PREVISÃO DE ESTATÍSTICAS DE PERDAS DE PACOTES USANDO MODELOS DE MARKOV OCULTOS

Fernando Jorge Silveira Filho

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

Prof. Edmundo Albuquerque de Souza e Silva, Ph.D.

Prof.ª Rosa Maria Meri Leão, Dr.

Prof. Valmir Carneiro Barbosa, Ph.D.

Prof. Virgílio Augusto Fernandes Almeida, Ph.D.

Prof. Donald Fred Towsley, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

MAIO DE 2006

SILVEIRA FILHO, FERNANDO JORGE

Previsão de estatísticas de perdas de pacotes usando modelos de Markov ocultos [Rio de Janeiro] 2006

XII, 71 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 2006)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Desempenho de redes de comunicação
2. Perdas de pacotes
3. Previsão de séries temporais
4. Modelos de Markov ocultos

I. COPPE/UFRJ II. Título (Série)

Agradecimentos

Esta dissertação é o resultado final de uma jornada longa e, às vezes, incerta. Muitas dúvidas surgiram nos últimos sete anos. Diante delas, o Professor Edmundo tem sido mais que apenas um orientador acadêmico. Com certeza, nunca deixarei de ser um de seus alunos. Mesmo que eu assim quisesse, ele, provavelmente, não o permitiria. Juntamente com a Professora Rosa, o lado mais sensato da dupla de professores do LAND, o Edmundo nunca mede esforços para que os seus alunos façam bons trabalhos. Sou muito grato a ambos, por cada um destes sete anos.

Se Edmundo e Rosa são o cérebro do grupo, então Carol é certamente o nosso coração. Comigo, o seu carinho e atenção sempre foram motivos para uma admiração sem limites de minha parte. Com muita atenção, Carol revisou a maior parte do texto desta dissertação e, além disso, me auxiliou em todos os detalhes burocráticos para a apresentação do meu trabalho. Onde quer que eu esteja, sempre guardarei um pedaço da minha atenção para essa flor de pessoa.

Algumas pessoas contribuíram muito, com discussões que, de uma forma ou de outra, tiveram algum impacto no conteúdo deste trabalho. Diana pode não ter sido uma dessas pessoas, mas com certeza, as tardes de sábado e domingo, perdidas no LAND para finalizar este texto, teriam sido muito entediantes, não fosse a sua presença, sempre animada, enquanto trabalhava na sua própria dissertação de mestrado.

Para dar o crédito a quem o merece, não posso deixar de mencionar o Flávio. Não seria nenhum exagero dizer que, se não fosse o seu trabalho anterior como

aluno de mestrado, esta dissertação não existiria na forma ou no conteúdo que tem hoje. Edson também foi extremamente generoso, desviando-se dos seus afazeres para discutir e até mesmo escrever comigo um artigo relacionado a esta dissertação. Por último, mas não menos importantes, David e Hugo também fizeram contribuições valiosas, tanto na implementação dos experimentos apresentados no Capítulo 5, quanto em uma discussão, onde chegamos a uma resposta para fechar a recursão da Seção 3.2.

Finalmente, gostaria de agradecer aos meus pais, por serem tão pacientes diante da minha insistência pelo caminho da pesquisa. Apesar de longo e incerto, provavelmente, é este o caminho no qual me manterei durante os próximos anos, e é bom saber que posso continuar contando com o seu carinho e compreensão.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

PREVISÃO DE ESTATÍSTICAS DE PERDAS DE PACOTES USANDO MODELOS DE MARKOV OCULTOS

Fernando Jorge Silveira Filho

Maio/2006

Orientador: Edmundo Albuquerque de Souza e Silva

Programa: Engenharia de Sistemas e Computação

Uma grande quantidade de aplicações de redes multimídia pode se beneficiar da estimativa de estatísticas de perdas de pacotes. Por exemplo, suponha que as características do processo de perdas em um caminho fim-a-fim possam ser bem aproximadas de antemão. Então, uma aplicação de transmissão de áudio ou vídeo em tempo real poderia adaptar sua taxa de transmissão e escolher a estratégia apropriada para recuperação de perdas de pacotes a fim de entregar os dados com uma qualidade aceitável. Mecanismos adaptativos para tais aplicações freqüentemente dependem de modelos de perda de pacotes. Estes devem ser suficientemente precisos para capturar as medidas de perda relevantes e ainda simples o bastante para serem usados em um protocolo de tempo real. A maior parte da pesquisa na literatura propõe modelos com o objetivo de aproximar descritores de perda do canal no longo prazo, sem considerar o processo em escalas de tempo curtas. Uma vez que fenômenos não estacionários podem ter um grande impacto nas estatísticas locais do caminho, uma aplicação que se concentra em escalas de tempo baseada apenas em médias de longo prazo pode realizar decisões ruins para o controle de curto prazo. Neste trabalho, avaliamos diferentes modelos de Markov ocultos como preditores de estatísticas de perda de curto prazo. Propomos um algoritmo para estimar perdas num futuro próximo baseado em medidas do passado recente e comparamos a acurácia de diferentes modelos utilizando este algoritmo.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

PREDICTING PACKET LOSS STATISTICS WITH
HIDDEN MARKOV MODELS

Fernando Jorge Silveira Filho

May/2006

Advisor: Edmundo Albuquerque de Souza e Silva

Department: Systems Engineering and Computer Science

There is a number of applications that can benefit from the estimation of packet loss statistics. For instance, suppose that the loss process characteristics in an end-to-end path can be well approximated in advance. Then, a real time audio or video streaming application could adapt its transmission rate and choose the appropriate packet loss recovery strategy in order to deliver data with an acceptable quality. Adaptive mechanisms for such applications often rely on packet loss models. These should be sufficiently accurate to capture the relevant loss measures and yet simple enough to be used in a real time protocol. Most research in the literature proposes models with the objective of fitting long-term loss descriptors of the channel under study, without considering the process in short time scales. Since non-stationarities can have a major impact on local path statistics, an application that focuses on time scales based only on long-term averages may make poor short-term control decisions. In this work, we evaluate different hidden Markov chain based models as predictors of short-term loss statistics. We propose an adaptive algorithm to estimate near future losses based on recent measurements and compare the accuracy of different underlying models.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Modelos de perda de pacotes desempenham um papel essencial na análise de redes de computadores. Estudos de avaliação de desempenho freqüentemente abstraem as características de perda e retardo de um caminho ou rede com um único modelo analítico. Idealmente, este modelo deve ser capaz de representar as características importantes do caminho e reproduzir, de maneira adequada, o impacto das características da rede no protocolo sendo estudado, mantendo a complexidade baixa.

