões para as curvas de serviço efetivas de três esquemas de escalonamento: SP, EDF e GPS. Na seção 8.7, inicialmente compara-se os envelopes efetivos obtidos para os modelos de tráfego fBm, tráfego regulado e multifractal. E mais especificamente, na seção 8.7.2 se propõe um esquema de controle de admissão com garantias estatísticas de retardo e de perda para tráfego multifractal, comparando os resultados de nossa proposta com o de outros modelos de tráfego sob os esquemas de escalonamento SP, EDF e GPS. Finalmente, a seção 8.8 expõe as conclusões e considerações finais deste capítulo.

8.2 Controle de Admissão

Em redes com suporte à qualidade de serviço, um algoritmo de controle de admissão (CAC- *Connection Admission Control*) determina se um novo fluxo de tráfego pode ser admitido na rede, tal que

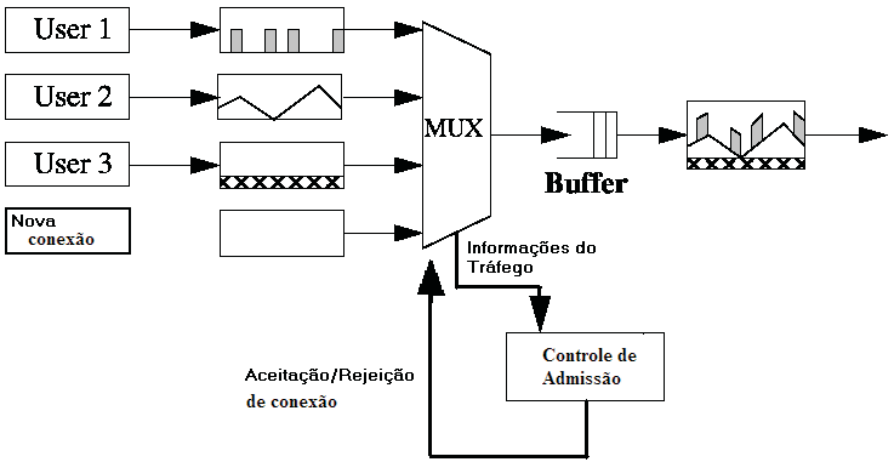


Fig. 8.1: Controle de admissão

os usuários obtenham o desempenho de rede requerido. O procedimento de controle de admissão é ilustrado na Figura 8.1. O controle de admissão efetua muitas funções importantes como dimensionar a banda necessária para os fluxos de tráfego em uma conexão. No contexto de redes ATM por exemplo, quando o fluxo de tráfego se insere na categoria de serviço do tipo taxa de bit constante CBR (*Constant Bit Rate*), sua taxa de pico é alocada para a conexão. Para tráfego sem taxa de bit especificada conhecido como UBR (*Unspecified Bit Rate*), a banda disponível é alocada, entretanto sem garantias de QoS (serviço do tipo 'melhor esforço'). Já para serviço com taxa variável VBR (*Variable Bit Rate*), destinado a aplicações que requerem banda variável garantida, a taxa alocada para a conexão poderia ser a banda efetiva, por ser a taxa realmente necessária para atender aos requisitos de QoS (Beran et al., 1995).

O desenvolvimento de algoritmos de controle de admissão e alocação de recursos tem encontrado dificuldades devido às características multifractais do tráfego e à complexa interação entre os fluxos e os multiplexadores (Lazar et al., 1994). Os algoritmos de controle de admissão podem ser divididos a grosso modo nas seguintes classes (Knightly & Shroff, 1999):

1) Baseado em taxa média e taxa de pico:

Nesta classe, a fonte de tráfego é caracterizada por sua taxa de pico e ou por sua taxa média. Um dos principais exemplos de controle de admissão desta classe é apresentado em (Ferrari & Verma, 1990). Neste artigo, os autores consideram um modelo On-Off para as fontes de tráfego, estimando a probabilidade de perda para o multiplexador sem *buffer*. Esta probabilidade de perda é então usada em um esquema de controle de admissão, indicando se os fluxos de tráfego podem ser admitidos no multiplexador.

2) Baseado em banda efetiva aditiva:

Vários algoritmos de controle de admissão baseados em banda efetiva foram propostos na litera-

tura, usando Teoria dos Grandes Desvios (Kesidis et al., 1993), decomposição em valores singulares de fluxos Markovianos (Elwalid & Mitra, 1993) e o Cálculo de Rede (Chang, 1994). Para acomodar conexões de taxa variáveis VBR de modo a atender seus requisitos de QoS descritos por suas bandas efetivas, a seguinte condição deve ser satisfeita (Berger & Whitt, 1998):

$$\sum_{i=1}^{n} n_i c_i \leq C, \tag{8.1}$$

onde C é a capacidade do enlace, c_i é a banda efetiva para conexões do tipo i, e n é o número de conexões do tipo i.

3) Baseado na curva de perda:

A curva de perda se refere à relação entre a probabilidade de perda e o tamanho do *buffer*. A curva de perda derivada via Teoria dos Grandes Desvios é exponencial, que pode ser uma aproximação conservadora em muitos casos. Assim, alguns trabalhos buscam projetar o formato da curva de serviço de tal forma a refletir melhor essa relação observada experimentalmente. Entre estes trabalhos, podem ser citados (Elwalid et al., 1995)(Kim & Shroff, 2001).

4)Baseado na máxima variância:

Sejam $A_j[s,t]$, o fluxo j de tráfego de chegada no intervalo $[s,t]$, $X_t = \sum_j A_j[s-t,s] - Ct$ e a variância normalizada $\tilde{\sigma}^2$ dada por

$$\tilde{\sigma}^2 = \frac{varX_t}{(B - EX_t)^2}. \tag{8.2}$$

Nesta classe, os algoritmos de controle de admissão consideram que o processo X_t seja Gaussiano. Com essa suposição, pode-se dizer que o instante de tempo t em que variância normalizada $\tilde{\sigma}^2$ for máxima é o mesmo instante em que a probabilidade de perda $P(X_t > B)$ atinge seu valor máximo, onde B é o tamanho do *buffer*. Dessa forma, um fluxo de tráfego é aceito se a estimativa da probabilidade de perda $P(X_t > B)$ não ultrapassar um determinado valor (Shroff & Schwartz, 1998).

5) Baseado em refinamentos da banda efetiva:

Esta classe inclui algoritmos de controle de admissão que estimam a banda efetiva utilizando a teoria das muitas fontes apresentada no Capítulo 7 (Courcoubetis et al., 1999) e outras abordagens que estendem a relação entre a probabilidade de perda e o tamanho do *buffer* usando uma função $g(B)$ da seguinte maneira (Duffield & O'Connell, 1993b):

$$\log P(Q > B) \sim -g(B). \tag{8.3}$$

Verifica-se neste caso que as bandas efetivas calculadas não são aditivas, possivelmente acarretando um aumento do número de fluxos aceitos atendendo aos requisitos de QoS.

Além das classes acima citadas, outros esquemas de controle de admissão foram propostos que

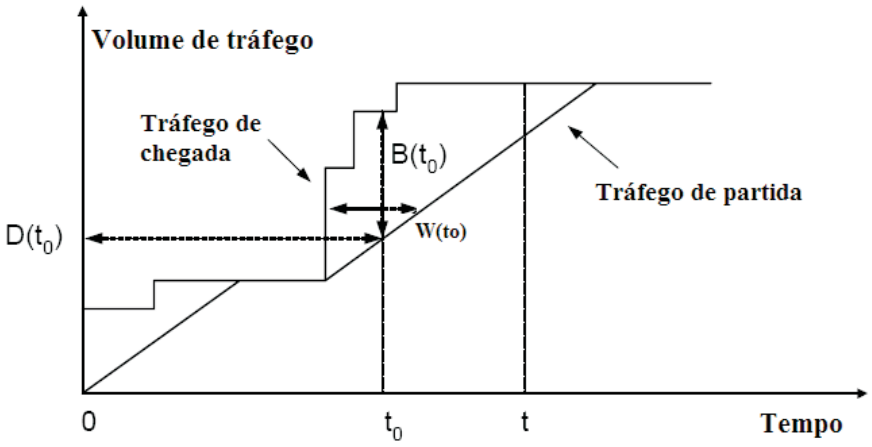


Fig. 8.2: Backlog e retardo

poderiam se encaixar em novas classes, como por exemplo, algoritmos baseados em medição de tráfego (Gibbens & Kelly, 1991) e baseado em estatísticas de pior caso para fluxos de tráfego policiados (Elwalid et al., 1995) (Rajagopal et al., 1998). Neste capítulo, a banda efetiva do modelo MMW é usada para calcular o envelope efetivo dos fluxos de tráfego e assim obter um controle de admissão com garantias estatísticas de perda e retardo.

8.3 Cálculo de Rede Determinístico

Os resultados do Cálculo de Rede Determinístico proporcionaram o desenvolvimento de novos algoritmos de escalonamento (Cruz, 1998)(Parekh & Gallager, 1993) e têm sido utilizados na especificação de novos serviços de rede (Blake, 1998)(Braden et al., 1994). Com o uso de envelopes de tráfego de chegada e de curvas de serviço, limitantes de retardo e de tamanho da fila no *buffer* podem ser expressos através do Cálculo Determinístico usando a álgebra 'min-plus' (Agrawal et al., 1999)(Berger & Whitt, 1998)(Boudec, 1998). A vantagem de se usar a álgebra 'min-plus' é que garantias fim-a-fim podem ser expressas pela concatenação das garantias em cada nó da rede.

No Cálculo de Rede, considera-se que as chegadas de pacotes (*bytes*) em um nó da rede e as saídas deste nó no intervalo de tempo $[0, t]$ podem ser representadas por funções não-negativas, não-decrescentes $A(t)$ e $D(t)$, respectivamente. Assim, o tamanho da fila no *buffer* (*backlog*) no instante de tempo t é dado por $B(t) = A(t) - D(t)$ e o retardo nesse mesmo instante por $W(t) = \inf\{d \geq 0 | A(t-d) \leq D(t)\}$ conforme é explicitado na Figura 8.2.

Sejam duas funções não-negativas e não-decrescentes $f(t)$ e $g(t)$ tais que $f(t), g(t) = 0$ se $t \leq 0$. A álgebra 'min-plus' aplicada à formulação do Cálculo de Rede define o seguinte para o operador de

convolução $*$ e o operador de desconvolução $\oslash$:

$$f * g(t) = \inf_{\tau \in [0,t]} \{f(t-\tau) + g(\tau)\} \tag{8.4}$$

$$f \oslash g(t) = \sup_{\tau \geq 0} \{f(t+\tau) - g(\tau)\} \tag{8.5}$$

Se f for não-decrescente, tem-se as seguintes relações:

$$f(t-\tau) = f * \delta_\tau(t) \tag{8.6}$$

$$f(t+\tau) = f \oslash \delta_\tau(t) \tag{8.7}$$

onde δ_τ é a função impulso definida como:

$$\delta_\tau = \begin{cases} \infty, & \text{se } t = \tau \\ 0 & \text{se } t \neq \tau \end{cases} \tag{8.8}$$

Estes operadores são usados na obtenção de garantias de serviço e de desempenho da rede. Denota-se por $A(x,y)$ e $D(x,y)$ as funções de chegada e partida no intervalo $[x,y)$, onde $A(x,y) = A(y) - A(x)$ e $D(x,y) = D(y) - D(x)$. As seguintes considerações são naturalmente feitas para as funções de chegada:

(A1) Não-negatividade: As chegadas em um intervalo de tempo são não-negativas. Ou seja, para qualquer $x < y$, temos que $A(y) - A(x) \geq 0$.

(A2) Limitante superior: o processo de chegada de tráfego $A(t)$ de um fluxo é limitado por uma função subaditiva A^*, chamada de processo envelope de chegada tal que $A(t+\tau) - A(t) \leq A^*(\tau)$ para quaisquer $t, \tau \geq 0$.

Uma curva mínima de serviço S para um fluxo, que define um limitante inferior para o serviço conferido a este fluxo, é tal que para todo $t \geq 0$, tem-se

$$D(t) \geq A * S(t). \tag{8.9}$$

Já uma curva de serviço máxima para um fluxo é uma função $\bar{S}$ que especifica um limitante superior para o serviço dado a um fluxo tal que para todo $t \geq 0$ pode ser dito que:

$$D(t) \leq A * \bar{S}(t). \tag{8.10}$$

Curvas de serviço mínimas têm um papel importante no Cálculo de Rede pois provêem garantias de serviços. Quando o processo de chegada é limitado por um processo envelope de chegada A^*,

tal que $A(t+\tau) - A(t) \geq A^*(\tau)$ para quaisquer $t, \tau \geq 0$, a garantia dada pela curva de serviço (8.9) implica em limitantes de piores casos para o retardo e 'backlog'. A curva de serviço mínima é freqüentemente referida na literatura como simplesmente curva de serviço. Quando uma curva de serviço máxima não é especificada, pode-se usar $\bar{S}(t) = ct$, onde c é a capacidade do enlace. De acordo com os trabalhos (Agrawal et al., 1999)(Boudec, 1998)(Chang, 2000), um processo envelope para o processo de partida de um nó oferecendo uma curva de serviço S é dado por $A^* \oslash S$, o 'backlog' é limitado por $A^* \oslash S(0)$ e o retardo em um nó, $W(t)$ é limitado por d, se este satisfaz $\sup_{\tau \geq 0}\{A^*(\tau - d) - S(\tau)\} \leq 0$. Os seguintes teoremas resumem alguns resultados do cálculo de rede determinístico (Agrawal et al., 1999)(Boudec, 1998)(Chang, 2000).

Teorema 8.3.1 *Dado um fluxo com envelope de chegada A^* e uma curva mínima de serviço S, pode-se verificar que:*

1. Envelope de saída: A função $D^* = A^* \oslash S$ é um envelope para o processo de partida, no sentido que para quaisquer $t, \tau \geq 0$,

$$D^*(t) \geq D(t+\tau) - D(\tau). \tag{8.11}$$

2. Limitante para o processo de tamanho da fila (backlog): Um limitante superior b_{max} para o backlog é dado por

$$b_{max} = A^* \oslash S(0). \tag{8.12}$$

3. Limitante de retardo: Um limitante superior de retardo, denotado por d_{max} é dado por

$$d_{max} = \inf\{d \geq 0 | \forall t \geq 0 : A^*(t-d) - S(t)\}. \tag{8.13}$$

O seguinte teorema afirma que as curvas de serviço de um fluxo nos nós de seu caminho podem ser concatenadas para definir uma curva de serviço de rede que expressa as garantias de serviço oferecidas aos fluxos da rede como um todo.

Teorema 8.3.2 *Suponha que um fluxo de tráfego passe por H nós em série e que em cada nó $h = 1, \ldots, H$, são oferecidas as curvas de serviço mínima e máxima S^h e $\bar{S}^h$ ao fluxo, respectivamente. Então, a seqüência de nós provê curvas de serviço mínima e máxima S^{net} e $\bar{S}^{net}$, dadas por:*

$$S^{net} = S^1 * S^2 * \ldots * S^H \tag{8.14}$$

$$\bar{S}^{net} = \bar{S}^1 * \bar{S}^2 * \ldots * \bar{S}^H \tag{8.15}$$

onde S^{net} e $\bar{S}^{net}$ são referidas como curvas de serviço de rede.

Aplicando-se os teoremas acima, as curvas de serviço de rede podem ser usadas para determinar limitantes de retardo e *backlog* para fluxos individuais na rede. Entretanto, como o Cálculo de Rede determinístico considera cenários de pior caso para perda e retardo, ignora assim, os efeitos da multiplexação estatística, implicando em uma superestimação dos requisitos de recursos dos fluxos multiplexados.

8.4 Cálculo de Rede Estatístico

O Cálculo de Rede Estatístico estende o conceito do Cálculo Determinístico a um contexto probabilístico no qual se pode explorar o ganho da multiplexação estatística. Os processos de chegada e de partida de tráfego no intervalo $[0, t]$ são descritos por processos estocásticos $A(t)$ e $D(t)$, que devem satisfazer as seguintes condições:

1) Para qualquer $\tau \geq 0$, o processo de chegada A_i^{net} de qualquer fluxo i da rede, deve satisfazer a seguinte equação:

$$\lim_{x \to \infty} \sup_{t \geq 0} Pr\{A_i^{net}(t+\tau) - A_i^{net}(t) > x\} = 0. \tag{8.16}$$

2) Os processos de chegada A_i^{net} e A_j^{net} de diferentes fluxos ($i \neq j$) são estatísticamente independentes.