Da mesma forma, tem crescido o interesse da comunidade de pesquisa por protocolos adaptativos para tarefas como *path-switching*, controle de taxa para tráfego multimídia, e recuperação de perdas de pacotes, para citar alguns exemplos. Tais mecanismos devem ser capazes de inferir perdas de pacotes futuras e ajustar seus comportamentos para lidar com mudanças nas condições da rede, [Bolot et al. 1999], [Duarte 2003], [Karol et al. 2004], [Tao e Guerin 2004], [Tao et al. 2005]. Estes mecanismos do controle freqüentemente dependem de modelos de perda que devem ser precisos e, no entanto, simples bastante para a análise em tempo real. Infelizmente, a dinâmica exata de processos de perdas de pacotes na Internet pode ser excepcionalmente variável através do espaço — isto é, entre os diferentes caminhos fim-a-fim — e do tempo, em um mesmo caminho, [Zhang et al. 2000]. Esse fato aumenta a dificuldade de inferir perdas.

Diferenças nas demandas de tráfego e nas capacidades dos canais contribuem para a natureza complexa e imprevisível da Internet. A recente popularização de tecnologias de redes sem fio acrescenta a este cenário a baixa confiabilidade inerente ao seu meio de transmissão, onde as taxas de erros em *bits* são costumeiramente maiores que aquelas vistas em fios. Entretanto, o que talvez seja mais digno de atenção é o fato de que mesmo em canais, cujas estatísticas permaneçam estacionárias ao longo do tempo, é possível encontrar indicações de correlações significativas entre o que ocorre a pacotes que estão separados por segundos entre suas transmissões, [Yajnik et al. 1999].

Por essas razões, é de suma importância não apenas construir modelos de perda *simples* e *flexíveis*, mas também desenvolver estratégias que permitam a um protocolo adaptativo prever corretamente estatísticas futuras de perdas de pacotes, levando em conta os efeitos de medições recentes.

Por *simples*, queremos dizer que o modelo deve ser computacionalmente tratável para ser usado pela aplicação em tempo real. *Flexibilidade*, por outro lado, implica que o modelo deve aproximar um conjunto razoável de características observáveis do processo real. Mais importante para os nossos propósitos, o modelo deve ser capaz de se adaptar a mudanças no processo de perda ao longo do tempo e prever o desempenho futuro, condicionado nesses efeitos localizados.

Provavelmente, os modelos mais aplicados a processos de perdas de pacotes são o *processo de Bernoulli* e a cadeia de Markov com 2 estados, usualmente referenciada como *modelo de Gilbert*. Recentemente, modelos de Markov ocultos (hidden Markov models — HMMs) também se tornaram comuns no contexto de modelagem de perdas, [Salamatian e Vaton 2001]. Avaliar a acurácia de HMMs para prever perdas de pacotes é um dos assuntos centrais discutidos neste trabalho.

Tentaremos explorar mais adiante o uso de modelos de Markov ocultos como ferramentas para prever perdas. Desenvolvemos um novo algoritmo que avalia a distribuição do número de pacotes perdidos em uma janela de tempo futura, dado um

histórico recente das estatísticas do canal. Avaliamos a qualidade dessas previsões usando diferentes modelos de Markov ocultos.

Também propomos uma variação da abordagem básica de HMMs, restringindo a estrutura do modelo. Nosso modelo resultante pode ser visto como um HMM hierárquico, com uma cadeia de Markov de 2 estados operando dentro de cada estado da cadeia oculta. A estrutura deste modelo possui duas propriedades interessantes. Primeiramente, restringindo o modelo, nós visamos a diminuir o número total de parâmetros a serem estimados, reduzindo assim a complexidade da fase de treinamento. Em segundo, supondo tais padrões no conjunto de parâmetros, nós tentamos capturar as dependências de curto prazo nos eventos da perda com um modelo de Gilbert, enquanto a dinâmica de longo prazo é governada por uma cadeia de Markov oculta.

Examinamos a literatura relacionada seguida por uma revisão de modelos ocultos de Markov e previsão linear no Capítulo 2. O Capítulo 3 introduz o algoritmo proposto para previsão e a metodologia usados em nossas experiências. No Capítulo 4, apresentamos três modelos que podem ser usados em conjunto com nosso algoritmo de previsão. Dois desses modelos são HMMs já estudados em outros contextos na literatura. O terceiro é uma proposta original desta dissertação, a qual consideramos mais apropriada para a tarefa da previsão de perdas de pacote. Em seguida, o Capítulo 5 mostra resultados experimentais baseados em *traces* reais coletados na Internet. Finalmente, o Capítulo 6 conclui este trabalho resumindo os nossos resultados e discutindo as direções futuras para esta pesquisa.

Capítulo 2

Conceitos e Trabalhos Relacionados

Dedicamos este capítulo a uma revisão dos conceitos relevantes ao desenvolvimento do nosso trabalho. Em paralelo à exposição de cada conceito, fazemos menção a outros trabalhos na literatura que, por razões diversas, têm relação com o tema desta dissertação.

Mais especificamente, na Seção 2.1, mostraremos abordagens para modelar perdas de pacotes e analisaremos algumas considerações a respeito do uso desses modelos em protocolos de rede adaptativos. Na Seção 2.2, faremos uma síntese de definições relacionadas a modelos de Markov ocultos. A Seção 2.3 mostra alguns resultados básicos relacionados à técnica de predição linear que também serão importantes ao desenvolvimento desta dissertação. Através das definições neste capítulo, apresentaremos também as convenções de notação que serão adotadas no material dos próximos capítulos.

2.1 Modelos para Perdas de Pacotes

Uma ferramenta simples e amplamente aplicada na modelagem de perdas de pacotes é a cadeia de Markov com dois estados, usualmente referida na literatura como o modelo de *Gilbert* ou modelo de *Gilbert-Elliot*. Na verdade, os modelos de Gilbert

e Gilbert-Elliot são mais gerais, uma vez que seus trabalhos originais — respectivamente, [Gilbert 1960] e [Elliott 1965] — os descreviam como modelos de Markov ocultos de dois estados. A Figura 2.1 ilustra estas diferenças.

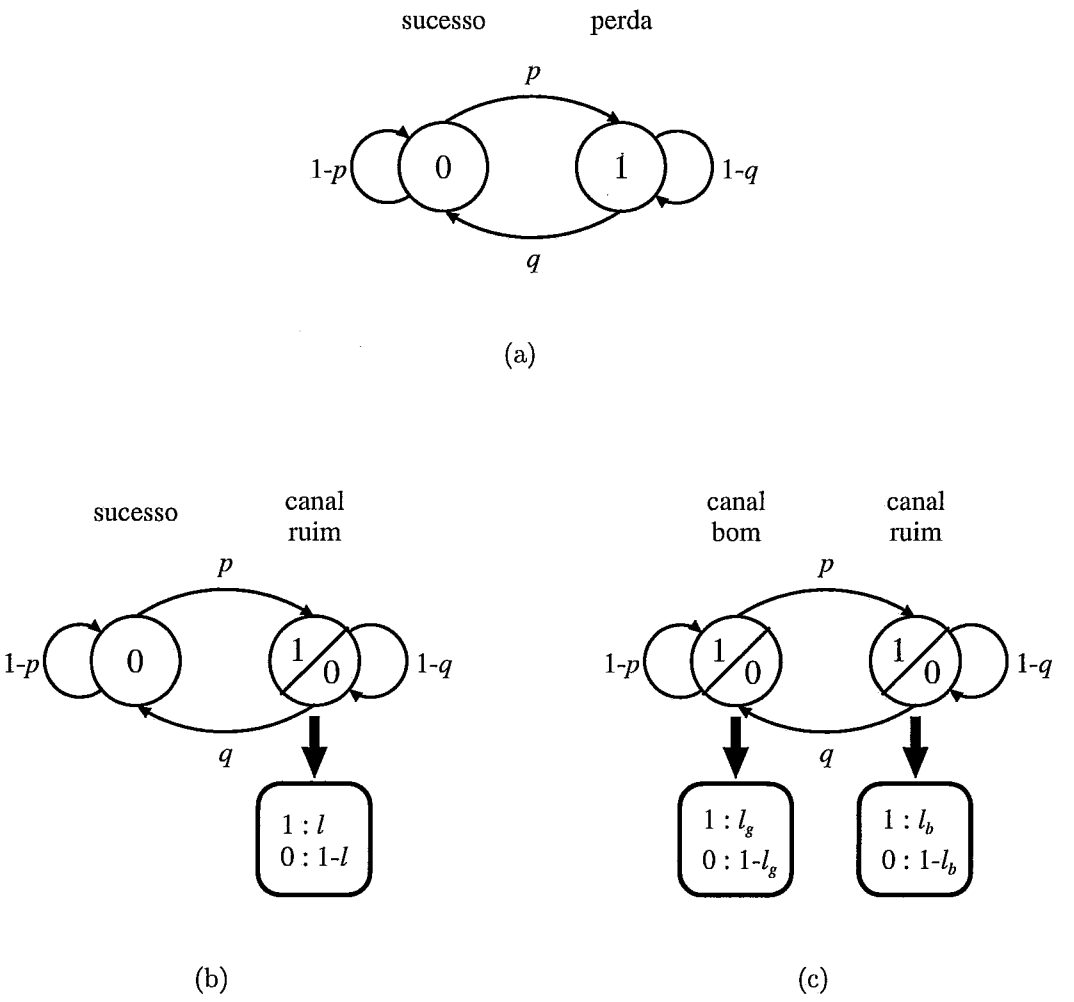


Figura 2.1: Modelos

A Figura 2.1(a) ilustra uma cadeia de Markov com dois estados, onde as probabilidades de transição a partir dos estados de sucesso e de perda são respectivamente p e q. No modelo de Gilbert, ilustrado na Figura 2.1(b), o estado denominado *bad*, produz símbolos de perdas com probabilidade l. Finalmente, no modelo de Gilbert-Elliot, visto na Figura 2.1(c), os estados são caracterizados como *good* e *bad*, com probabilidades de perdas respectivamente l_g e l_b, onde $l_g < l_b$. Os modelos de Gilbert e Gilbert-Elliot foram originalmente propostos no contexto de erros de *bits*

em transmissões de dados.

Ainda no contexto de modelos de Gilbert, o trabalho em [Elliott 1963] desenvolve uma recursão para computar $P(i, j)$, a probabilidade de observar exatamente i erros em j tentativas de transmissão. Mais recentemente, [Su et al. 2004] apresentou um método para calcular o valor de $P(i, j)$ condicionado em medições recentes da taxa de perda no canal, também para modelos de Gilbert. Nesta dissertação, apresentaremos um algoritmo que calcula $P(i, j)$ condicionado em medições recentes para modelos de Markov ocultos em geral. Embora nosso procedimento possa ser aplicado a modelos de Gilbert, ele não possui relação com aquele em [Su et al. 2004].

Apesar de sua extensa aplicabilidade, é sabido que a cadeia de Markov com 2 estados possui uma habilidade limitada em modelar dependências de longo prazo. Em [Yajnik et al. 1999], é argumentado que, em alguns cenários, é possível encontrar correlações estatisticamente significativas entre pacotes separados de 1 segundo. Nesse mesmo trabalho, uma cadeia de Markov de ordem k com memória suficiente foi usada para capturar essas correlações de longo prazo. Entretanto, a complexidade exponencial do espaço de estados nesses modelos os torna uma alternativa menos atraente para uso em aplicações de tempo real.