Essas condições são impostas apenas na entrada da rede. A condição 1 reflete que limitantes estacionários são necessários para que se possa fazer conclusões que não dependam de instantes de tempo específicos, estendendo assim os resultados ao regime permanente. Ou seja, os valores esperados podem ser calculados como médias a longo prazo. A condição 2 de independência entre os fluxos na entrada da rede, nos permite explorar os ganhos obtidos com a multiplexação estatística.

Segundo os conceitos do Cálculo de Rede Estatístico apresentado em (Boorstyn et al., 2000)(Burchard et al., 2001) e assumindo as condições 1 e 2, a definição para o envelope efetivo de um processo de tráfego é dada a seguir.

Definição 8.4.1 *O envelope efetivo para um processo de chegada $A(t)$ é definido com sendo uma função não-negativa Γ^{ε}, tal que para todo t e τ, tem-se:*

$$Pr\{A(t+\tau) - A(t) \leq \Gamma^{\varepsilon}(\tau)\} > 1 - \varepsilon. \tag{8.17}$$

Pode-se dizer simplesmente que o envelope efetivo fornece um limitante estacionário para um processo de chegada de tráfego. Envelopes efetivos podem ser obtidos tanto para fluxos individuais como para multiplexados. Para caracterizar o serviço disponível a um ou mais fluxos são usadas curvas de serviço efetivas, que podem ser vistas como uma medida probabilística do serviço disponível

(Li et al., 2003). Dado um processo de chegada A, uma curva de serviço efetiva (mínima) é uma função não-negativa S^{ε} para a qual existe uma escala de tempo T, satisfazendo para todo $t \geq 0$ a seguinte equação:

$$Pr\{D(t) \geq \inf_{\tau \leq T}\{A(t-\tau) + S^{\varepsilon}(\tau)\} \geq 1 - \varepsilon. \tag{8.18}$$

Note que fazendo $\varepsilon = 0$ nas equações (8.17) e (8.18), obtém-se os envelopes de chegada e as curvas de serviço descritas pelo Cálculo Determinístico, onde esses eventos ocorrem com probabilidade igual a 1. Em geral, o valor da escala de tempo T depende do processo de chegada, assim como da curva de serviço. Sejam $A_{\mathcal{C}}(t)$, $D_{\mathcal{C}}(t)$ e $B_{\mathcal{C}}(t)$ as agregações dos processos de chegada, de partida e do tamanho da fila de um conjunto $\mathcal{C}$ de fluxos em um escalonador. Define-se período de ocupação (*busy period*) para um dado tempo $t \geq 0$, o intervalo máximo de tempo contendo t durante o qual o *backlog* devido aos fluxos em $\mathcal{C}$ permanece com valor positivo. Para escalonadores com uma disciplina de serviço conservativa, o valor da escala de tempo T pode ser limitado pelo tamanho do período de ocupação (*busy period*) do escalonador no instante de tempo $t \geq 0$ (Li et al., 2003). O começo desse período de ocupação *(busy period*) de t é o último instante ocioso (sem saída de dados) antes de t, dado por

$$\underline{t} = \max\{\tau \leq t : B_{\mathcal{C}} = 0\}. \tag{8.19}$$

Ao supor que os *buffer*s estejam vazios no instante $t = 0$, se garante que $0 \leq \underline{t} \leq t$. O lema enunciado a seguir estabelece que T é o limite probabilístico para o 'busy period' em enlaces com capacidade do servidor igual a C.

Lema 8.4.1 *Seja um escalonador obedecendo a uma disciplina de serviço conservativa com capacidade C. Para $\varepsilon > 0$ e supondo que exista uma função $\Gamma^{\varepsilon}_{\mathcal{C}}$ (envelope efetivo) tal que para todos os $t, \tau \geq 0$:*

$$\sum_{\tau=0}^{\infty} Pr\{A_{\mathcal{C}}(t+\tau) - A_{\mathcal{C}}(t) > \Gamma^{\varepsilon}_{\mathcal{C}}\} \leq \varepsilon. \tag{8.20}$$

Seja

$$T = \sup\{\tau | \Gamma^{\varepsilon}_{\mathcal{C}} > C\tau\}, \tag{8.21}$$

então, pode-se dizer que T satisfaz a seguinte equação (Li et al., 2003):

$$Pr\{t - \underline{t} \leq T\} \geq 1 - \varepsilon. \tag{8.22}$$

O seguinte teorema estabelece limitantes estatísticos para o retardo e o 'backlog' em termos de envelopes efetivos e curvas de serviço efetivas utilizando o Lema 8.4.1. Há 2 tipos de probabilidades de violação sendo consideradas neste teorema: ε_g é a probabilidade do tráfego de chegada violar o envelope efetivo e ε_s é a probabilidade do serviço fornecido ao tráfego violar a curva de serviço

efetiva ou a condição (8.18).

Teorema 8.4.2 *Seja Γ^{ε_g} um envelope efetivo para um processo de chegada A em um nó da rede e S^{ε_s}, uma curva de serviço efetiva que satisfaz a equação (8.18) para alguma escala de tempo $T < \infty$. Definindo-se ε como sendo*

$$\varepsilon = \varepsilon_s + T\varepsilon_g, \tag{8.23}$$

tem-se os seguintes resultados (Li et al., 2003):

1) Envelope de saída: A função $\Gamma^{\varepsilon_g} \oslash S^{\varepsilon_s}$ é um envelope efetivo para o tráfego de saída do nó.

*2) Limitante para o tamanho de fila: A função $\Gamma^{\varepsilon_g} \oslash S^{\varepsilon_s}(0)$ é um limitante probabilístico para o tamanho da fila, no sentido de que, para todo $t \geq 0$, temos $Pr\{B(t) \leq \Gamma^{\varepsilon_g} \oslash S^{\varepsilon_s}(0)\} \geq 1 - \varepsilon$, onde $B(t)$ é o tamanho da fila (*backlog*).*

3) Limitante de retardo: Se $d \geq 0$ é tal que satisfaz $\sup_{\tau \leq T}\{\Gamma^{\varepsilon_g}(\tau - d) - S^{\varepsilon_s}(\tau)\} \leq 0$, então d é um limitante de retardo probabilístico no sentido de que para todo $t \geq 0$, $Pr\{W(t) \leq d\} \geq 1 - \varepsilon$, onde $W(t)$ é o retardo no instante de tempo t.

Pode-se verificar que este teorema generaliza os resultados do cálculo determinístico que são obtidos ao se fazer $\varepsilon_s = \varepsilon_g = 0$. Ainda pode-se observar que quando apenas $\varepsilon_g = 0$, ou seja, para o caso em que a probabilidade do tráfego de chegada violar o envelope efetivo é zero, o limite T de escala de tempo desaparece da equação (8.23). Neste caso, se obtém o Cálculo de Rede Estatístico estabelecido em (Burchard et al., 2001), o qual considera processos envelopes determinísticos (onde $\varepsilon_g = 0$) e curvas de serviço efetivas S^{ε_s}.

Considere um fluxo que passa por um caminho composto de H nós. Para cada nó h, é alocada uma curva de serviço denotada por S^{h,ε_s}. De forma similar à equação (8.18), supõe-se que a seguinte relação seja válida para cada nó h:

$$Pr\{D^h(t) \geq \inf_{\tau \leq T^h}\{A^h(t-\tau) + S^{h,\varepsilon_s}(\tau)\} \geq 1 - \varepsilon_s, \tag{8.24}$$

com $T^h < \infty$, $h = 1, \ldots, H$. Por conveniência de notação, assume-se que a probabilidade de violação ε_s é idêntica em cada nó. Assim, o seguinte teorema pode ser estabelecido.

Teorema 8.4.3 *Suponha que um fluxo de tráfego passe por H nós em série e que ao fluxo são oferecidas as curvas de serviço S^{h,ε_s} em cada nó $h = 1, \ldots, H$ satisfazendo a equação (8.24). Então, a curva efetiva de serviço de rede $S^{net,\varepsilon}$ para o fluxo é dada por*

$$S^{net,\varepsilon} = S^{1,\varepsilon_s} * S^{2,\varepsilon_s} * \ldots * S^{H,\varepsilon_s}, \tag{8.25}$$

com probabilidade de violação limitada por

$$\varepsilon = \varepsilon_s \sum_{h=1}^{H} (1 + (h-1)T^h). \tag{8.26}$$

Nota-se que a probabilidade de violação (equação (8.26)) cresce a cada nó em $\varepsilon_s T^h$. Portanto, é importante controlar o limitante de escala de tempo T^h.

8.5 Envelope Efetivo e Banda Efetiva

Esta seção estabelece uma estratégia para combinar dois importantes métodos de caracterização probabilística de tráfego, o envelope efetivo e a banda efetiva. Expressões para a banda efetiva foram derivadas para vários modelos de tráfego. Ao se prover uma relação entre a banda efetiva e o envelope efetivo, estas expressão de banda efetiva podem ser aplicadas ao Cálculo de Rede, incluindo a banda efetiva do MMW apresentada no Capítulo 6.

8.5.1 Relação entre Envelope Efetivo e Banda Efetiva

O nome envelope efetivo por si só já sugere uma conexão com a banda efetiva. O seguinte lema apresentado em (Li et al., 2003) provê uma relação formal entre esses dois conceitos.

Lema 8.5.1 *Seja um processo de chegada $A(t)$ com banda efetiva $\alpha(s,\tau)$, o envelope efetivo para esse processo é dado pela seguinte equação:*

$$\Gamma^{\varepsilon}(\tau) = \inf_{s>0} \left\{ \tau\alpha(s,\tau) - \frac{\log \varepsilon}{s} \right\}. \tag{8.27}$$

De forma recíproca, para toda probabilidade de violação $\varepsilon \in (0,1)$, sendo a função Γ^{ε}, o envelope efetivo para o mesmo processo de chegada de tráfego $A(t)$, então pode-se dizer que a sua banda efetiva é limitada por (Li et al., 2003):

$$\alpha(s,\tau) \leq \frac{1}{s\tau} \log \left(\int_0^1 e^{\Gamma^{\varepsilon}(\tau) d\varepsilon} \right). \tag{8.28}$$

É importante salientar que o envelope efetivo é um conceito mais geral do que a banda efetiva no sentido de que a expressão de banda efetiva pode ser imediatamente expressa em termos do envelope efetivo, e o contrário nem sempre é possível. O Lema 8.5.1 permite construir os envelopes efetivos para todas as classes de tráfego cujas expressões de banda efetiva sejam conhecidas. Nas próximas

seções, com o uso do Lema 8.5.1, são obtidos os envelopes efetivos através das bandas efetivas para processos multifractais (MMW), para tráfegos regulados e para processos monofractais (fBm).

8.5.2 Envelope Efetivo para Cascatas Multiplicativas Multifractais

No Capítulo 6 derivou-se a banda efetiva para o MMW, que é perfeitamente aplicável a processos multifractais baseados em cascatas por ser expressa em termos dos multiplicadores $A_{i,j}$. Por conveniência, a equação de banda efetiva obtida é repetida a seguir:

$$\alpha(s,\tau) = \frac{1}{s\tau}\left(\log\frac{\sum_{k=1}^{2^j-1}|wg_{j,k}|}{2^j}\right) \tag{8.29}$$

onde

$$wg_{j,k} = 2^{-j/2}[e^{sv1} - e^{sv2}] \tag{8.30}$$

$$v1 = 2^{-j/2}U_{0,0}\prod_{i=0}^{j-1}[1+(-1)^{k'_i}A_{i,2k_i}] \tag{8.31}$$

$$v2 = 2^{-j/2}U_{0,0}\prod_{i=0}^{j-1}[1+(-1)^{k'_i}A_{i,2k_i+1}] \tag{8.32}$$

O envelope efetivo para tráfego multifractal através do MMW pode ser obtido pela aplicação do Lema 8.5.1 e da Proposição 6.4.1, ou seja, expressando o envelope efetivo (equação (8.27)) por meio das equações (8.29) e (8.30). Adicionalmente, se supõe que o parâmetro de tempo τ seja dado na forma diádica $\tau = 2^j$, onde $j = 1,\ldots,N$. Assim, o envelope efetivo para o MMW pode ser expresso como

$$\Gamma^{\varepsilon}(\tau) = \inf_{s>0}\left\{\frac{1}{s}\left(\log\frac{\sum_{k=1}^{\tau-1}|wg_{j,k}|}{\tau}\right) - \frac{\log\varepsilon}{s}\right\}, \tag{8.33}$$

onde $wg_{j,k}$ é dado pela equação (8.30).

Em resumo, o envelope efetivo para o MMW em um dado instante de tempo τ é obtido pelo valor atualizado da banda efetiva calculada com os dados da série de tráfego analisada até esse instante de tempo τ considerado, conforme descreve a equação (8.33).

8.5.3 Envelope Efetivo para Tráfego Regulado

Processos de tráfego de chegada limitados por algum processo envelope A^* são denominados de tráfegos regulados (Cruz, 1991b). Espera-se que um modelo de tráfego regulado seja capaz de

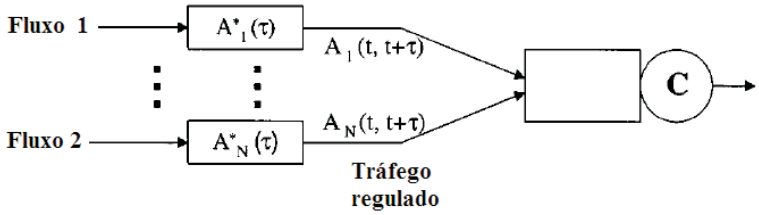


Fig. 8.3: Reguladores e escalonador em um enlace

descrever o comportamento do tráfego limitado na entrada da rede por algum algoritmo que desempenha esta função, por exemplo, pelo algoritmo de balde furado. A Figura 8.3 representa fluxos de tráfego sendo regulados e escalonados em um nó da rede. Mais formalmente, seja A^* uma função não-decrescente, não-negativa e subaditiva. Um processo de chegada A é regulado por A^*, se para quaisquer $t, \tau \geq 0$, tem-se

$$A(t+\tau) - A(t) \leq A^*(\tau). \tag{8.34}$$

A taxa de pico P e a taxa média ρ para esse tráfego regulado são dadas respectivamente pelas equações (Li et al., 2003):

$$P = A^*(1) \tag{8.35}$$

e

$$\rho = \lim_{t\to\infty} \frac{A^*(t)}{t}. \tag{8.36}$$

Seja um conjunto $\mathcal{C}$ de fluxos para os quais A_i^*, P_i e ρ_i são o processo envelope de chegada, a taxa de pico, e a taxa média do fluxo i, respectivamente. Assumindo que os fluxos sejam independentes e que cada fluxo i satisfaça a seguinte condição:

$$E[A_i(t+\tau) - A_i(t)] \leq \rho_i \tau, \tag{8.37}$$

a banda efetiva para este conjunto $\mathcal{C}$ de fluxos obedece a seguinte desigualdade (Kelly, 1996):

$$\alpha_{\mathcal{C}}(s,t) \leq \frac{1}{st} \sum_{i\in\mathcal{C}} \log\left(1 + \frac{\rho_i t}{A_i^*(t)}(e^{sA_i^*(t)} - 1)\right). \tag{8.38}$$

Por meio do Lema 8.5.1, o correspondente envelope efetivo de (8.38) pode ser escrito como

$$\Gamma_{\mathcal{C}}^{\varepsilon}(t) = \inf_{s>0}\left\{\sum_{i\in\mathcal{C}} \frac{1}{s}\log\left(1 + \frac{\rho_i t}{A_i^*(t)}(e^{sA_i^*(t)} - 1)\right) - \frac{log\varepsilon}{s}\right\}. \tag{8.39}$$

8.5.4 Envelope Efetivo para o Modelo fBm

Norros propôs um modelo parcimonioso de tráfego denominado tráfego Browniano fracionário (fractional Brownian traffic, fBt), capaz de representar matematicamente as características auto-similares observadas nos dados de tráfego (Norros, 1995). Denomina-se tráfego Browniano fracionário, o processo $A(t)$ que descreve o volume de tráfego acumulado até o instante de tempo t, representado pela seguinte equação:

$$A(t) = \rho t + \beta Z(t), \tag{8.40}$$

onde ρ é a taxa média de chegada do tráfego, β é o desvio-padrão do volume de tráfego de chegada em uma unidade de tempo e o processo $Z(t)$ é o movimento Browniano fracionário normalizado com parâmetro de Hurst $H \in [1/2; 1)$.

Conforme discutido no Capítulo 2, o modelo fBm exibe longa-dependência, característica também constatada em traços de tráfego real (Erramilli et al., 1996), ou seja, a função de autocorrelação de um processo fBm tem um decaimento mais lento do que a exponencial. Porém, a distribuição de probabilidade gaussiana do modelo fBm dificulta o seu emprego na modelagem de tráfego em pequenas escalas de tempo, visto que nestas escalas, o tráfego apresenta uma forte característica não-gaussiana.

A banda efetiva para o tráfego Browniano fracionário (fBm) é dada por (Kelly, 1996):

$$\alpha_{\mathcal{C}}(s,t) = \rho_{\mathcal{C}} + \frac{1}{2}\beta_{\mathcal{C}}^2 s t^{2H-1}. \tag{8.41}$$

Novamente, com o uso do Lema 8.5.1, o envelope efetivo para o tráfego Browniano fracionário pode ser escrito como (Li et al., 2003):

$$\Gamma_{\mathcal{C}}^{\varepsilon}(t) = \rho_{\mathcal{C}} t + \sqrt{-2\log\varepsilon}\beta_{\mathcal{C}} t^{H}. \tag{8.42}$$

8.6 Curvas de Serviço Efetivas

Esta seção apresenta limitantes inferiores para os serviços oferecidos a uma classe de fluxos por meio de curvas de serviço efetivas. Mas antes disso, discute-se sobre as disciplinas de escalonamento de fluxos de dados para os quais serão derivadas as curvas de serviço efetivas.

8.6.1 Escalonamento de Fluxos de Dados

As redes de computadores permitem que as sessões ou fluxos de dados compartilhem recursos, tais como a taxa de transmissão de um enlace e os espaços em *buffers* nos pontos de multiplexação. Entretanto, a competição pelo uso desses recursos pode originar situações de contenção. Dada uma fila de usuários requisitando um determinado recurso, um servidor deve realizar uma tarefa de escalonamento, ou seja estabelecer uma ordem de atendimento destas requisições. Uma disciplina de escalonamento normalmente realiza duas funções:

1) Decide a ordem de atendimento das requisições de serviço;

2) Gerencia a fila de requisições de serviço.

As disciplinas de escalonamento são amplamente utilizadas na camada do protocolo de redes ou em sistemas onde pode ocorrer contenção de recursos. Elas desempenham um papel fundamental na provisão de qualidade de serviço às aplicações, permitindo controle diferenciado de atraso, de taxa de transmissão e de taxa de perda dos dados.

Em geral, espera-se que um algoritmo de escalonamento possua as seguintes características: facilidade de implementação, proteção aos fluxos presentes e alocação dos recursos de modo justo. O conceito de proteção refere-se ao isolamento dos fluxos de entrada do escalonador, para que um fluxo mal-comportado não afete a qualidade de serviço dos demais fluxos. Em relação ao conceito de justiça na alocação de recursos, costuma-se utilizar como base o critério *max-min* (*max-min fairness*), definido da seguinte maneira (Keshav, 2001):

1) Recursos são alocados em ordem crescente de valor solicitado, normalizado pela ponderação do fluxo;

2) Nenhum fluxo obtém uma quota de utilização superior à solicitada;

3) Fluxos cujos requisitos não possam ser satisfeitos, compartilham os recursos disponíveis na proporção de suas ponderações.

Na concepção de um mecanismo de escalonamento, duas questões devem ser consideradas: a prioridade de cada fluxo e a conservação do serviço pelo escalonador. Entre as disciplinas de escalonamento mais conhecidas estão: a disciplina *First-Come First-Served* (FCFS) e a *Generalized Processor Sharing* (GPS) (Keshav, 2001). A disciplina FCFS é caracterizada pela simplicidade na implementação. Basicamente esta disciplina é constituída por apenas uma fila e as requisições são atendidas na ordem de chegada. A desvantagem desta disciplina está no fato de não oferecer proteção aos fluxos, nem alocação justa de recursos. Estas duas últimas características são desejáveis para a diferenciação de serviço e estão presentes na disciplina GPS.

Seja A_q denotando as chegadas de fluxos da classe q, pertencente a uma das Q classes existentes e $A_{\mathcal{C}}$ as chegadas do conjunto $\mathcal{C}$ dos fluxos de todas as classes $q = 1, \ldots, Q$. Neste seção, considerou-se um enlace operando com disciplina de serviço conservativa de capacidade igual a C e três esquemas

de escalonamento diferentes: (**SP**-Static Priorities) (Boorstyn et al., 2000), (**EDF**-Earliest Deadline First) (Georgiadis et al., 1996), (**GPS**-Generalized Processor Sharing) (Parekh & Gallager, 1993), os quais são brevemente descritos a seguir.

1) **Escalonador SP**: a cada classe de fluxo é designado um índice de prioridade, onde um índice de baixo valor indica uma alta prioridade. O escalonador SP mantém uma fila FCFS para cada nível de prioridade. São transmitidos primeiramente os pacotes dos fluxos da fila FCFS de mais alta prioridade. Um problema típico da disciplina SP é que pacotes com baixa prioridade podem ficar esperando indefinidamente por serviço em suas filas correspondentes.

2) **Escalonador EDF**: a cada classe q associa-se um índice de retardo d_q. Para um pacote da classe q chegando em um instante de tempo t, é associado um 'prazo de atendimento' $t + d_q$ e o escalonador EDF sempre seleciona o pacote com menor 'prazo de atendimento'. O escalonador EDF é ótimo no sentido de minimização de latência em um contexto de serviços determinísticos (Georgiadis et al., 1996).

3) **Escalonador GPS**: O escalonador GPS (*Generalized Processor Sharing*) é uma das disciplinas de escalonamento conservativa mais estudadas (Parekh & Gallager, 1993). O WFQ (Weighed Fair Queueing) é a modalidade de GPS mais conhecida na prática (Keshav, 2001). Entre suas propriedades mais importantes estão a proteção aos diferentes fluxos e o compartilhamento da taxa de serviço. Devido a estas duas características, o escalonador GPS é altamente recomendado em redes que necessitem de suporte à diferenciação de serviços. Para cada classe q é designado um peso ϕ_q e se garante que esta classe receba uma taxa mínima de serviço g_q de pelo menos $\frac{\phi_q}{\sum_p \phi_p}$ da capacidade C disponível, ou seja,

$$g_q = \frac{\phi_q}{\sum_p \phi_p} C. \tag{8.43}$$

Um servidor GPS certifica-se que fluxos com dados enfileirados, com quantidade de taxa ainda insuficiente para suas necessidades, compartilhem a taxa de serviço remanescente de outros fluxos em proporção a seus pesos. Desta forma, pode-se dizer que a disciplina de escalonamento GPS é justa em relação ao critério de justiça *max-min* (Keshav, 2001).

Em seguida, o Lemma 8.6.1 estabelece as curvas de serviço efetivas (para cada classe de tráfego q) para os três esquemas de escalonamento descritos. Estas curvas de serviço efetivas são funções determinísticas e baseadas na banda não usada por outros fluxos de tráfego $p \neq q$ (Li et al., 2003).

Lema 8.6.1 *Sejam fluxos de Q classes chegando em um servidor com capacidade C. Para cada classe $q = 1, ..., Q$, $\Gamma_q^{\varepsilon_g}$ representa o envelope efetivo para o processo de chegada A_q. Seja T tal que satisfaça a equação (8.22), ou seja, $P\{t - \underline{t} \leq T\} \geq 1 - \varepsilon_b$, para um processo agregado A_C para algum valor $\varepsilon_b < 1$. Adicionalmente, assume-se que no caso do escalonador GPS, a função $\Gamma_p^{\varepsilon_g}$ seja côncava. Assim sendo, as curvas de serviço efetivas $S_q^{\varepsilon_s}$ podem ser escritas como (Li et al., 2003):*

1) **SP***:*

$$S_q^{\varepsilon_s}(t) = \left[Ct - \sum_{p<q} \Gamma_p^{\varepsilon_g}(t) \right]_+, \qquad \varepsilon_s = \varepsilon_b + (q-1)T\varepsilon_g \tag{8.44}$$

2) **EDF***:*

$$S_q^{\varepsilon_s}(t) = \left[Ct - \sum_{p\neq q} \Gamma_p^{\varepsilon_g}(t - [d_p - d_q]_+) \right]_+, \qquad \varepsilon_s = \varepsilon_b + (Q-1)T\varepsilon_g \tag{8.45}$$

3) **GPS***:*

$$S_q^{\varepsilon_s}(t) = \lambda_q \left(Ct + \sum_{p\neq q} [\lambda_p Ct - \Gamma_p^{\varepsilon_g}(t)]_+ \right), \qquad \varepsilon_s = \varepsilon_b + (Q-1)T\varepsilon_g \tag{8.46}$$

onde $\lambda_q = \frac{\phi_q}{\sum_p \phi_p}$ *é a taxa garantida para a classe* q*. Em cada caso,* $S_q^{\varepsilon_s}$ *é uma curva de serviço efetiva para a classe* q *que satisfaz a seguinte equação:*

$$Pr\left\{ D_q(t) \geq \inf_{\tau \leq T} \{ A_q(t-\tau) + S_q^{\varepsilon_s}(\tau) \} \right\} \geq 1 - \varepsilon_s. \tag{8.47}$$

Fazendo as probabilidades de violação ε_b e ε_g iguais a zero, obtém-se o limitante inferior para o serviço visto por cada classe de serviço igual ao da formulação determinística. As curvas de serviço efetivas do Lema 8.6.1 representam limitantes inferiores da capacidade não usada por outras classes diferentes de q (Li et al., 2003). Com isso, a equação (8.44) descreve com precisão o desempenho de um escalonador SP. Já para o escalonador EDF, no limite onde o retardo d é o mesmo para todas as classes ($d_p = d_q$ para todas classes $p \neq q$), a equação (8.45) se aproxima do serviço de um escalonador SP para a classe de mais baixa prioridade, enquanto na verdade, o escalonador EDF se aproxima da disciplina de serviço FIFO (FCFS) (Li et al., 2003).

8.7 Validação Experimental

Nesta seção são apresentados os resultados experimentais que ilustram o ganho de multiplexação de diferentes modelos de tráfego (tráfego regulado, fBm, multifractal) e dos algoritmos de escalonamento SP, EDF e GPS. Para isso, primeiramente compara-se os envelopes efetivos obtidos para os modelos de tráfego considerados. Em seguida, verifica-se o número de fluxos admitidos utilizando o esquema de controle de admissão proposto. Na obtenção dos resultados experimentais, considerou-se o algoritmo de balde furado com envelope de chegada dado por $A^*(t) = \min(Pt, \sigma + \rho t)$ para o tráfego regulado.

8.7.1 Comparacão de Envelopes Efetivos

Esta seção avalia os envelopes efetivos obtidos utilizando as bandas efetivas calculadas para tráfego regulado pelo algoritmo de balde furado, para tráfego fBm e para o modelo MMW. A análise foi realizada considerando o envelope efetivo normalizado pelo número de fluxos, ou seja $\Gamma^{\varepsilon}_{N}(t)/N$, onde $\Gamma^{\varepsilon}_{N}(t)$ é o envelope efetivo para N fluxos iguais.

Nas análises experimentais desta seção, considerou-se as características de 2 fluxos de tráfego diferentes (Fluxo 1 e Fluxo 2). Uma vez que as simulações envolvem eventos discretos no tempo, tomou-se 1ms como unidade de tempo. Os parâmetros da Tabela 8.1 foram escolhidos de acordo com as estatísticas das séries dec-pkt-2 (Fluxo 1) e lbl-pkt-5 (Fluxo 2) na escala de tempo de 1ms, as quais são séries de tráfego TPC-IP. Os diagramas multiescala linear dessas séries são apresentados na seção 3.3.1 do Capítulo 3, onde se pode encontrar uma discussão sobre suas características multifractais. As Figuras 8.4 e 8.5 mostram os envelopes efetivos por fluxo com $\varepsilon_g = 10^{-9}$ para os Fluxos 1 e Fluxos 2, respectivamente. Foram incluídas nos gráficos as taxas médias das fontes de tráfego e para o tráfego regulado foi inserido também seu envelope determinístico dado por $\min(Pt, \sigma + \rho t)$. Além disso, para o tráfego regulado, considerou-se como parâmetros para o algoritmo de balde furado a taxa de pico (P), a taxa média (ρ) das séries de tráfego dec-pkt-2 e lbl-pkt-5 (Tabela 8.1). O valor de σ^2 é igual à variância dos traços considerados.

Quanto menor for o envelope efetivo para o intervalo de tempo considerado, menor será a quantidade de *bytes* exigida por fluxo para que se garanta a probabilidade de violação desejada (10^{-9}). Pela análise das Figuras 8.4 e 8.5 pode-se observar que a medida que se aumenta o número de fluxos, uma menor quantidade de *bytes* por fluxo é estabelecida para que se atenda aos requisitos de QoS. Pode-se notar também que o envelope efetivo proposto é menor do que o do fBm, mas é maior do que o do modelo tráfego regulado. Isso mostra que a capacidade do servidor exigida pelo modelo fBm para atender aos requisitos de QoS é maior do que a que realmente é necessária. Para o modelo tráfego regulado, uma menor capacidade é exigida do que nos outros 2 casos, porém o tráfego regulado corresponde a uma situação na qual se força o tráfego a ter um comportamento menos intenso (maior controle das rajadas) do que o tráfego multifractal.

Tab. 8.1: Parâmetros de Tráfego para Cálculo de Envelope Efetivo

Fluxo	P(bytes)	ρ(bytes)	σ(bytes)	β^2	H
1	1536	$92,26$	$232,3772$	$5.4\text{x}10^4$	$0,8084$
2	1024	$10,72$	$82,5621$	$6.8165\text{x}10^3$	$0,7810$

8.7.2 Controle de Admissão por Envelope Efetivo

Seja um enlace com um único servidor com capacidade de 1Mbps para atender aos fluxos de tráfego Fluxo 1 e Fluxo 2. Nesta seção, as estatísticas destes fluxos de tráfego são extraídas das séries dec-pkt-2 e lbl-pkt-5, agora na escala de 100ms (Tabela 8.2). Avalia-se aqui o serviço oferecido ao Fluxo 1 sob os 3 algoritmos de escalonamento citados: SP, EDF e GPS. Supõe-se que os Fluxos 1 devem satisfazer um limite de retardo de d =200ms. Dado um certo número de Fluxos 2 no enlace de 1Mbps, com o uso dos resultados do Lema 8.6.1, determina-se o número máximo de fluxos do Fluxos 1 que podem ser admitidos no enlace de modo a não violar o limitante probabilístico de retardo d estabelecido, tal que d satisfaça a seguinte equação:

$$\sup_{t \leq T} \{\Gamma^{\varepsilon_g}(\tau - d) - S^{\varepsilon_s}\} \leq 0. \tag{8.48}$$

Os parâmetros dos algoritmos de escalonamento são os seguintes: os índices de prioridade para a disciplina de escalonamento SP, os índices de retardo para o EDF e os pesos para o GPS. Para o escalonamento SP, os Fluxos 1 possuem índices de prioridade mais altos. Para o escalonamento EDF, considerou-se com relação aos Fluxos 1, o índice de retardo $d_1 = 200$ms e para os Fluxos 2, $d_2 = 100$ms. No escalonamento GPS, fixou-se os pesos em $\phi_1 = 0.25$ e $\phi_2 = 0.75$. Assim como feito na seção anterior, o envelope efetivo foi calculado de 3 formas: para tráfego regulado, para tráfego fBm e através da nossa proposta de banda efetiva. Os parâmetros de tráfego utilizados são apresentados na Tabela 8.2. Para efeito de comparação, são incluídos também nos gráficos desta seção, o número de fluxos que podem ser acomodados no enlace considerado através da alocação de banda usando a média e a taxa de pico dos fluxos.