Em [Salamatian e Vaton 2001], foi mostrado que um modelo de Markov oculto com poucos estados é capaz de aproximar as estatísticas dos mesmos *traces* de [Yajnik et al. 1999] que requerem modelos de Markov de ordens altas. Apesar de o modelo de Gilbert-Elliot ser considerado um modelo de Markov oculto, o trabalho em [Salamatian e Vaton 2001] é original, uma vez que permitiu que o modelo tenha um número arbitrário de estados e utilizou o algoritmo EM para estimar parâmetros.

O trabalho em [Wei et al. 2002] considera o uso de um HMM de tempo contínuo, criado a partir de medições de perda e retardo, para simular características de canais fim-a-fim em estudos de avaliação de desempenho.

Uma propriedade de grande relevância na modelagem e previsão de uma série temporal é a estacionaridade. Lidar com variações bruscas nas estatísticas de in-

teresse ou ainda fenômenos determinísticos como periodicidade em longa escala são tarefas não triviais em problemas de pesquisa operacional [Brockwell e Davis 2002]. O trabalho de [Yajnik et al. 1999] descarta 52 de 128 horas de dados coletados que exibem características não estacionárias. Posteriormente, [Zhang et al. 2001] apresentou um tratamento mais detalhado de diferentes critérios de estacionaridade aplicados a medições em caminhos fim-a-fim na Internet.

Enlaces sem fio são particularmente mais suscetíveis a eventos transientes, como interferência de radio-freqüência, que seus correspondentes com fios. Em conjunto com fenômenos como atenuação de sinal e sombreamento (*shadowing*), isso leva a taxas de erros em bits mais altas além de maior variabilidade nas estatísticas de perda de pacotes, [Rappaport 2001]. Em [Konrad et al. 2001], é argumentado que essas diferenças podem ter um papel crucial em estudos de avaliação de desempenho, uma vez que a modelagem inapropriada destas características pode levar à estimação errada de parâmetros ótimos para protocolos de rede. Esse mesmo trabalho propõe o algoritmo *Markov-based Trace Analysis (MTA)*, que agrupa perdas de pacotes próximas em estados de contenção, e apenas então modela estes períodos como cadeias de Markov de ordem k.

Posteriormente, [Ji et al. 2004] retornou ao problema de modelagem de erros ao nível de *frames* em redes sem fio, utilizando uma cadeia semi-Markoviana de 2 estados com tempo de permanência em cada estado estimado a partir de *traces* através de misturas de distribuições geométricas. Isto é feito, uma vez que uma mistura de variáveis geométricas pode aproximar arbitrariamente qualquer distribuição discreta, uma vez que haja elementos suficientes na mistura. O artigo se refere a esse modelo como um *modelo de Gilbert estendido*, e avalia sua capacidade de aproximar estatísticas de longo prazo em medições de uma rede GSM. Comparações são traçadas entre o modelo de Gilbert estendido, a cadeia de Markov de ordem k de [Yajnik et al. 1999], os HMMs de [Salamatian e Vaton 2001] e o MTA de [Konrad et al. 2001].

O trabalho de [Tao e Guerin 2004] desenvolve um modelo hierárquico de dois níveis para prever o desempenho de perdas em caminhos fim-a-fim e realizar *path-*

switching para aplicações multimídia. No nível mais alto, uma cadeia de Markov com N estados é utilizada, a cada minuto, para selecionar um de N HMMs no nível mais baixo que, por sua vez, modelam os eventos de perdas para cada pacote. A previsão feita baseia-se em determinar o estado da cadeia de Markov de alto nível para o próximo minuto e obter a fração de perda em estado estacionário do HMM correspondente no nível inferior. A partir destas previsões, o mecanismo de *path-switching* pode escolher o canal cujo modelo prevê a menor taxa de perda.

Embora existam trabalhos que utilizam modelos Markovianos para prever estatísticas de perdas de pacotes, eles o fazem considerando apenas medidas de estado estacionário. Se as condições da rede forem muito variáveis em escalas de tempo relativamente curtas, essa hipótese pode levar a erros significativos, conforme mostraremos no decorrer desta dissertação.

Em [Duarte 2003] e [Duarte et al. 2003], os autores consideram um modelo de Markov oculto, que agrega o número total de pacotes perdidos em uma seqüência de δ tentativas de transmissão. O modelo também restringe as transições entre estados ocultos de forma similar a uma cadeia absorvente, com tempo para absorção seguindo uma distribuição *phase-type*. Essa restrição foi feita para tentar capturar um padrão específico de correlações periódicas que ocorrem, com freqüência significativa, nos *traces* estudados no trabalho de [Duarte 2003].

O modelo de [Duarte 2003] foi utilizado para a previsão da quantidade de perdas em intervalos de 1 segundo. A previsão é aplicada à seleção de um esquema de FEC para recuperação de perdas em simulações com *traces* de tráfego de voz. Recentemente, esse algoritmo foi implementado e avaliado em uma ferramenta de transmissão de Voz sobre IP real, [de Vielmond e de Souza e Silva 2005].

2.2 Modelos de Markov Ocultos

Uma referência concisa sobre modelos de Markov ocultos pode ser encontrada em [Rabiner 1989], enquanto que um tratamento mais profundo do assunto está em [Elliot et al. 1995]. A seguir, apresentaremos definições que serão usadas no desenvolvimento dos próximos capítulos. A maior parte da notação adotada segue aquela definida em [Rabiner 1989].

Neste trabalho, consideramos apenas modelos em tempo discreto, com ambos espaços de estados — ocultos e observáveis — também discretos. As referências que acabamos de mencionar incluem informações sobre modelos com características contínuas.

De forma geral, um modelo de Markov oculto é composto por dois processos estocásticos dependentes entre si. O primeiro desses é uma cadeia de Markov. Uma excelente referência para os principais aspectos de cadeias de Markov é [Kemeny e Snell 1969]. O segundo componente de um HMM é um processo de observações, cuja distribuição, a qualquer instante de tempo, é completamente determinada pelo estado atual da cadeia.

Seja $\{Y_t\}$ a cadeia de Markov de N estados. A distribuição do estado inicial é dada pelo vetor de N dimensões π, com:

$$\pi_i = P(Y_1 = i). \tag{2.1}$$

As probabilidades de transição entre estados são controladas pela matriz $N \times N$, $\mathbf{A} = \{a_{ij}\}$, onde:

$$a_{ij} = P(Y_t = j | Y_{t-1} = i). \tag{2.2}$$

O processo de observações, $\{X_t\}$, tem M estados e é governado pela matriz $N \times M$, $\mathbf{B} = \{b_{ij}\}$, i.e.:

$$b_{ij} = P(X_t = j | Y_t = i). \tag{2.3}$$

Dados os significados probabilísticos de π, $\mathbf{A}$ e $\mathbf{B}$, as restrições a seguir serão

sempre satisfeitas:

$$\sum_{i=1}^{N} \pi_i = 1, \tag{2.4a}$$

$$\sum_{j=1}^{N} a_{ij} = 1, \quad \forall i, \tag{2.4b}$$

$$\sum_{j=1}^{M} b_{ij} = 1, \quad \forall i. \tag{2.4c}$$

Sempre que possível, por brevidade, iremos nos referir ao conjunto de parâmetros do modelo como a tripla $\lambda = (\pi, \mathbf{A}, \mathbf{B})$.

Um primeiro passo em criar um modelo é especificar os espaços de estados sobre os quais $\{X_t\}$ e $\{Y_t\}$ estão definidos. Uma vez que $\{X_t\}$ é o processo de observações, seus estados são geralmente determinados pelo que está sendo modelado. A maneira mais direta de modelar eventos de perdas de pacotes é representar cada pacote individual com um símbolo binário. Usamos o símbolo 1 para representar uma perda, e 0 como um indicador de que o pacote é entregue com sucesso. O trabalho de [Duarte 2003] considera um modelo de observações diferente que consiste de 51 símbolos: os inteiros de 0 a 50, para representar o número total de perdas em um grupo de 50 pacotes. Ambas abordagens serão consideradas mais tarde em nossos experimentos de previsão.

Por outro lado, caracterizar os estados da cadeia oculta, $\{Y_t\}$, pode ser um pouco mais abstrato. Em modelos para perdas de pacotes, os estados ocultos podem ser vistos como "estados da rede", guardando informação sobre as estatísticas de perdas em um dado momento.

Consideremos um vetor com T valores para o processo de observações, $\mathbf{x} = [x_1, \ldots, x_T]$. Sempre que não houver ambigüidade, usaremos a forma abreviada $X_{i:j}$ (ou a correspondente, $Y_{i:j}$) para denotar o evento composto que cada variável X_t (Y_t) no alcance $t = i, \ldots, j$ assume o valor x_t (y_t). No caso particular em que $i = j$, escreveremos X_i (ou de maneira equivalente, Y_i). Por outro lado, usaremos $\mathbf{X}$ (ou $\mathbf{Y}$) quando os sub-índices se referem a todas as variáveis $1, \ldots, T$, i.e., $X_{1:T}$ ($Y_{1:T}$).

Finalmente, também definimos as seguintes medidas de probabilidade, seguindo a notação de [Rabiner 1989]:

$$\alpha_t(i) = P(X_{1:t}, Y_t = i|\lambda), \tag{2.5a}$$

$$\beta_t(i) = P(X_{t+1:T}|Y_t = i, \lambda), \tag{2.5b}$$

$$\gamma_t(i) = P(Y_t = i|\mathbf{X}, \lambda), \tag{2.5c}$$

$$\xi_t(i,j) = P(Y_t = i, Y_{t+1} = j|\mathbf{X}, \lambda). \tag{2.5d}$$

Onde $\alpha_t(i)$ e $\beta_t(i)$ são calculados através do algoritmo *forward-backward* (veja a Seção A.2 para detalhes) e as seguintes identidades podem ser estabelecidas:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}, \tag{2.6a}$$

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_{jx_{t+1}}\beta_{t+1}(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}. \tag{2.6b}$$

A principal dificuldade da abordagem de modelos de Markov ocultos é o problema de estimação de parâmetros, isto é, como inferir valores para λ dado um caminho amostral do processo de observações. A dificuldade reside no fato de que não há fórmulas fechadas para estimadores de *máxima verossimilhança* (*maximum likelihood*) dos parâmetros de um HMM, como existem para cadeias de Markov. De fato, a função de verossimilhança para um HMM é, por si só, muito complexa para ser analiticamente otimizada, i.e., para ter suas condições de otimalidade verificadas.

Apesar dessa dificuldade, o algoritmo *Baum-Welch* é uma técnica muito bem sucedida para estimação iterativa de parâmetros de máxima verossimilhança para modelos de Markov ocultos. O método começa a partir de uma atribuição arbitrária de valores para λ e produz estimativas sucessivamente melhores, garantindo convergência para um máximo local na função de verossimilhança, sempre que um existir, [Dempster et al. 1977]. A seguir, iremos rever as fórmulas da estimação do método Baum-Welch para motivar discussões posteriores neste trabalho.

A função de verossimilhança dos parâmetros, $\lambda = (\pi, \mathbf{A}, \mathbf{B})$, para uma amostra,

$\mathbf{X}$, pode ser escrita como:

$$\begin{aligned} L(\lambda|\mathbf{X}) &= P(\mathbf{X}|\lambda) \\ &= \sum_{\forall \mathbf{y}} P(\mathbf{X}, \mathbf{Y}|\lambda). \end{aligned} \tag{2.7}$$

Onde a medida $P(\mathbf{X}, \mathbf{Y}|\lambda)$ é chamada *verossimilhança dos dados completos*, uma vez que envolve dados observáveis e ocultos, representados por $x_{1:T}$ e $y_{1:T}$, respectivamente. Para um modelo de Markov oculto, essa função é definida como:

$$P(\mathbf{X}, \mathbf{Y}|\lambda) = P(Y_1|\lambda)P(X_1|Y_1, \lambda)\prod_{t=2}^{T} P(Y_t|Y_{t-1}, \lambda)P(X_t|Y_t, \lambda), \tag{2.8}$$

onde temos as correspondências:

$$P(Y_1|\lambda) = \pi_{y_1}, \tag{2.9a}$$

$$P(Y_t|Y_{t-1}, \lambda) = a_{y_{t-1}, y_t}, \tag{2.9b}$$

$$P(X_t|Y_t, \lambda) = b_{y_t, x_t}. \tag{2.9c}$$

Logo, a Equação (2.8) pode ser reescrita como:

$$P(\mathbf{X}, \mathbf{Y}|\lambda) = \pi_{y_1} b_{y_1, x_1} \prod_{t=2}^{T} a_{y_{t-1}, y_t} b_{y_t, x_t}. \tag{2.10}$$