A Figura 8.6 apresenta o número de Fluxos 1 que pode ser admitido sem violação do limitante de retardo requerido $(d = 200ms)$ e da probabilidade de violação $\varepsilon_g = 10^{-6}$ em função do número de Fluxos 2 já encontrados no sistema. Pode-se observar que a escolha do modelo de tráfego tem um impacto significativo no número admissível de Fluxos 1 no enlace. O número de Fluxos 1 que pode ser admitido com o modelo fBm é menor do que o com a nossa proposta e o para tráfego regulado. Para o tráfego regulado, por ser uma 'versão controlada' do tráfego de entrada, foi admitido um número ligeiramente maior de fluxos. Entretanto, novos reguladores de tráfego têm sido propostos, sendo mais eficientes no cumprimento dos contratos de tráfego estabelecidos do que o algoritmo de balde furado (Fonseca et al., 2000) (Roberts, 2000). Pode-se constatar também que resultados similares são obtidos para os escalonadores SP e GPS. Um maior número de Fluxos 1 admitidos pode ser observado para o escalonador EDF com relação a todos os modelos de tráfego considerados. Esse número se reduz com a diminuição dos valores de retardo d_1 e d_2. Outro fato importante é que, dado um determinado modelo de tráfego, para o escalonador GPS, o número mínimo de Fluxos 1 admitido

é independente do número de Fluxos 2. Isto se deve ao fato de que o GPS garante um número mínimo de Fluxos 1.

Tab. 8.2: Parâmetros de Tráfego para o Controle de Admissão

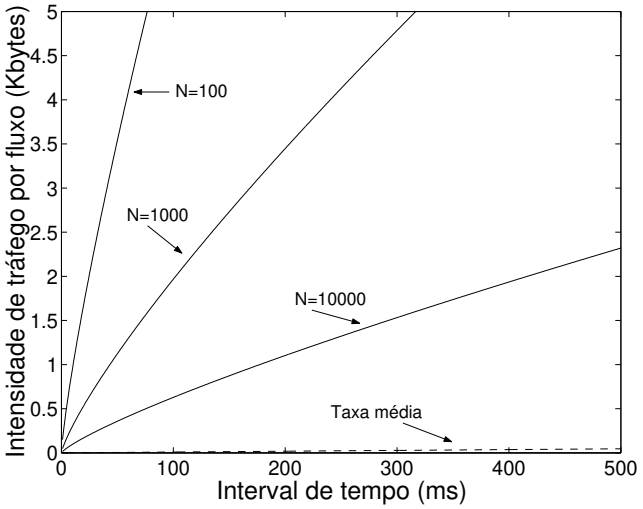
Tipo	P(Kbytes)	ρ(Kbytes)	σ(Kbytes)	β^2	H
1	$81,4$	$26,5$	$12,3$	$151730,40$	$0,8080$
2	$45,9$	$2,70$	$3,20$	$10032,70$	$0,7820$

8.8 Considerações Finais

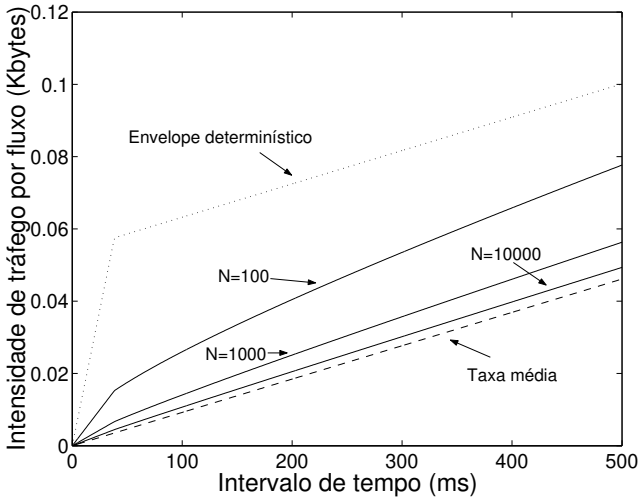
Neste capítulo, usando a banda efetiva do modelo multifractal MMW, obtivemos seu envelope efetivo correspondente. Na análise dos envelopes efetivos, o modelo monofractal fBm apresentou um maior envelope efetivo do que o do tráfego multifractal MMW e o do tráfego regulado. Isso mostra que o modelo monofractal fBm exige mais recursos do que realmente é necessário para se garantir uma dada QoS. A análise realizada se restringiu às escalas de tempo em que se constatou a presença de 'multifractalidade' no tráfego (Erramilli et al., 1996). Nestas escalas o envelope efetivo proposto, por ser baseado em um modelo multifractal, é mais adequado, já que o modelo MMW possui distribuição aproximadamente lognormal assim como o tráfego real, diferente do modelo fBm que é Gaussiano.

A união entre o Cálculo de Rede Estatístico (envelope efetivo e curvas de serviço efetivas) com o conceito de banda efetiva, nos permitiu propor um esquema de controle de admissão de fluxos para tráfego não-regulado. Limitantes de probabilidade de perda e retardo são mais fáceis de serem obtidos para tráfego regulado e foram bastante explorados na literatura (Rajagopal et al., 1998) (Reisslein et al., 2002). Este capítulo mostra que é possível atingir um número de fluxos admitidos comparado ao do tráfego regulado. Esta característica de admissão de fluxos é desejável pois busca-se por algoritmos eficientes de controle de admissão, mas que não apresentem os problemas dos algoritmos de policiamento existentes. O algoritmo de balde furado, como exemplo de regulador de tráfego, não descreve adequadamente a variabilidade do tráfego e pode não assegurar o correto policiamento dos fluxos de tráfego (Roberts, 2000).

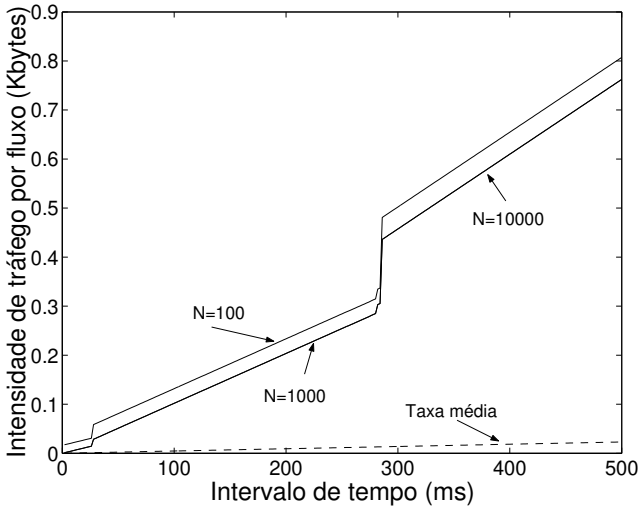
O esquema de controle de admissão proposto controla a admissão de fluxos de modo que sejam atendidos ambos requisitos probabilísticos de perda e retardo, considerando um modelo de tráfego mais preciso do que os modelos monofractal e tráfego regulado. Como este esquema de controle de admissão se baseia em Cálculo de Rede Estatístico, com isso, generaliza a sua aplicação a vários tipos de escalonadores e contextos, podendo ser empregado em diferentes arquiteturas de rede. Assim



(a) fBm

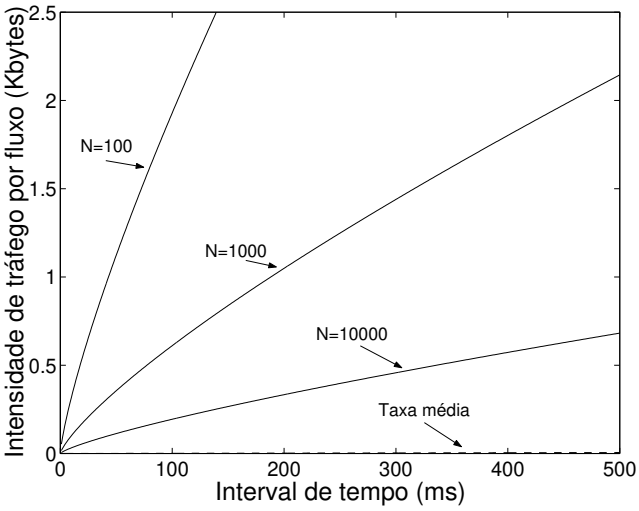


(b) Tráfego regulado

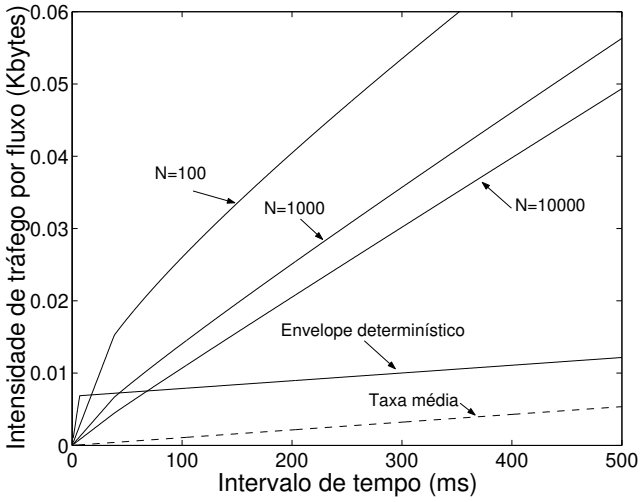


(c) Tráfego multifractal

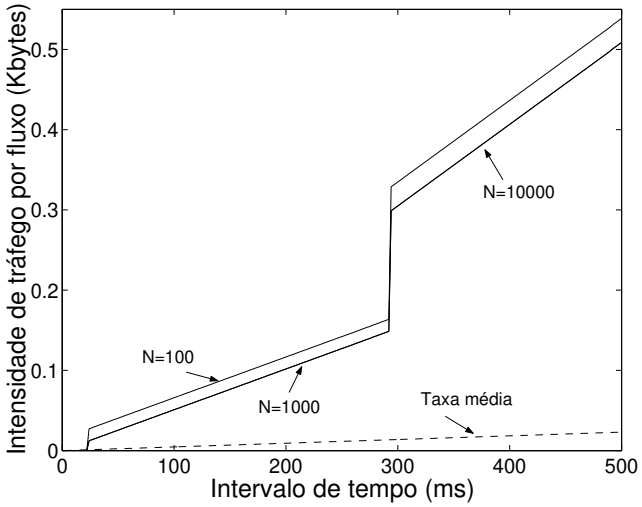
Fig. 8.4: Envelope efetivo por fluxo ($\mathcal{G}_N^{\varepsilon}(t)/N$) para os Fluxos 1



(a) fBm

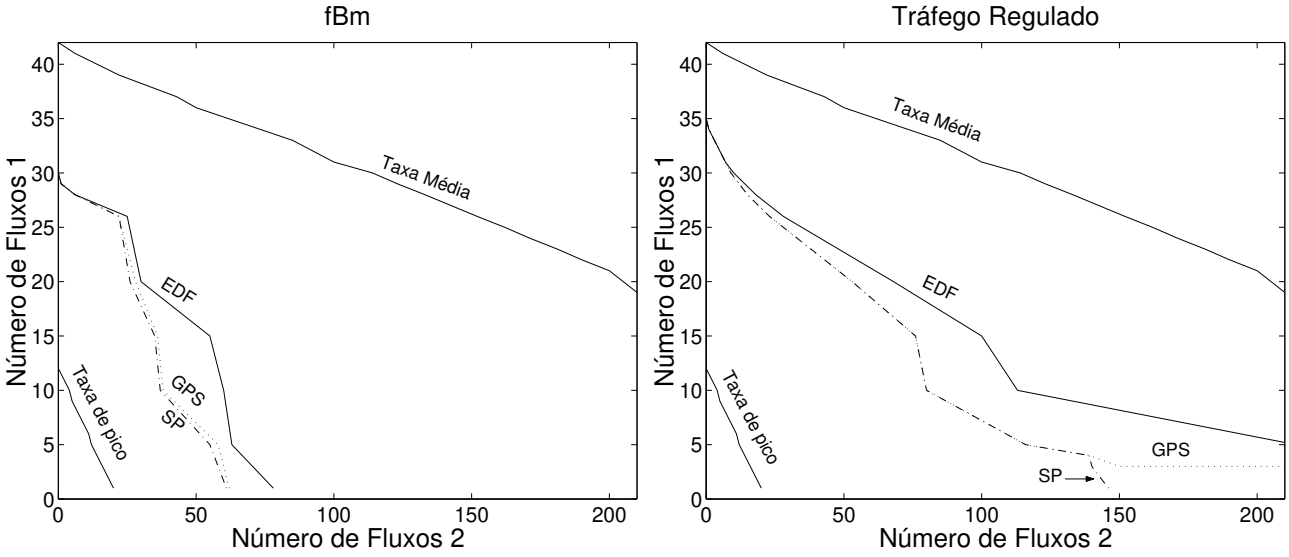


(b) Tráfego regulado



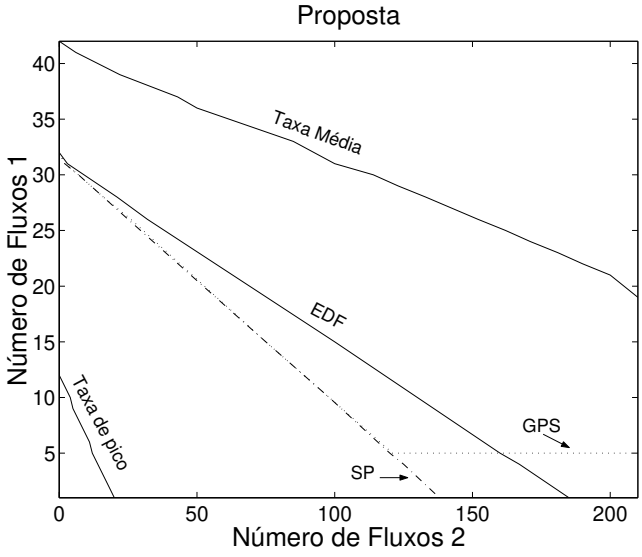
(c) Multifractal

Fig. 8.5: Envelope efetivo por fluxo ($\mathcal{G}_N^{\varepsilon}(t)/N$) para os Fluxos 2



(a) fBm

(b) Tráfego regulado



(c) Multifractal

Fig. 8.6: Número de Fluxos 1 em função do número de Fluxos 2 (C=1Mbps) para diferentes escalonadores, $\varepsilon_g = 10^{-6}$, $d = 200$ms, $\phi_1 = 0.25$ e $\phi_2 = 0.75$.

sendo, pode ser visto como uma solução eficaz para se prover QoS em redes como a Internet.

Capítulo 9

Conclusões

O tráfego de redes possui características que são descritas mais adequadamente por modelos multifractais (Riedi & Véhel, 1997) (Ribeiro et al., 2000). Sabe-se que algumas das propriedades dos processos multifractais têm impacto direto no desempenho das redes (Grossglauser & Bolot, 1999) (Erramilli et al., 1996). Com isso, métodos mais eficientes de alocação de recursos podem ser obtidos ao se considerar modelos multifractais.

Em sua essência, esta tese abordou soluções para alocação de banda e estimação de perda que podem ser utilizadas para a provisão de QoS nas redes multiserviços atuais. Investigou-se o problema de alocação de banda para tráfego multifractal sob vários aspectos. Sob a luz da análise multifractal, analisou-se de modo geral os seguintes temas: predição de tráfego, estimação de banda efetiva e de probabilidade de perda e controle de admissão com garantia de QoS.

Neste estudo, desenvolveu-se um modelo multifractal a partir de propriedades das *wavelets* aplicadas a processos multifractais baseados em cascatas multiplicativas. Através de simulações com séries de tráfego reais, mostrou-se que as características do tráfego podem ser eficientemente capturadas por este modelo multifractal de tráfego. Projetou-se um modelo capaz de ajustar seus parâmetros à medida que dados de tráfego são recebidos e disponibilizados em tempo real. Isto possibilita que se tenha uma 'visão' sempre atualizada da dinâmica do tráfego e como conseqüência, os parâmetros estimados para o modelo refletem melhor o real comportamento dos fluxos de tráfego no momento.

Um parâmetro relacionado a processos auto-similares bastante mencionado na literatura é o parâmetro de Hurst, que mede o grau de auto-similaridade de um processo. Neste trabalho, um parâmetro semelhante foi derivado com base nos processos multifractais. Este parâmetro, denominado de parâmetro de escala global, é capaz de indicar o grau de auto-similaridade de um processo, porém é calculado a partir dos parâmetros do MMW.