A cada iteração, o algoritmo Baum-Welch maximiza a *função auxiliar*, $Q(\lambda|\overline{\lambda})$, em relação a λ, fazendo uso da estimativa atual dos parâmetros, $\overline{\lambda}$:

$$Q(\lambda|\overline{\lambda}) = \sum_{\forall \mathbf{y}} \log P(\mathbf{X}, \mathbf{Y}|\lambda) P(\mathbf{Y}|\mathbf{X}, \overline{\lambda}). \tag{2.11}$$

Embora esse procedimento tenha sido apresentado pela primeira vez em 1970, [Baum et al. 1970], mais tarde ele foi generalizado por uma classe de métodos estatísticos para estimação de máxima verossimilhança, conhecida como *Expectation-Maximization*, ou simplesmente EM. O trabalho clássico sobre o método EM é [Dempster et al. 1977], onde seus principais resultados de convergência são apresentados.

A abordagem EM consiste em realizar dois passos em cada iteração. O passo *Expectation* (ou passo-E) avalia a função auxiliar (2.11), enquanto o passo *Maximization* (ou passo-M) obtém o valor de λ que a maximiza.

Usando a *desigualdade de Jensen* (veja o Teorema A.1 na Seção A.3), é facilmente demonstrado que:

$$\begin{aligned} Q(\lambda|\overline{\lambda}) &= \sum_{\forall \mathbf{y}} \log P(\mathbf{X}, \mathbf{Y}|\lambda) P(\mathbf{Y}|\mathbf{X}, \overline{\lambda}) \\ &\leq \log \left[\sum_{\forall \mathbf{y}} P(\mathbf{X}, \mathbf{Y}|\lambda) P(\mathbf{Y}|\mathbf{X}, \overline{\lambda}) \right] \\ &= \log P(\mathbf{X}|\lambda), \end{aligned} \tag{2.12}$$

isto é, a função auxiliar é um limite inferior para o logaritmo da função de verossimilhança, com a igualdade ocorrendo quando $\lambda = \overline{\lambda}$. Logo, maximizar $Q(\lambda|\overline{\lambda})$ nunca leva a um decréscimo na verossimilhança.

Usando a definição da verossimilhança conjunta, $P(\mathbf{X}, \mathbf{Y}|\lambda)$, a Equação (2.11) pode ser dividida em três termos independentes:

$$Q(\lambda|\overline{\lambda}) = Q_1(\pi|\overline{\lambda}) + Q_2(\mathbf{A}|\overline{\lambda}) + Q_3(\mathbf{B}|\overline{\lambda}), \tag{2.13}$$

onde $Q_1(\pi|\overline{\lambda})$, $Q_2(\mathbf{A}|\overline{\lambda})$, $Q_3(\mathbf{B}|\overline{\lambda})$ são dados por:

$$Q_1(\pi|\overline{\lambda}) = \sum_{i=1}^{N} \log \pi_i \, \gamma_1(i), \tag{2.14a}$$

$$Q_2(\mathbf{A}|\overline{\lambda}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \log a_{ij} \sum_{t=1}^{T-1} \xi_t(i,j), \tag{2.14b}$$

$$Q_3(\mathbf{B}|\overline{\lambda}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \log b_{ij} \sum_{t=1}^{T} \mathbb{I}\{x_t = j\} \gamma_t(i). \tag{2.14c}$$

Na Equação (2.14c), utilizamos a notação $\mathbb{I}\{c\}$, para representar a *função indicadora* de uma condição c, que vale 1 quando a condição é satisfeita, ou 0 no caso contrário.

Maximizando cada termo de (2.14) e levando em consideração as restrições es-

tocásticas de (2.4), chegamos às fórmulas de estimação de parâmetros:

$$\pi_i = \gamma_1(i), \tag{2.15a}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \tag{2.15b}$$

$$b_{ij} = \frac{\sum_{t=1}^{T} \mathbb{I}\{x_t = j\}\gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}. \tag{2.15c}$$

O leitor, interessado nos passos da derivação dessas fórmulas, pode remeter à Seção A.4.

As expressões em (2.15) são tradicionalmente interpretadas como expressões envolvendo valores esperados, da seguinte forma:

$$\pi_i = E[\,\#\text{ de observações no estado } i \text{ em } t = 1\,], \tag{2.16a}$$

$$a_{ij} = \frac{E[\,\#\text{ de transições de } i \text{ para } j\,]}{E[\,\#\text{ de transições a partir de } i\,]}, \tag{2.16b}$$

$$b_{ij} = \frac{E[\,\#\text{ de observações de } j \text{ em } i\,]}{E[\,\#\text{ de observações no estado } i\,]}. \tag{2.16c}$$

2.3 Previsão Linear

Previsão linear é uma técnica comum em processamento digital de sinais e em modelagem matemática. No domínio de processamento de voz, previsão linear é mais conhecida como *linear predictive coding* (LPC — conforme visto, por exemplo, em [Rabiner e Schafer 1978]). A previsão linear está também relacionada ao conceito de modelo autoregressivo (AR), que é uma parte fundamental de ferramentas mais tradicionais para previsão de séries temporais como modelos ARMA, ARIMA e FARIMA, para citar alguns exemplos, [Brockwell e Davis 2002].