A característica de dependência de longo prazo, comum a processos mono e multifractais e

presente nas séries de tráfego reais, interfere fortemente no desempenho das redes (Beran, 1995). Comprovou-se neste estudo pela análise da função de autocorrelação derivada para o MMW, a presença de dependência de longo prazo nos dados de tráfego sintéticos gerados segundo o MMW. Portanto, demonstrou-se de forma analítica que o MMW apresenta as características de auto-similaridade e dependência de longo prazo.

Nesta tese, além de se realizar a análise do MMW e de outros modelos multifractais, também se verificou a melhoria obtida em termos de desempenho de predição com a inserção de informação provinda do MMW em um algoritmo de predição adaptativo baseado na modelagem *fuzzy* TSK. A modelagem *fuzzy* TSK foi escolhida como base para esta proposta por sua facilidade em considerar funções de base ortonormais em sua estrutura e pelo seu excelente desempenho de predição. Como contribuição original, foram incorporados conceitos multifractais em um preditor adaptativo *fuzzy*. Para que isso fosse possível, desenvolveu-se um algoritmo de treinamento adaptativo para o preditor *fuzzy* com funções de base ortonormais. A incorporação destas funções de base ortonormais proporcionou um melhor desempenho de predição ao modelo preditor adaptativo *fuzzy* proposto. Estas funções de base ortonormais foram obtidas a partir do pólo sugerido neste trabalho, o qual também é calculado adaptativamente.

Uma das propostas expostas neste trabalho considera o controle preditivo da capacidade do servidor como método de garantir perda zero aos fluxos. Este tipo de controle é importante por amenizar a perda de pacotes e por propiciar que os protocolos de rede possam realizar ações para diminuir o congestionamento, até mesmo antes que ele ocorra. Devido a isso, analisou-se a eficiência de algoritmos adaptativos de predição de tráfego e a aplicação destes em esquemas de alocação de banda. Através das simulações, verificou-se que o preditor *fuzzy* proposto é mais robusto e preciso do que os outros preditores comparados, alocando bandas mais apropriadas pelo esquema adaptativo de alocação de banda proposto do que os esquemas baseados em preditores lineares.

Os esquemas de alocação de banda baseados em predição em sua maioria objetivam ter perda zero, pois a banda alocada é freqüentemente a dada pelas predições do valor da taxa de pico da série de tráfego considerada. Porém, muitas aplicações em redes reais permitem uma pequena perda ou retardo. Neste caso, a banda efetiva é usada para se estimar a taxa realmente necessária para atender aos requisitos de perda e retardo, fazendo com que uma maior utilização do enlace seja obtida e mais fluxos de tráfego possam ser multiplexados. Os resultados experimentais mostraram que a banda efetiva do MMW atingiu os objetivos de QoS impostos a uma maior utilização da rede, comparada à banda efetiva para processos monofractais. Mas em ambientes dinâmicos é interessante que a banda efetiva seja também estimada adaptativamente. Deve-se ressaltar que um esquema de provisão adaptativa de banda efetiva evita que mudanças bruscas no comportamento do tráfego façam com que a alocação de banda seja ineficiente, como pode ocorrer no caso de alocação de banda

estática. Verificou-se para o algoritmo de estimação adaptativa de banda efetiva proposto que, além dos requisitos de QoS serem atendidos, uma maior utilização do enlace foi obtida.

A taxa de perda de pacotes ou *bytes* é um dos principais parâmetros para se avaliar a qualidade de serviço oferecido a um fluxo de tráfego. Neste estudo, o comportamento de fila, principalmente descrito pela probabilidade de perda, foi caracterizado analiticamente através dos parâmetros e propriedades do MMW em termos de sua banda efetiva e parâmetro de escala global. A análise não-assintótica proposta para a estimação da probabilidade de perda de pacotes produziu resultados mais realistas e robustos do que os da Análise de Fila Multiescala (Ribeiro et al., 2000), o que a torna uma alternativa analítica promissora no projeto de redes. Análise assintótica do comportamento de fila também foi realizada, onde utilizou-se a banda efetiva do MMW para se obter limitantes para a probabilidade de perda e o tamanho médio de fila no *buffer*, conduzindo a resultados mais precisos do que os obtidos simplesmente pela Teoria dos Grandes Desvios (Duffield & O'Connell, 1993a).

Outra ferramenta para análise de tráfego, tão poderosa quanto a banda efetiva é o envelope efetivo, oriundo do cálculo de rede estatístico. Neste trabalho, estabeleceu-se pela primeira vez o envelope efetivo para um modelo multifractal. Através deste envelope efetivo foi possível propor um esquema de controle de admissão, o qual pode ser aplicado a vários contextos de redes para garantir que os fluxos atendam simultaneamente a requisitos de perda e de retardo.

Uma das possíveis extensões deste trabalho é o de avaliar em um simulador de rede, o controle de admissão proposto considerando diferentes tipos de escalonadores. Pretende-se também analisar melhor o impacto das variações dos parâmetros do MMW no desempenho de rede. Ainda como trabalho futuro, vislumbra-se aplicar as ferramentas aqui desenvolvidas em diversas tecnologias de redes. De forma geral, pode-se concluir que as propostas apresentadas nesta tese contribuem intensamente para a pesquisa na área de engenharia de tráfego e o desenvolvimento de novas tecnologias de controle de tráfego.

Referências Bibliográficas

Abry, P. & Veitch, D. (1998). Wavelet analysis of long-range dependent traffic. *IEEE Trans. on Information Theory*, *44*, 2–15.

Adas, A. (1996). Supporting real time VBR video using dynamic reservation based on linear prediction. In *Proc.IEEE INFOCOMM'96*, (pp. 1476–1483).

Adas, A. & Mukerjee, A. (1995). On resource management and QoS guarantees for long range dependent traffic. In *Proc. IEEE INFOCOMM*, (pp. 779–787).

Adas, A. M. (1998). Using adaptive linear prediction to support real-time VBR video under RCBR network service model. *IEEE/ACM Trans. Net.*, *6*(5), 635–645.

Addie, R. G. & Zukerman, M. (1994). An approximation for performance evaluation of stationary single server queues. *IEEE Trans. Commun.*, *42*, 3150–3160.

Addie, R. G., Zukerman, M., & Neame, T. D. (1995). Fractal traffic: Measurements, modeling and performance evaluation. In *Proceedings of INFOCOM*, (pp. 977–984).

Agrawal, R., Cruz, R. L., Okino, C., & Rajan, R. (1999). Performance bounds for flow control protocols. *IEEE/ACM Transactions on Networking*, *7*(3), 310–323.

Aquino, V. A. & Barria, J. A. (2006). Multiresolution FIR neural-network-based learning algorithm applied to network traffic prediction. *IEEE Transactions on Systems, Man and Cybernetics-C*, *36*(2), 208–220.

Aspirot, L., Belzarena, P., Bermolen, P., Ferragut, A., Perera, G., & Simon, M. (2005). Quality of service parameters and link operating point estimation based on effective bandwidths. *Performance Evaluation*, *59*(Issues 2-3), 103–120.

Awduche, D., Berger, L., Gan, D., Li, T., & Srinivasan, V. (2001). RSVP-TE: extensions to RSVP for LSP tunnels. RFC3209.

Baiocchi, A., Melazzi, N., Listani, M., Roveri, A., & Winkler, R. (1991). Loss performance analysis of an ATM multiplexer loaded with high-speed on-off sources. *IEEE J. Select. Areas Commun.*, *9*, 388–393.

Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (1993). *Nonlinear Programing: Theory and Algorithms* (second ed.). New York, NY: John Wiley and Sons.

Beran, J. (1995). *Statistics for Long-Memory Processes*. New York, NY 10119: Chapman and Hall.

Beran, J., Sherman, R., Taqqu, M., & Willinger, W. (1995). Variable-bit-rate video traffic and long-range dependence. *IEEE Trans. Commun.*, *43*, 1566–1579.

Berger, A. W. & Whitt, W. (1998). Extending the effective bandwidth concept to networks with priority classes. *IEEE Communications Magazine*, *36*(8), 78–84.

Bezdek, J. C. (1993). Fuzzy models-what are they and why? *IEEE Trans. Fuzzy Syst.*, 1–6.

Bianchi, G. R., Vieira, F. H. T., & Ling, L. L. (2004a). A novel network traffic predictor based on multifractal caracteristic. In *GLOBECOM'04*, Dallas, Texas, EUA.

Bianchi, G. R., Vieira, F. H. T., & Ling, L. L. (2004b). Predictive dynamic bandwidth allocation based on multifractal traffic characteristics. In *ICT (International Conference on Telecomunications)*, Fortaleza,Ceará.

Bianchi, G. R., Vieira, F. H. T., & Ling, L. L. (2004c). Um modelo multifractal aplicado à predição de tráfego de redes. In *XXI Simpósio Brasileiro de Telecomunicações*, Belém, Pará.

Bickel, D. R. & West, B. J. (1998). Multiplicative and fractal processes in DNA evolution. *Fractals*, *6*, 211–217.

Blake, S. (1998). An architecture for differentiated services. RFC 2745.

Boorstyn, R., Burchard, A., Liebeherr, J., & Oottamakorn, C. (1999). Statistical multiplexing gain of link scheduling algorithms in QoS networks. Cs-99-21, Univ. Virginia, Computer Science Dep.

Boorstyn, R. R., Burchard., A., Liebeherr, J., & Oottamakorn, C. (2000). Statistical service assurances for traffic scheduling algorithms. *IEEE Transactions on Selected Areas in Communications*, *18*(12), 2651–2664.

Boudec, J. Y. L. (1998). Application of network calculus to guaranteed service networks. *IEEE/ACM Transactions on Information Theory*, *44*(3), 1087–1097.

Boudec, J. Y. L. & Thiran, P. (2001). *Network Calculus*. Lecture Notes in Computer Science. New York: Springer-Verlag.

Box, G. E. & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.

Braden, R., Clark, D., & Shenker, S. (1994). Integrated services in the internet architecture: an overview. IETF RFC 1633.

Brockwell, P. J. & Davis, R. A. (1996). *Introduction to Time Series and Forecasting*. Springer.

Brockwell, R. & R.Davies (1991). *Time series: Theory and methods*. New York: Springer.

Bucklew, J. A. (1990). *Large deviation techniques in decision, simulation and estimation*. New York: J.Wiley.

Burchard, A., Liebeherr, J., & Patek, S. D. (2001). A calculus for end-to-end statistical service guarantees (revised). Cs-2001-19, University of Virginia Computer Science Dep.

Burrus, C. S., Gopinath, R. A., & Guo, H. (1998). *Intoduction to Wavelets and Wavelet Transforms*. Prentice Hall.

Cappe, O., Moulines, E., Pesquet, J. C., Petropulu, A., & Yang, X. (2002). Long-range dependence and heavy-tail modeling for teletraffic data. *IEEE Signal Procesing Magazine*, *19*(3), 14–27.

Chang, C. (1994). Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. Automat. Contr.*, *39*, 913–931.

Chang, C. S. (2000). *Performance guarantees in comunication networks*. Springer.

Chang, C. S. & Thomas, J. A. (1995). Effective bandwidth in high-speed digital networks. *IEEE J. Select. Areas Commun.*, *13*(6), 913–931.

Chen, B. S., Peng, S. C., & Wang, K.-C. (2000). Traffic modeling, prediction, and congestion control for high-speed networks: A fuzzy AR approach. *IEEE Transactions on Fuzzy Sytems*, *8*(5).

Chen, B. S., Yang, Y. S., Lee, B. K., & Lee, T. H. (2003). Fuzzy adaptive predictive flow control of atm network traffic. *IEEE Trans. on Fuzzy Syst.*, *11*(4), 568–581.

Chong, S., Li, S. Q., & Ghosh, J. (1995). Predictive dynamic bandwidth allocation for efficient transport of real-time VBR video over ATM. *IEEE JSAC*, *13*(1), 12–23.

Choudhury, G. L., Lucantoni, D. M., & Whitt, W. (1994). On the effectiveness of effective bandwidths for admission control in ATM networks. In *In Proc. of the 14th International Teletraffic Congress (ITC-14)*, (pp. 411–420)., North Holland. Elsevier Science B. V.

Chuang, C. C., Hsiao, C. C., & Jeng, J. T. (2003). Adaptive fuzzy regression clustering algorithm for TSK fuzzy modeling. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*.

Courcoubetis, C., Kelly, F. P., & Weber, R. (1997). Measurement-based usage charges in communications networks. 1997-19, Statistical Laboratory, University of Cambridge.

Courcoubetis, C., Siris, V. A., & Stamoulis, G. D. (1999). Application of the many sources asymptotic and effective bandwidths to traffic engineering. *Telecommunication Systems*, *12*, 167–191.

Courcoubetis, C. & Weber, R. (1994). Effective bandwidths for stationary sources. *University of Crete, Greece*.

Courcoubetis, C. & Weber, R. (1996). Buffer overflow asymptotics for a switch handling many traffic sources. *Journal of Applied Probability*, *33*, 886–903.

Crovella, M. & Taqqu, M. (1999). Estimating the heavy tail index from scaling properties. *Methodology and Computing in Applied Probability*, *1*(1), 55–79.

Crovella, M. E. & Bestavros, A. (1996). Self-similarity in world wide web traffic - evidence and possible causes. In *Proceedings of ACM Sigmetrics* (pp. 160–169).

Cruz, R. (1998). Sced+: efficient management of quality of service guarantees. In *Proceedings of IEEE INFOCOM 98*, San Francisco, CA,.

Cruz, R. L. (1991a). A calculus for network delay part II: Network analysis. *IEEE Trans. Information Theory*, *37*, 132–141.

Cruz, R. L. (1991b). A calculus for network delay part I:network elements in isolation. *IEEE Trans. Information Theory*, *37*, 114–131.

Dai, L. (1997). Effective bandwidths and performance bounds in high-speed communication systems. In *Decision and Control Proceedings of the 36th IEEE Conference*, (pp. 4580–4585).

Dang., T. D., Molnár, S., & Maricza, I. (2002). Capturing the complete characteristics of multifractal network traffic. In *GLOBECOM 2002*, Taipei, Taiwan.

Dang, T. D., Molnár, S., & Maricza, I. (2003). Queuing performance estimation for general multifractal traffic. *Int. J. Commun. Syst.*, *16*(2), 117–136.

Daoudi, K., Lévy-Véhel, J., & Meyer, Y. (1998). Construction of continuous functions with prescribed local regularity. *Journal of constructive approximation*, *14*(3), 349–385.

Daoudi, K. & Véhel, J. L. (1995). Speech modeling based on regularity analysis. In *Proceedings of the IASTED/IEEE international conference on signal and image processing*, Las Vegas.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. New York: SIAM.

de Prycker, M. (1995). *Asynchronous Transfer Mode Solution For Broadband ISDN*. Prentice-Hall.

de Veciana, G., Kesidis, G., & Walrand, J. (1995). Resource management in wide-area ATM networks using effective bandwidths. *IEEE J. Select. Areas Commun.*, *13*, 1081–1090.

Dembo, A. & Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Springer-Verlag.

Devroye, L. (1989). The double kernel method in density estimation. In *Anais do Instituto Henri Poincaré*, volume 25, (pp. 533–580).

Drummond, A. C. (2005). *Alocação de banda em redes auto-ajustáveis*. Tese de mestrado, Unicamp.

Duan, Z., Zhang, Z. L., & Hou, Y. T. (2002). Service overlay networks: SLA, QOS and bandwidth provisioning. In *Proceedings of IEEE International Conference on Network Protocols (ICNP)*, (pp. 334–343).

Duffield, N. G. (1994). Exponential bounds for queues with markovian arrivals. *Queueing systems*, *17*, 413–430.

Duffield, N. G., Goyal, P., & Greenberg, A. (1999). A flexible model for resorce management in virtual private networks. In *Proceedings of SIGCOMM*, (pp. 95–108).

Duffield, N. G., Lewis, J. T., O'Connell, N., Russel, R., & Tomey, F. (1995). Entropy of ATM traffic streams: a tool for estimating QoS parameters. *IEEE Journal on Selected Areas in Commun.*, *13*(6), 981–990.

Duffield, N. G. & O'Connell, N. (1993a). Large deviations and overflow probabilities for the general single-server queue, with applications. Technical Report 1, Dublin Institute for Advanced Studies-Applied Probability Group, DIAS-STP-93-30.