Seja r_t a taxa de perdas medida no t-ésimo intervalo de tempo. O preditor linear $\hat{r}_t$ é definido como:

$$\hat{r}_t = \sum_{i=1}^{k} a_i r_{t-i}. \tag{2.17}$$

Onde k é a ordem do preditor e os a_i's são os coeficientes da combinação linear que define $\hat{r}_n$. Consideramos um preditor de primeira ordem, i.e., um da forma $\hat{r}_t = ar_{t-1}$. Se definirmos a medida de erro $e_t = r_t - \hat{r}_t$, então o *erro médio quadrático* (*mean squared error* — MSE) de um preditor linear é simplesmente o valor esperado $E[e_t^2]$, que pode ser escrito como uma função de a usando:

$$E[e_t^2] = E[r_{t-1}^2] - 2E[r_{t-1}r_t]a + E[r_{t-1}^2]a^2. \tag{2.18}$$

O mínimo de (2.18) é alcançado em $a^* = E[r_{t-1}r_t]/E[r_{t-1}^2]$, e o MSE para esse preditor linear de primeira ordem ótimo é dado por:

$$\min_a E[e_t^2] = E[r_t^2] - \frac{E[r_{t-1}r_t]^2}{E[r_{t-1}^2]}. \tag{2.19}$$

Uma vez que $E[r_t^2]$, $E[r_{t-1}^2]$ e $E[r_{t-1}r_t]$ não podem ser determinados de antemão, o preditor linear ótimo não é prático para o nosso propósito de previsão em tempo real. Seja um caso especial dessa técnica com $a = 1$, isto é, um preditor da forma $\hat{r}_t = r_{t-1}$. Nos referimos a esta técnica como a estratégia replicadora, uma vez que simplesmente repete o valor da medição anterior como a próxima previsão. O erro de previsão para essa estratégia é dado por:

$$E[e_t^2] = E[r_t^2] + E[r_{t-1}^2] - 2E[r_{t-1}r_t]. \tag{2.20}$$

Assumindo a estacionaridade fraca de r_t, denotamos por $r(i)$ a correlação produto entre as variáveis r_t e r_{t-i}, isto é, $r(i) = E[r_t r_{t-i}]$. Se d é a diferença $r(0) - r(1)$, então o erro do replicador, (2.20), é dado por $2d$, e o erro mínimo do preditor linear, (2.19), pode ser escrito como:

$$\min_a E[e_t^2] = \frac{r(0) + r(1)}{r(0)}\, d. \tag{2.21}$$

Uma vez que $0 \leq r(1) \leq r(0)$, o erro do preditor linear ótimo está limitado segundo a expressão:

$$d \leq \min_a E[e_t^2] \leq 2d. \tag{2.22}$$

Capítulo 3

Algoritmo de Previsão

Medidas de estado estacionário fornecem somente médias de longa duração do processo de perdas que está sendo observado. Um de nossos objetivos é estimar a habilidade de um modelo de prever estatísticas de perda em uma curta duração de tempo. Para atingir este objetivo, calculamos estimativas para a taxa da perda a partir de medições recentes.

Primeiramente, na Seção 3.1, apresentamos a metodologia de previsão adaptativa que usamos em nossos experimentos. Na Seção 3.2, desenvolvemos um algoritmo para avaliar a distribuição da taxa de perda enxergada por um fluxo de pacotes em uma janela de tempo futura. Finalmente, na Seção 3.3, apresentamos um teorema que nos permite escolher o valor esperado desta distribuição como uma métrica de previsão ótima, sob o ponto de vista do erro médio quadrático.

3.1 Metodologia de Previsão Adaptativa

Para fazer uso das previsões geradas por um algoritmo tal como o que será apresentado na Seção 3.2, é necessário especificar quando as métricas de previsão devem ser avaliadas e também quando os parâmetros do modelo devem ser re-estimados. Nesta seção, descrevemos um mecanismo simples para previsão adaptativa que utilizamos

em nossos experimentos.

A Figura 3.1 mostra um esquema geral de nossa metodologia em duas camadas. Na camada de *treinamento do modelo*, os parâmetros do modelo são periodicamente re-estimados a cada τ unidades de tempo. Em cada treino, apenas as amostras das últimas T unidades de tempo são usadas para o procedimento de estimação de parâmetros. Cada época de treinamento também é dividida em intervalos de previsão de tamanho ψ, conforme ilustrado na camada de *previsão da medida*. Cada previsão individual pode ser condicionada nas amostras de pacotes das H unidades de tempo mais recentes. Uma vez que estamos interessados em calcular medidas para intervalos finitos no futuro, também introduzimos um parâmetro F que especifica o tamanho da janela de previsão. Na Seção 3.2, estaremos interessados em calcular a taxa de perdas no intervalo de tempo que vai de um instante atual t, até o instante no futuro, $t + F$.

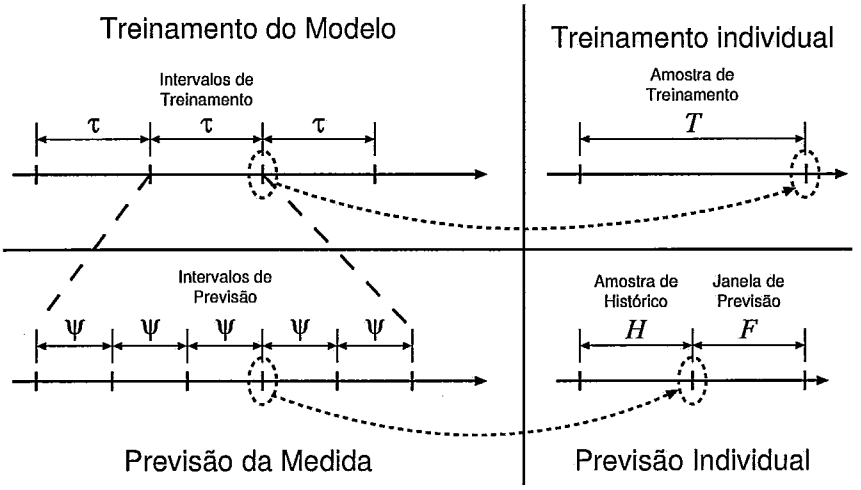


Figura 3.1: Mecanismo de previsão adaptativo em duas camadas: treinamento e previsão.