Duffield, N. G. & O'Connell, N. (1993b). Large deviations and overflow probabilities for the general single-server queue, with applications. *DIAS STP, UK*.

Duffy, K. (2000). On the large deviations of a class of stationary on/off sources which exhibit long range dependence. *Dublin Institute for Advanced Studies - Dublin, Ireland*.

Dumont, G. A. & Fu, Y. (1993). Non-linear adaptive control via Laguerre expansion of Voltera kernels. *Int. J. Adaptive Control and Signal Processing*, *7*, 367–382.

Ellis, R. (1984). Large deviations for a general class of random vectors. *Ann.Prob.*, *12*, 1–12.

Elwalid, A. & Mitra, D. (1993). Effective bandwidth of general markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking*, *1*(3).

Elwalid, A., Mitra, D., & Wentworth, R. (1995). A new approach for allocating buffers and bandwidth to heterogeneous regulated traffic in an ATM node. *IEEE J. Select. Areas Commun.*, *13*, 1115–1127.

Embrechts, P. & Maejima, M. (2000). An introduction to the theory of self-similar stochastic processes. *International Journal of Modern Physics B*, *14*, 1399–1420.

Erramilli, A., Narayan, O., Neidhardt, A., & Saniee, I. (2000). Performance impacts of multi-scaling in wide area TCP/IP traffic. In *Proc. Infocom*.

Erramilli, A., Narayan, O., & Willinger, W. (1996). Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. on Net.*, *4*(2).

Euntai, K., Minkee, P., Seunghwan, J., & Mignon, P. (1997). A new approach to fuzzy modeling. *IEEE Trans. Fuzzy Syst.*, *5*(3), 328–337.

Falconer, K. (1990). *Fractal geometry: mathematical foundations and applications*. Nova York: John Wiley and Sons.

Feldmann, A., Gilbert, A. C., & W.Willinger (1998). Data networks as cascades: Investigating the multifractal nature of internet WAN traffic. (pp. 25–38)., Vancouver. ACM/SIGCOMM'98.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons / Mei Y A Publications Inc.

Fengyuan, R. (2002). A robust queue management algorithm based on sliding mode variable structure control. In *Proc. IEEE Infocom'02*, New York, NY.

Ferrari, D. (1992). Design and application of a delay jitter control scheme for packet-switching internetworks. *Computer Communications*, *15*(6), 367–373.

Ferrari, D. & Verma, D. (1990). A scheme for real-time channel establishment in wide-area networks. *IEEE J. Select. Areas Commun.*, *8*, 368–379.

Fitzek, F. H. P. & Reisslein, M. (2000). MPEG-4 and H.263 video traces for network performance evaluation (extended version). Tkn-00-06, Telecommunication Networks Group, Technical University of Berlin.

Fitzek, F. H. P. & Reisslein, M. (2001). MPEG-4 and H.263 video traces for network performance evaluation. *IEEE Trans. on Networking*, *15*(6), 40–54.

Fonseca, N. L. S., Mayor, G. S., & Neto, C. A. V. (2000). On the equivalent bandwidth of self-similar sources. *ACM Transactions on Modeling and Computer Simulation*, *10*(3), 104–124.

Fu, Y. & Dumont, G. A. (1993). An optimum time scale for discrete Laguerre network. *IEEE Transactions on Automatic Control*, *38*(6), 934–938.

Gao, J. & Rubin, I. (1999a). Multifractal analysis and modeling of long range-dependent traffic. In *Proceedings of ICC'99*.

Gao, J. & Rubin, I. (1999b). Multifractal modeling of counting processes of long-range dependent network traffic. In *SCS Advanced Simulation Technologies Conf.*, San Diego, CA.

Gao, J. & Rubin, I. (2000). Superposition of multiplicative multifractal traffic streams. In *In Proceedings of ICC 2000*.

Gao, J. & Rubin, I. (2001). Multiplicative multifractal modelling of long-range dependent network traffic. *International Journal of Telecommucication Systems*, *14*, 783–801.

Garret, M. W. & Willinger, W. (1994). Analysis, modeling and generation of self-similar vbr vide traffic. In *Proceedings of ACM Sigcomm* (pp. 269–280).

Georgiadis, L., Guérin, R., Peris, V., & Sivarajan, K. N. (1996). Efficient network QoS provisioning based on per node traffic shaping. *IEEE/ACM Trans. Networking*, *4*, 482–501.

Gibbens, R. J. (1996). *Stochastic Networks: Theory and Applications*, chapter Traffic characterization and effective bandwidths for broadband network traces, (pp. 169–179). Oxford Univ.Press.

Gibbens, R. J. & Hunt, P. J. (1991). Effective bandwidths for the multi-type UAS channel. *Queue Syst .*, *9*, 17–28.

Gibbens, R. J. & Kelly, F. P. (1991). Measurement-based connection admission control. In *15th International Teletraffic Symposium*, (pp. 879–888).

Glynn, P. W. & Whitt, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Probab. A (special issue)*, *31*, 131–156.

Groskinsky, B., D.Medhi, & D.Tipper (2001). An investigation of adaptive capacity control schemes in a dynamic traffic environment. *IEICE Trans. on Comm*, *E84-B*(2), 263–274.

Grossglauser, M. & Bolot, J. C. (1999). On the relevance of long-range dependence in network traffic. *IEEE/ACM Transactions on Networking*, *7*(5), 629–640.

Guérin, R., Ahmadi, H., & Naghshineh, M. (1991). Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE JSAC*, *9*(7968-981).

Gupta, V. & Waymire, E. (1993). A statistical analysis of mesoscale rainfall as a random cascade. *Journal of Applied Meteorology*, *32*, 251–267.

Gyorgy, A. & Borsos, T. (2001). Estimates on the packet loss ratio via queue tail probabilities. In *GLOBECOM'01*, volume 4, (pp. 2407–2411).

Harmantzis, F. C., Hatzinakos, D., & Lambadaris, I. (2003). Effective bandwidths and tail probabilities for gaussian and stable self-similar traffic. volume 3 (pp. 1515–1520). IEEE International Conference on Communications.

Hayes, M. H. (1996). *Statistical Digital Signal Processing and Modeling*. John Wiley of Sons.

Haykin, S. S. (1989). *Modern filters*. New York: Macmillan Publishing Company.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of statistics*, *3*, 1163–1174.

Hiramatsu, A. (1990). ATM communications netwok control by neural networks. *IEEE Transations on Neural Networks*, *1*(1), 120–130.

Hirchoren, G. A. (1999). *Predição e estimação de parametros de processos auto-similares para redes de faixa larga*. Tese de doutorado, Universidade Estadual de Campinas. Faculdade de Engenharia Eletrica e de Computação., Campinas, SP.

Hirchoren, G. A. & Arantes, D. S. (1998). Predictors for the discrete time fractional gaussian process. In *Proceedings of SBT/IEEE International Telecommunications Symposium ITS'98*, volume 1 (pp. 49–53).

Hobson, E. W. (1958). *The theory of functions of a real variable and the theory of Fourier's series*, volume I. Nova York: Dover Publications Inc.

Jain, R. (1990). Congestion control in computer networks: Issues and trends. *IEEE Network Magazine*, 24–30.

Joo, Y., Ribeiro, V., Feldman, A., Gilbert, A. C., & Willinger, W. (2001). TCP/IP traffic dynamics and network performance: a lesson in workload modeling, flow control and trace driven simulations. *Computer Communications Review*, *31*, 25–37.

Jorge, C., Vieira, F. H. T., & Ling, L. L. (2005a). Escalonamento GPS baseado na regularidade local do tráfego internet. In *SBrT'05*, Campinas, SP.

Jorge, C., Vieira, F. H. T., & Ling, L. L. (2005b). Predição adaptativa do expoente de Hölder para tráfego multifractal de redes. In *XXVIII Congresso nacional de matemática aplicada e computacional*.

Karasaridis, A. & Hatzinakos, D. (2001). A network heavy traffic modeling using alpha stable self-similar processes. *IEEE Transactions on Communications*, *7*, 1203–1214.

Kelly, F. (1996). *Notes on effective bandwidths*. In Stochastic Networks:. Oxford University Press.

Kelly, F. P. (1991). Effective bandwidth at multi-class queues. *Queue Syst.*, *9*, 5–16.

Keshav, S. (2001). *An engineering approach to computer networking: ATM networks, the internet, and the telephone network*. Boston: Addison-Wesley.

Kesidis, G. (1999). Bandwidth adjustments using on-line packet-level measurements. In *SPIE Conf. Performance and Control of Network Systems*, Boston, MA.

Kesidis, G., Walrand, J., & Chang, C. S. (1993). Effective bandwidths for multiclass markov fluids and other ATM sources. *IEEE/ACM Trans. on Networking*, *1*(3), 424–428.

Kim, E., Park, M., Ji, S., & Park, M. (1997). A new approach to fuzzy modeling. *IEEE Trans. Fuzzy Syst*, *5*, 328–337.

Kim, H. S. & Shroff, N. B. (2001). Loss probability calculations and asymptotic analysis for finite buffer multiplexers. *IEEE/ACM Trans. on Networking*, *9*(6).

Kleinrock, L. (1975). *Queueing systems: VolI: Theory*. John Wiley, Inc.

Knightly, E. (1998). Enforceable quality of service guarantees for bursty traffic streams. In *Proc. IEEE INFOCOM 98* (pp. 635–642). San Francisco.

Knightly, E. & Shroff, N. (1999). Admission control for statistical QoS: Theory and practice. *IEEE Network*, *13*(2), 20–29.

Kolmogorov, A. N. (1962). A refinement of previous hypothses concerning the local structure of turbulence in a viscous incompressible fluid a high Reynolds number. *J.Fluid Mech.*, *13*, 82–85.

Krishna, M. P., Gadre, V. M., & Dessay, U. B. (2003). *Multifractal based network traffic modeling*. Kluwer Academic Publishers.

Krishna, P. M., Gadre, V. M., & Desai, U. B. (2001). Global scaling exponent for variable variance gaussian multiplicative (VVGM) multifractal cascades. In *SPCOM'01* (pp. 19–25).

Kroll, A. (1996). Indentification of funciotnal fuzzy models using multi-dimensional reference fuzzy sets. *Fuzzy sets and Systems*, *80*, 149–158.

Krunz, M. & Ramasamy, A. M. (2000). The correlation structure for a class of scene-based video models and its impact on the dimensioning of video buffers. *IEEE Trans. Multimedia*, *2*, 27–36.

Kurose, J. (1992). On computing per-session performance bounds in high-speed multi-hop computer networks. In *Proc. ACM Sigmetrics*, (pp. 128–139).

Lazar, A., Pacifici, G., & Pendarakis, D. (1994). Modeling video sources for real time scheduling. *ACM multimedia Syst.J.*, *1*, 253–266.

Lee, I. W. C. & Fapojuwo, A. O. (2005). Stochastic processes for computer network traffic modeling. *Computer Communications*, *29*, 1–23.

Leland, W. E., Taqqu, M. S., Willinger, W., & Wilson, D. V. (1994). On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, *2*, 1–15.

Lewis, J. T. & Russell, R. (1998). An introduction to large devaitions for teletraffic engineers. *Dublin Institute for Advanced Studies - Dublin, Ireland*.

Li, C., Burchard, A., & Liebeherr, J. (2003). A network calculus with effective bandwidth. Cs-2003-20, University of Virginia.

Li, Q. & Mills, D. (1999). Investigating the scaling behavior, crossover. In *Proc. of Globecom*, (pp. 1843–1852).

Liebeherr, J. (2000). End-to-end quality of service quarantees. In *IwQoS*.

Liebeherr, J., Patek, S. D., & Burchard, A. (2003). Statistical per-flow service bounds in a network with aggregate provisioning. In *IEEE INFOCOM*, volume 3 (pp. 1680–1690).

Likhanov, L. & Mazumdar, R. R. (1998). Cell loss asymptotics for buffers fed with a large number of independent stationary processes. In *Proc. of IEEE INFOCOM'98*, San Francisco USA.

Lim, L. K., Gao, J., Eugene, T. S., Chandra, P. R., Steenkiste, P., & Zhang, H. (2001). Customizable virtual private network service with QoS. *Computer Networks*, *36*, 137–151.

Low, S. H., Paganini, F., & Doyle, J. C. (2002). Internet congestion control. *IEEE Control Systems Magazine*, *22*(1), 28–43.

LRPRC (2002). Laboratório de reconhecimento de padrões e redes de comunicações (LRPRC) projeto ericson UNI-20. a computational tool and optimization methods for multimedia traffic characterization and effective bandwidth estimation on modern communication networks. Technical report, Unicamp.

Mallat, S. & Hwang, W. (1992). Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, *38*(8), 617–643.

Mandelbrot, B. (1977). *The fractal geometry of nature*. Nova York: W.H.Freeman e Co.

Mandelbrot, B. B., Calvet, L., & Fisher, A. (1997). Large deviations and the distribution of price changes. Discussion paper No 1165 of the Cowles Foundation for Economics at Yale University.

Mandelbrot, B. B., Fisher, A., & Calvet, L. (1997). A multifractal model of asset return. Technical report, Yale University.

Mandelbrot, B. B. & Ness, J. W. V. (1968). Fractional brownian motions, fractional noises and applications. *SIAM Review*, *10*, 422–437.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*, 431–441.

Melo, C. A. V. & da Fonseca, N. L. S. (2005). Envelope process and computation of the equivalent bandwidth of multifractal flows. *Computer Networks*, *48*, 351–375.

Mendel, J. M. (1995). Fuzzy logic systems for engineering: A tutorial. *Proceedings of the IEEE*, *83*(3), 345–377.

Molnár, S., Dang, T. D., & Maricza, I. (2002). On the queue tail asymptotics for general multifractal traffic. In *Em Proc., IFIP Networking'02*, Pisa, Italia.

Molnar, S. & Terdik, G. (2001). A general fractal model of internet traffic. Tampa, Florida. In Proc. IEEE LCN 2001.

Muzy, J. F., Delour, J., & Bacry, E. (2000). Modeling fluctuations of financial time series: from cascade process to stochastic volatility model. *European Physics Journal B*, *17*, 537–548.

Neves, J. E., Leitão, M. J., & Almeida, B. L. (1995). Neural networks in b-isdn flow control: Atm traffic prediction or network modeling. *IEEE Communications Magazine*, *33*(10), 50–56.

Ninness, B. & Gustafsson, F. (1995). Orthonormal bases for system identification. In *Proc. of 3rd European Control Copnference*, (pp. 13–18)., Roma, Itália.

Norros, I. (1994). A storage model with self-similar input. *Queueing Systems*, *16*, 387–396.

Norros, I. (1995). On the use of fractional brownian motion in the theory of connectionless networks. *IEEE Journal on Selected Areas in Communications*, *13*(6), 953–962.

O'Connell, N. (1999). Large deviations with applications to telecommunications. *BRIMS - Hewlett-Packard Labs - Bristol, UK*.

Oliveira, G. H. C., Campello, R. J. G. B., & Amaral, W. C. (1999). Fuzzy models within orthonormal basis function framework. In *IEEE International Fuzzy Systems Conference Proceedings*, Seoul, Korea.

Onvural, R. O. (1995). *An Introduction to Probability Theory and Its Applications*. Asynchronous Transfer Mode Networks - Performance Issues.

Ossiander, M. E. & Waymire, E. C. (2002). Statistiscal estimation theory for multiplicative cascades. *Ann.Statistic.*, *28*, 1533–1560.

Pancha, P. & Eizarki, M. (1994). Variable bit rate video transmission. *IEEE Commun.Mag*, *32*(5), 54–66.

Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes* (3rd ed.). New York: McGraw-Hill.

Parekh, A. & Gallager, R. (1993). A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM Transactions on Networking*, *1*(3), 344–357.

Park, K. & Willinger, W. (2000). *Self-similar Network Traffic and Performance Evaluation*. New York: John Wiley and Sons.

Paxson, V. & Floyd, S. (1995). Wide-area traffic: The failure of poisson modeling. *IEEE/ACM Transactions on Networking*, *3*(3), 226–244.

Pedrycz, W. & Gomide, F. (1998). *An Introduction to Fuzzy Sets: Analysis and Design*. Cambridge, MA: MIT Press.

Peitgen, H., Jurgens, H., & Saupe, D. (1994). *Chaos and Fractals*. Inglaterra: Springer-Verlag.

Peltier, R. & Véhel, J. L. (1995). Multifractional brownian motion. INRIA Research Project 2645.

Perlingeiro, F. R. & Ling, L. L. (2005). Estudo de estimação de banda efetiva para tráfego auto-similar com varância infinita. In *SBrT*, (pp. 326–331)., Campinas.

Pitsillides, A., Ioannou, P., & Rossides, L. (2001). Congestion control for differentiated services using non-linear control theory. In *IEEE 6th Proc. Symp. Comp. and Comm.*, (pp. 726–33).

Postel, J. (1981). Transmission control protocol. RFC 793 (STD 7).

Qiu, B. (1999). Analysis of fuzzy logic and autoregressive video source predictors using t-tests. In *International Symposium on Signal Processing and its Applications (ISSPA)*. Brisbane, Austrália.

Rajagopal, S., Reisslein, M., & Ross, K. W. (1998). Packet multiplexers with adversarial regulated traffic. In *Proc. IEEE INFOCOM 98* (pp. 347–355).

Ramamurthy, G. & Sengupta, B. (1996). A predicitve congestion control policy for broadband integrated wide area networks. *Computer Networks and ISDN Systems*, *28*, 811–834.

Rananand, N. (1996). *Traffic Modeling and Performance Evaluation for ATM Networks: Short and Long Range Dependent Models*. Tese de doutorado, University of Maryland.

Reisslein, M., Ross, K. W., & Rajagopal, S. (2002). A framework for guaranteeing statistical qos. *IEEE/ACM Transactions on Networking*, *19*(1).

Resnik, S., Samorodnitsky, G., Gilbert, A., & Willinger, W. (2003). Wavelet analysis of conservative cascades. *Bernoulli*, *9*(1), 97–135.

Ribeiro, V. J., Riedi, R. H., Crouse, M. S., & Baraniuk, R. G. (2000). Multiscale queueing analysis of long-range dependent traffic. In *IEEE INFOCOM*, (pp. 1026–1035)., Tel Aviv, Israel.

Riedi, R. (2003). Multifractal processes. In G. O. P. Doukhan & M. S. Taqqu (Eds.), *Theory and Aplications of Long-range Dependence*. Boston, MA.

Riedi, R. H. (1997). Introduction to multifractals. Technical report, Rice University Department of ECE, Houston, TX, USA.

Riedi, R. H., Crouse, M. S., Ribeiro, V. J., & Baraniuk, R. G. (1999). A multifractal wavelet model with application to network traffic. *IEEE Trans. on Information Theory*, *45*(3).

Riedi, R. H. & Véhel, J. L. (1997). Tcp traffic is multifractal: a numerical study. Technical Report 3129, INRIA Research report.

Roberts, J. W. (2000). Engineering for quality of service. In *Self-similar network traffic and performance evaluation*. John Wiley and Sons.

Rolls, D. A., Michailidis, G., & Hernández-Campos, F. (2005). Queueing analysis of network traiffc: methodology and visualization tools. *Computer Networks*, *48*, 447–473.

Ross, S. M. (1989). *Introduction to Probability Models*. Academic Press Inc.

Ross, T. J. (1997). *Fuzzy Logic with Engineering Applications*. McGraw-Hill International Editions.

Roughan, M. & Veitch, D. (1999). Measuring long-range dependence under changing traffic conditions. In *Proceedings of the INFOCOM* (pp. 1513–1521).

Ryu, S. (2003). Advances in internet congestion control. *IEEE Communications Surveys*, *5*(1), 28–39.

Samorodnistiky, G. & Taqqu, M. (1994). *Stable Non Gaussain Processes: Stochastic Models with Infinite Variance*. Chapman and Hall.

Sang, A. & Li, S. (2002). A predictability analysis of network traffic. *Computer networks*, *39*, 329–345.

Seuret, S. & Gilbert, A. C. (2000). Pointwise hölder exponent estimation in data network traffic. In *ITC Specialist Semina*, Monterey.

Shiomoto, K., Yamanaka, N., & Takahashi, T. (1999). Overview of measurement-based connection admission control in ATM networks. *IEEE Comm. Surveys*, 2–13.

Shroff, N. & Schwartz, M. (1998). Improved loss calculations at an ATM multiplexer. Technical report, Sch. Elec. Comput. Eng., Purdue Univ., West Lafayette,.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Siripongwutikorn, P., Banerjee, S., & Tipper, D. (2002). Adaptive bandwidth control for efficient aggregate QoS provisioning. In *IEEE Globecom'02*, Taipen, Taiwan.

Siris, V. A. (2000). Large deviation techniques for traffic engineering. www.ics.forth.gr/netgroup/msa.

Solo, V. (1996). Adaptive estimation of effective bandwidth in ATM networks. In *Proc. 35th IEEE CDC*.

Sousa, L. M. C., Vieira, F. H., & Lee, L. L. (2006a). A fuzzy approach for adaptive control of MPLS network traffic flows. In *International Telecommunications Simposium - ITS*, Fortaleza, CE.

Sousa, L. M. C., Vieira, F. H. T., & Lee, L. L. (2006b). Controle fuzzy adaptativo de fluxos de tráfego. In *Congresso Brasileiro de Automática*, Salvador, Bahia, Brasil.

Stark, H. & Woods, J. W. (1994). *Probability, Random Processes, and Estimation Theory for Engineers*. Prentice Hall.

Struzik, Z. R. (2000). Determining local singularity strengths and their spectra with the wavelet transform. *Fractals*, *8*(1), 163–179.

Sugeno, M. & Yasukawa, T. (1993). A fuzzy-logic-based approach to qualitative modeling. *IEEE Trans. Fuzzy Syst*, *1*, 7–31.

Takagi, T. & Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Syst., Man, Cybern.*, *15*, 116–132.

Takens, F. (1981). *Detecting Strange Attractors in Turbulance. Dynamical Systems and Turbulance*, (pp. 366–381). Lectures Notes of Mathematics. Springer-Verlag.

Tartarelli, S., Falkner, M., Devetsikiotis, M., Lambadaris, I., & Giordano, S. (2000). Empirical effective bandwidths. In *Proc. of IEEE Globecom'00*. San Francisco.

Tran, H. T. & Ziegler, T. (2005). Adaptive bandwidth provisioning with explicit respect to QoS requirements. *Computer Communications*, *28*, 1862–1876.

Tuan, T. & Park, K. (1998). Congestion control for self-similar network traffic. Csd-tr 98-014, Dept.of Comp. Science, Purdue Univ.

Véhel, J. L. & Riedi, R. (1997). *Fractional Brownian motion and data traffic modeling: The other end of the spectrum*. Fractals in Engineering. Springer.

Veitch, D. & Abry, P. (1999). A wavelet based joint estimator for the parameters of LRD. *IEEE Trans. Info. Theory.*, *45*(3).

Veitch, D., Hohn, N., & Abry, P. (2005). Multifracatlity in TCP/IP traffic: The case against. In *Computer Networks*, (pp. 293–313). 48.

Vieira, F. H. T., Bianchi, G. R., Ling, L. L., & Lemos, R. P. (2004a). Estimação de banda efetiva dinâmica em redes de computadores utilizando uma modelagem auto-regressiva nebulosa. In *XXI Simpósio Brasileiro de Telecomunicações (SBrT)*, Belém, Pará.

Vieira, F. H. T., Bianchi, G. R., Ling, L. L., & Lemos, R. P. (2004b). Fuzzy-AR modeling for dynamic effective bandwidth estimation in high-speed networks. In *WSEAS Transactions on Systems*, (pp. 2680–2685).

Vieira, F. H. T. & Lee, L. (2004a). Fuzzy modeling and prediction with confidence bound estimation for traffic rate allocation in high-speed networks. São Luís: Simpósio Brasileiro de Redes Neurais.

Vieira, F. H. T. & Lee, L. L. (2004b). Uma nova arquitetura neural combinada utilizando teoria de ressonância adaptativa e algoritmo de Kalman estendido para alocação dinâmica de taxa de transmissão em redes de computadores. In *Simpósio Brasileiro de Redes Neurais- SBRN*, São Luís, Maranhão, Brasil.

Vieira, F. H. T. & Lee, L. L. (2004c). Uma nova arquitetura neural combinada utilizando teoria de ressonância adaptativa e aprendizagem RTRL para predição e controle de tráfego de dados em tempo real. In *Congresso Brasileiro de Automática (CBA)*, Gramado, RS.

Vieira, F. H. T., Lemos, R. P., & Lee, L. (2003a). Alocação dinâmica de taxa de transmissão em redes de pacotes utilizando redes neurais recorrentes treinadas com algoritmos em tempo real. *IEEE Latin America*, *1*(1).

Vieira, F. H. T., Lemos, R. P., & Lee, L. L. (2003b). Aplicação de redes neurais RBF treinadas com algoritmo ROLS and análise wavelet na predição de tráfego em redes ethernet. In *VI Congresso Brasileiro de Redes Neurais*, São Paulo, SP.

Vieira, F. H. T. & Ling, L. L. (2006a). Análise de fila para tráfego multifractal utilizando cálculo de rede e parâmetro de escala global. In *Simpósio Brasileiro de Redes de Computadores*, Curitiba-PR.

Vieira, F. H. T. & Ling, L. L. (2006b). Multifractal traffic modeling using a multiplicative cascade with generalized multiplier distributions. In *International Conference on Communications -ICC'06*, Istanbul, Turquia.

Vieira, F. H. T. & Ling, L. L. (2006c). Queueing analysis for multifractal traffic through network calculus and global scaling parameter. In *International Telecommunications Simposium - ITS*, Fortaleza CE.

Vijayan, L., Chakrabarti, S., Petr, D. W., & Khan, S. (2002). Extensions to multifractal wavelet model for synthesing network traffic. In *IEEE International Conference on Communications ICC'02*, volume 4 (pp. 2400–2404).

Wahlberg, B. & Ljung, L. (1992). Hard frequency-domain model error bounds from least-squares like identification techniques. *IEEE Trans. on Automatic Control*, *37*(7), 900–912.

Wang, L. X. & Mendel, J. M. (1992). Fuzzy basis functions, universal aproximation and orthogonal least squares learning. *IEEE Transactions on Neural Networks*, *3*, 807–814.

Wang, W., D.Tipper, & S.Banerjee (1996). A simple approximation for modeling nonstationary queues. (pp. 255–262). IEEE INFOCOM.

Wang, Y. & Zhu, Q. (1998). Error control and concealment for video communication: A review. *Proc. IEEE*, *86*, 974–997.

Waymire, E. & Williams, S. (1995). Multiplicative cascades: Dimension spectra and dependence. *Jour. Fourier Anal. and Appl.*, 589–609.

Weiss, A. (1995). An introduction to large deviations for communication networks. *IEEE Journal on Selected Areas in Communications*, *13*(6).

West, B. J., Scafeta, N., Cooke, W. H., & Balocchi, R. (2004). Influence of progressive central hypovolemia on hölder exponent distributions of cardiac interbeat intervals. *Annals of biomedical engineering*, *32*(8), 1077–1087.

Yager, R. R. & Filev, D. P. (1994). *Essentials of Fuzzy Modeling*. John Wiley and Sons.

Young, P. (1984). *Recursive Estimation and Time Series Analysis: An Introduction*. Communications and Control Engineering Series. New York: Springer-Verlag.

Yuang, M. C. (1997). Intelligent video smoother for multimedia communications. *IEEE J. Select. Areas Commun.*, *15*(2), 136–146.

Yule, G. (1927). On a method of investigating periodicity in disturbed series with special reference to wofer's sunspot numbers. Phil.Trans.Roy.Soc.London.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338–353.

Zhang, Q., Wu, J., & H.Xi. Dynamic bandwidth allocation over ATM based neural network prediction: analysis and simulation. Technical report, School of Information Science and Technology University of Science and Technology of China.

Zhang, Z. L., Ribeiro, V., Moon, S., & Diot, C. (2003). Small-time scaling behaviors of internet backbone traffic: an empirical study. In *IEEE Infocom*, San Francisco.

Apêndice A

Análise Wavelet

A análise *wavelet* tem sido utilizada para estudar as propriedades de dependência em escala dos dados diretamente através dos coeficientes da decomposição escala-tempo *wavelet*. Quando há alguma evidência de longa-dependência (LRD), a análise *wavelet* é capaz de oferecer estimativas não-polarizadas e que podem ser implementadas usando técnicas da análise de multiresolução.

A.1 Análise de Multiresolução e Transformada Wavelet Discreta

A análise de multiresolução (Multiresolution Analysis - (MRA)) é formulada com base em subespaços $\{\nu_j\}_{j\in\mathbb{Z}}$, satisfazendo as seguintes propriedades (Abry & Veitch, 1998):

1. $\underset{j\in\mathbb{Z}}{\cap}\nu_j = \{0\}$, $\underset{j\in\mathbb{Z}}{\cup}\nu_j$ em $\mathbb{L}^2$ (espaço de funções integráveis quadráticas);

2. $\nu_j \subset \nu_{j-1}$;

3. $x(t) \in \nu_j \longleftrightarrow x(2^j t) \in \nu_0$;

4. Existe uma função $\varphi_0(t)$ em ν_0, denominada de função escala, tal que $\{\varphi_0(t-k), k\in\mathbb{Z}\}$ é uma base para ν_0. De forma análoga, o conjunto de funções transladadas no tempo e modificadas em escala:

$$\left\{\varphi_{j,k}(t) = 2^{-j/2}\varphi_0\left(2^{-j}t - k\right), k\in\mathbb{Z}\right\} \tag{A.1}$$

constituem uma base para o espaço vetorial ν_j. Portanto, o espaço que contém sinais de mais alta resolução também contém aqueles de menor resolução.

Pode-se pensar na análise de multiresolução de um sinal $x(t)$ como uma sucessiva projeção do sinal em cada subespaço ν_j. Assim, estes subespaços são vistos como subespaços de aproximação, ou seja,

$$approx_j(t) = \left(proj_{v_j} x\right)(t) = \sum_k a_x(j,k)\,\varphi_{j,k}(t). \tag{A.2}$$

Uma vez que $\nu_j \subset \nu_{j-1}$, $approx_j(t)$ é uma aproximação mais grosseira do sinal $x(t)$ do que, por exemplo, $approx_{j-1}$. Portanto, a idéia chave da MRA consiste em examinar as diferenças, ou seja, os detalhes entre os subespaços de aproximação varridos por várias escalas da função escala, que podem ser expressos por:

$$detail_j(t) = approx_j(t) - approx_{j-1}(t). \tag{A.3}$$

As funções que descrevem as diferenças entre subespaços ν_j são as *wavelets* $\psi_{j,k}(t)$. Como resultado, os subespaços *wavelets* ω_j podem ser definidos de tal forma que:

$$\nu_2 = \upsilon_3 \oplus \omega_3$$
$$\nu_1 = \upsilon_2 \oplus \omega_2$$
$$\nu_0 = \upsilon_1 \oplus \omega_1$$

$$\nu_0 = \upsilon_3 \oplus \omega_3 \oplus \omega_2 \oplus \omega_1 \tag{A.4}$$

Em geral, isso resulta em $\mathbb{L}^2 = \upsilon_j \oplus \omega_j \oplus \omega_{j-1} \oplus \omega_{j-2} \cdots \oplus \omega_1$. Como conseqüência da equação $(A.4)$ tem-se que o sinal de detalhes $detail_j$ pode se obtido diretamente pelas projeções de x em subespaços *wavelets*, já que estes subespaços residem no espaço varrido pelas funções escala. A análise de multiresolução mostra que existe uma função ψ_0, denominada de *wavelet*-mãe, que é derivada de φ_0 tal que o conjunto de funções $\left\{\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi_0\left(2^{-j}t - k\right), k \in \mathbb{Z}\right\}$ constitui uma base para os subespaços ω_j e assim:

$$detail_j(t) = \left(proj_{\omega_j} x\right)(t) = \sum_k d_x(j,k)\,\psi_{j,k}(t). \tag{A.5}$$

Basicamente, como pode-se notar pela equação $(A.4)$, a análise de multiresolução (MRA) permite reescrever a informação x como uma coleção de detalhes em diferentes resoluções e uma aproximação

de mais baixa resolução da seguinte forma:

$$x(t) = approx_J(t) + \sum_{j=1}^{j=J} detail_j(t) \tag{A.6}$$

$$= \sum_k a_x(J,k)\,\varphi_{J,k}(t) + \sum_{j=1}^{J}\sum_k d_x(j,k)\,\psi_{j,k}(t). \tag{A.7}$$

A aproximação mais grosseira $approx_j(t)$ de um sinal x significa que φ_0 pode ser visto como um filtro passa-baixas. Os detalhes do sinal na escala j representado por $detail_j(t)$ indica que ψ_0 tem função de filtro passa-faixa. Dada uma função escala φ_0 e uma *wavelet*-mãe ψ_0 a transformada *wavelet* discreta consiste de um mapeamento de $\mathbb{L}^2(\mathbb{R}) \rightarrow \mathbb{L}^2(\mathbb{Z})$ dado por

$$x(t) \rightarrow \{\{a_x(J,k), k \in \mathbb{Z}\}\{d_x(j,k), j = 1,\ldots,J, k \in \mathbb{Z}\}\}. \tag{A.8}$$

Os coeficientes a_x e d_x podem ser calculados através dos produtos internos de x com dois conjuntos de funções:

$$\left.\begin{array}{l} a_x(j,k) = \langle x, {\varphi^\circ}_{j,k}\rangle \\ d_x(j,k) = \langle x, {\psi^\circ}_{j,k}\rangle \end{array}\right\} \tag{A.9}$$

onde ${\psi^\circ}_{j,k}$ e ${\varphi^\circ}_{j,k}$ são versões dilatadas e transladadas de $\psi_{j,k}$ e $\varphi_{j,k}$. Estes coeficientes podem ser facilmente calculados usando banco de filtros como é mostrado em (Burrus et al., 1998).

Apêndice B

Estimação do Parâmetro de Hurst: Método usando Wavelets

Esta seção apresenta um estimador computacionalmente eficiente para o parâmetro de Hurst. Os métodos paramétricos requerem a escolha de um modelo de tráfego a priori, sendo por isso, de difícil implementação na prática para grande volume de dados devido a alta complexidade computacional e necessidade de armazenamento.

A dependência a longo prazo está relacionada com o decaimento em lei de potência da função de autocorrelação de um processo estacionário. Um afirmação equivalente e que se refere a densidade espectral $f_x(\lambda)$ de x é dada a seguir:

Definição B.0.1 *Seja X_t um processo estacionário para o qual existe um número real $\beta \in (0,1)$ e uma constante $c_f > 0$ tal que*

$$\lim_{\lambda \to 0} f(\lambda) / \left[c_f |\lambda|^{-\beta} \right] = 1. \tag{B.1}$$

Então X_t é chamado de processo estacionário com longa memória ou longa dependência.

Ao se escrever o sinal X_t de acordo com a equação $(A.6)$, pode-se interpretar o coeficiente $|d_x(j,k)|^2$ como uma medida da quantidade de energia do sinal analisado no instante de tempo $2^j k$ e freqüência $2^{-j}\lambda_0$, onde λ_0 é uma freqüência arbitrária selecionada pela escolha de ψ_0. Em (Abry & Veitch, 1998), os autores sugerem que um estimador prático para a densidade espectral de x pode ser realizado efetuando-se uma média temporal de $|d_x(j,k)|^2$ em um dada escala, ou seja:

$$\widehat{f}_x\left(2^{-j}\lambda_0\right) = \frac{1}{n_j} \sum_k |d_x(j,k)|^2, \tag{B.2}$$

onde n_j é o número de coeficientes *wavelet* disponíveis na oitava j, em que, $n_j = 2^{-j}n$, e n é o número de amostras do processo. Portanto, $\widehat{f}_x(\lambda)$ é uma medida da quantidade de energia em uma

dada largura de banda em torno de uma freqüência λ e pode ser considerada uma estimativa estatística para o espectro $f_x(\lambda)$ de x. De fato, pode-se demonstrar que quando x é um processo estacionário em sentido amplo a esperança matemática de $\widehat{f_x}(\lambda)$ é

$$E\widehat{f_x}\left(2^{-j}\lambda_0\right) = \int f_x(\lambda)\, 2^j \left|\psi_0\left(2^j\lambda\right)\right|^2 \, d\lambda \tag{B.3}$$

Da relação acima, nota-se que $\widehat{f_x}$ sofre de polarização de convolução padrão, pois a densidade espectral a ser estimada se mistura com a gama de frequencias na janela analisada na escala j. O importante fato neste ponto é que para sinal com longa dependência (LRD) a polarização se reduz a um forma simples. Em outras palavras, a equação $(B.3)$ pode ser reescrita da seguinte forma:

$$E\widehat{f_x}\left(2^{-j}\lambda_0\right) = \begin{cases} c_f \left|2^{-j}\right|^{(1-2H)} \int \lambda^{(1-2H)} \left|\psi_0(\lambda)\right|^2 \, d\lambda \\ f_x\left(2^{-j}\lambda_0\right) 2^{-j} \lambda_0^{(1-2H)} \int \left|\psi_0\left(2^j\lambda\right)\right|^2 \, d\lambda \end{cases} \tag{B.4}$$

Portanto, é possível se implementar um estimador $\hat{H}$ para o parâmetro de Hurst H através de uma simples regressão linear de $\log_2\left(\widehat{f_x}\left(2^{-j}\lambda_0\right)\right)$ em j:

$$\log_2\left(\frac{1}{n_j}\sum_k \left|d_x(j,k)\right|^2\right) = \left(2\hat{H}-1\right) j + \hat{c} \tag{B.5}$$

É sabido que na presença de LRD, o estimador amostral clássico $\frac{1}{n}\sum_{t=1}^{n} X_t^2$ para estatísticas de segunda ordem como a variância do processo X_t tem propriedades estatísticas fracas, porque uma média temporal é efetuada em dados altamente correlacionados. Quando os dados são representados através de coeficientes *wavelets*, a estrutura de correlação resultante não é de longa dependência, assim é possível se estimar com precisão o parâmetro de Hurst H (Abry & Veitch, 1998).

Apêndice C

Estimação Não-Paramétrica de Distribuição de Probabilidade: Método de Kernel

A função de densidade de probabilidade é um conceito fundamental em estatística. Estimação de densidade de probabilidade a partir de dados reais pode ser feita de forma paramétrica, ou seja, assumindo que os dados são oriundos de uma distribuição conhecida. Ou de forma não-paramétrica que é descrita nesta seção, a qual foi utilizada neste trabalho a fim de encontrar uma distribuição mais apropriada para os multiplicadores da cascata conforme discutido na seção 2.5.3 do Capítulo 2.

O método de Kernel para estimação da função distribuição de probabilidade possui larga aplicação e suas propriedades são bem conhecidas. A não ser pelos histogramas, a estimação por Kernel é provavelmente a mais usada e certamente a mais estudada. Sejam as amostras $X_i, \ldots, X_n$ obtidas com uma distribuição de probabilidade contínua *f(x)* a qual se quer estimar. O método de Kernel usa uma soma de funções localizadas nas observações para as quais se obterá uma distribuição de probabilidade. Um estimador com Kernel *K* é definido por:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{C.1}$$

onde *h* é a largura da janela, também chamada de parâmetro de suavização. Neste método, determina-se o formato do Kernel *K* e sua largura *h*. Deve ser notado que se um valor de *h* muito pequeno é escolhido, pode resultar numa distribuição com muitos 'espúrios', enquanto que se *h* é demasiado grande então uma suavização muito intensa pode ocorrer.

A função Kernel *K* deve satisfazer a seguinte condição:

$$\int_{-\infty}^{\infty} K(x)\mathrm{d}x = 1 \tag{C.2}$$

Ao se considerar a discrepância entre a densidade de probabilidade estimada $\hat{f}(x)$ e a real *f(x)*, uma medida comum de ser empregada é o erro quadrático integral médio (EQIM) dado por:

$$EQM(\hat{f}) = \int_{-\infty}^{\infty} E\{\hat{f}(x) - f(x)\}^2 \mathrm{d}x + \int_{-\infty}^{\infty} var\hat{f}(x)\mathrm{d}x \tag{C.3}$$

O valor ideal para a largura da janela *h* do ponto de vista da minimização do erro quadrático integral médio (EQIM) pode ser dado por (Silverman, 1986):

$$h_{ot} = K_2^{\frac{-2}{5}} \left\{ \int_{-\infty}^{\infty} K(t)^2 \mathrm{d}x \right\}^{\frac{1}{5}} \left\{ \int_{-\infty}^{\infty} f''(x)^2 \mathrm{d}x \right\}^{\frac{-1}{5}} n^{\frac{-1}{5}} \tag{C.4}$$

Na seção 2.5.3 do Capítulo 2, utilizou-se o Kernel gaussiano. Neste caso, o valor ótimo para a largura da janela *h* é dado por (Silverman, 1986):

$$h_{ot} = \frac{4}{3}^{\frac{1}{5}} \sigma n^{\frac{-1}{5}} = 1.06\sigma n^{\frac{-1}{5}} \tag{C.5}$$

Apêndice D

Algoritmo de Levenberg-Marquardt

O algoritmo de Levenberg-Marquardt é um algoritmo geral de minimização linear para o caso onde as derivadas da função objetivo são conhecidas. Este algoritmo mistura dinamicamente iterações dos métodos de Gauss-Newton e Gradiente Descendente, sendo o algoritmo de otimização mais usado, pois possui melhor desempenho do que os métodos de gradiente descendente e outros métodos de gradiente conjugado. O seu objetivo é prover solução para o problema da minimização por mínimos quadrados não-linear. Seja o vetor $\vec{x}$ que representa os parâmetros desconhecidos tal que:

$$z(j) = h(j; \vec{x}) + r(j), \qquad j = 1, \ldots, k \tag{D.1}$$

onde $h(j)$ é a função medida e $r(j)$ corresponde ao erro, ou resíduo. A função a ser minimizada tem a forma:

$$f(x) = \frac{1}{2} \sum_{j=1}^{m} r_j^2(x) \tag{D.2}$$

Escrevendo em forma matricial o vetor de erro $\vec{r}(x) = (r_1(x), r_2(x), \ldots, r_m(x))$, pode-se representar f como $f(x) = \frac{1}{2}||\vec{r}(x)||^2$. As derivadas de f podem ser escritas usando a matriz Jacobiana J de r definida como $J(x) = \frac{\partial r_j}{\partial x_i}$, $1 \leq j \leq m$, $1 \leq i \leq n$. O mínimo da função f é encontrado fazendo o gradiente de f igual a zero, i.e., $\nabla f(x) = 0$. O gradiente da função $f(x)$ assim como a Hessiana ($\nabla^2 f(x)$) podem ser dados em função da Jacobiana de $r_j(x)$:

$$\nabla f(x) = \sum_{j=1}^{m} r_j(x) \nabla r_j(x) = J(x)^T \vec{r}(x) \tag{D.3}$$

$$\nabla^2 f(x) = J(x)^T J(x) + \sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x) \tag{D.4}$$

Neste caso, demonstra-se que a matriz Hessiana pode ser aproximada por $\nabla^2 f(x) = J(x)^T J(x)$ para resíduos pequenos. Levenberg propôs uma combinação entre os métodos de gradiente descendente e de Newton pra resolver a equação $\nabla f(x) = 0$, resultando na seguinte equação de atualização:

$$x_{i+1} = x_i - (H + \lambda diag[H])^{-1} \nabla f(x_i) \tag{D.5}$$

onde H é a matriz Hessiana avaliada em x_i. Em síntese, o algoritmo de Levengerg-Marquardt consiste de:

Algoritmo D.0.1 *Algoritmo de Levenberg-Marquardt*

1. *Atualize a estimativa dos parâmetros usando a equação D.5.*
2. *Calcule o erro para o novo vetor de parâmetros.*
3. *Se o erro aumentou como resultado da atualização, reinicie os pesos com seus valores prévios e aumente λ de um fator de 10 ou outro fator significativo, senão vá para o passo 4. Então vá pra o passo 1 e tente uma nova atualização.*
4. *Se o erro diminuiu como resultado da atualização, mantenha os pesos com seus novos valores e diminua λ por exemplo de um fator de 10.*

O algoritmo de Levenberg-Marquardt tem taxa de convergência alta por incorporar o método de Newton. Uma vez que a matriz Hessiana é proporcional à curvatura da função $f(x)$, a equação (D.5) implica em passos mais largos na direção onde o gradiente é menor. Dessa forma, o problema clássico de 'vale de erro', não ocorre. Não se pode dizer que o algoritmo de Levenberg-Marquardt seja sempre ótimo mas é uma heurística que funciona extremamente bem na prática. A inversão matricial envolvida na equação (D.5) é geralmente implementada usando métodos de obtenção de pseudo-inversas, como decomposição em valores singulares.

Apêndice E

Algoritmo de Levinson-Durbin

Consideremos o problema de prever os valores X_{n+h}, $h > 0$, de uma série temporal estacionária com média conhecida μ e autocovariância γ, em termos dos valores $\{X_n, ..., X_1\}$ até um instante de tempo n. O objetivo aqui é achar a combinação linear de $1, X_n, X_{n-1}..., X_1$, para se fazer a predição de X_{n+h} com erro quadrático médio mínimo. O melhor preditor linear em termos de $1, X_n, X_{n-1}..., X_1$ será denotado por P_nX_{n+h}, tendo a seguinte forma:

$$P_nX_{n+h} = a_0 + a_1X_n + ... + a_nX_1 \tag{E.1}$$

Propriedades de P_nX_{n+h}

1. $P_nX_{n+h} = \mu + \sum_{i=1}^{n} a_i \left(X_{n+1-i} - \mu\right)$, onde $\mathbf{a}_n = (a_1, a_2, ..., a_n)'$ satisfaz $\Gamma_n\mathbf{a}_n = \gamma_n(h)$, onde $\Gamma_n = \left[\gamma(i-j)\right]_{i,j=1}^{n}$ é a matriz de covariância n-dimensional,
2. $E\left(X_{n+h} - P_nX_{n+h}\right)^2 = \gamma(0) - \mathbf{a}_n'\gamma_n(h)$, onde $\gamma_n(h) = \left(\gamma(h), ..., \gamma(h+n-1)\right)$,
3. $E\left(X_{n+h} - P_nX_{n+h}\right) = 0$,
4. $E\left(\left(X_{n+h} - P_nX_{n+h}\right)X_j\right) = 0$, $j = 1, ..., n$.

A determinação do melhor estimador linear P_nX_{n+h} de X_{n+h} em termos de $(X_n, X_{n-1}, ..., X_1)$ requer a solução de um sistema de n equações lineares, a qual para valor grande de n pode ser difícil e demorada de ser encontrada. O algoritmo de Levinson-Durbin consiste de um preditor a um passo que se baseia em $n+1$ observações prévias, ou seja,

$$P_nX_{n+1} = \phi_n'\mathbf{X}_n = \phi_{n1}X_n + ... + \phi_{nn}X_1 \tag{E.2}$$

onde $\phi_n = \Gamma_n^{-1}\gamma_n$, com $\gamma_n = \left(\gamma(1), ..., \gamma(n)\right)$ e o correspondente erro quadrático médio é:

$$v_n \stackrel{\Delta}{=} E\left(X_{n+1} - P_n X_{n+1}\right)^2 = \gamma\left(0\right) - \phi_n' \gamma_n \tag{E.3}$$

Assim sendo, o algoritmo de Levinson-Durbin é apresentado a seguir (Brockwell & Davis, 1996).

Algoritmo E.0.2 *Os coeficientes* $\phi_{n1}, ..., \phi_{nn}$ *podem ser calculados recursivamente pelas equações:*

$$\phi_{nn} = \left[\gamma\left(n\right) - \sum_{j=1}^{n-1} \phi_{n-1,j} \gamma\left(n-j\right)\right] v_{n-1}^{-1} \tag{E.4}$$

$$\begin{bmatrix} \phi_{n1} \\ \vdots \\ \phi_{n,n-1} \end{bmatrix} = \begin{bmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{bmatrix} - \phi_{nn} \begin{bmatrix} \phi_{n-1,n-1} \\ \vdots \\ \phi_{n-1,1} \end{bmatrix} \tag{E.5}$$

$$v_n = v_{n-1}\left[1 - \phi_{nn}^2\right] \tag{E.6}$$

onde $\phi_{11} = \gamma\left(1\right)/\gamma\left(0\right)$ *e* $v_0 = \gamma\left(0\right)$.