Através de experimentos, verificamos que cada um desses parâmetros pode ter diferentes impactos na qualidade de previsão. Os valores de T e H, por exemplo, desempenham papéis importantes em perceber os efeitos das mudanças recentes nas estatísticas do canal. Se pelo menos um desses parâmetros é muito alto, as previsões se tornam muito mais suaves, basicamente refletindo as medidas de es-

tado estacionário. Valores que são muito baixos, por outro lado, irão falhar em incluir informação suficiente para permitir ao modelo estimar corretamente os seus parâmetros ou realizar uma previsão precisa. Claramente, em um cenário real se deseja ter os valores de τ e ψ tão grandes quanto possível para minimizar a perda de eficiência da aplicação.

A informação de perdas de pacotes usada para condicionar as estatísticas de previsão, na prática, não estão disponíveis ao transmissor imediatamente após a ocorrência dessas perdas. Uma vez que o transmissor deve esperar um *round-trip time* (RTT) até que a previsão possa ser feita, uma parte das estatísticas previstas será inútil na tomada de decisões de controle. Por causa disso, no mecanismo de previsão de taxa de perda, o valor de F não deve ser muito pequeno em relação ao RTT estimado. Por outro lado, é fácil perceber que ao fazer F tender ao infinito, a taxa de perdas prevista será independente do histórico dado por H e convergirá para a probabilidade de perda em estado estacionário.

Em nossos experimentos, tentamos uma série de variações para cada um desses parâmetros. Reconhecemos que uma análise mais profunda, da sensibilidade de cada parâmetro, seria necessária para facilitar o uso da metodologia de previsão, no caso geral. Entretanto, uma apresentação extensa dessas comparações tornaria a exposição por demais cansativa e, por essa razão, no Capítulo 5, apresentaremos apenas resultados baseados em valores de parâmetros que descobrimos funcionar bem na prática de nossos experimentos. Na Seção 5.3, apresentaremos os parâmetros escolhidos, justificando cada escolha.

3.2 Previsão de Taxas de Perda

Estimar a fração a curto prazo de pacotes perdidos no canal pode ser extremamente importante, especialmente se esta medida convergir lentamente para o estado estacionário. Dada uma janela fixa de F unidades de tempo, a taxa de perdas no curto prazo é simplesmente a fração dos pacotes transmitidos nesta janela que não

chegam ao seu destino. Para simplificar a análise, iremos assumir que a quantidade de pacotes transmitidos em F unidades de tempo é igual a f.

Similarmente aos trabalhos de [Elliott 1963] e de [Su et al. 2004], nós computamos a distribuição exata de i erros em j transmissões. Entretanto, nosso trabalho é mais geral do que essas referências, uma vez que computamos esta medida para qualquer HMM, enquanto aqueles se aplicam somente a canais de Gilbert ou Gilbert-Elliot. Também, nossa distribuição para o número das perdas é condicionada no resultado de medidas recentes de pacotes. O algoritmo apresentado a seguir é uma contribuição original deste trabalho.

Seja R_t^f a variável aleatória denotando o número total de eventos de perda que ocorrem nos f pacotes transmitidos a partir da t-ésima observação. Em outras palavras, R_t^f é a soma de cada um dos f valores de observação, de t até $t+f-1$, i.e.:

$$R_t^f = \sum_{i=1}^{f} X_{t+i-1}. \tag{3.1}$$

Os resultados a seguir podem ser aplicados a qualquer modelo em que as observações representam o número de perdas em uma unidade de tempo, com o modelo 0-1 sendo apenas um caso especial. Por exemplo, nossos resultados podem ser aplicados ao modelo de Markov oculto apresentado em [Duarte 2003], onde as observações podem variar de 0 a 50 perdas observadas em um segundo. Consideramos então que as observações no intervalo dos inteiros de 0 a um dado máximo r. Como conseqüência, a variável aleatória R_t^f estará entre 0 e rf, inclusive.

Seja h a quantidade de pacotes transmitidos no intervalo de histórico, de duração H unidades de tempo. Queremos calcular a distribuição de R_t^f dadas as h amostras mais recentes das observações passadas, $X_{t-h:t-1}$. Esta é a base do nosso preditor para a taxa de perdas no curto prazo, em um canal de comunicação. Condicionando

no valor do estado oculto na t-ésima observação, Y_t, podemos escrever:

$$\begin{aligned} P(R_t^f = j|X_{t-h:t-1}) &= \sum_{\forall y_t} P(R_t^f = j, Y_t|X_{t-h:t-1}) \\ &= \sum_{\forall y_t} P(R_t^f = j|Y_t, X_{t-h:t-1})P(Y_t|X_{t-h:t-1}) \\ &= \sum_{\forall y_t} P(R_t^f = j|Y_t)P(Y_t|X_{t-h:t-1}), \end{aligned} \tag{3.2}$$

onde, na última igualdade, utilizamos a independência condicional entre R_t^f, que é uma função das observações futuras, e as observações passadas, $X_{t-h:t-1}$, dado o estado oculto no tempo t, Y_t (veja o Corolário A.2). Primeiramente, notamos que este problema, assim como a maioria daqueles relacionados à previsão usando modelos de Markov ocultos, pode ser dividido em dois passos: (a) prever o estado oculto no início da janela no futuro, condicionado nas observações passadas; e (b) calcular a distribuição da métrica no futuro condicionada no estado atual.

Defina $\mathbf{r}_t^{f,h}$ como o vetor de probabilidades para R_t^f dado o histórico passado, e $\mathbf{R}^f$ como a matriz cujo elemento na linha i e coluna j é $P(R_t^f = j|Y_t = i)$. Uma vez que um modelo de Markov oculto é um processo estocástico homogêneo no tempo, $P(R_t^f = j|Y_t = i)$ é a mesma medida para todo t. Seja:

$$\pi_{t,h}(i) = P(Y_t = i|X_{t-h:t-1}); \tag{3.3}$$

i.e. $\pi_{t,h}$ é o vetor de probabilidades dos estados ocultos em t dadas as observações passadas. Podemos reescrever a Equação (3.2) como:

$$\mathbf{r}_t^{f,h} = \pi_{t,h}\mathbf{R}^f. \tag{3.4}$$

A distribuição do estado oculto, $\pi_{t,h}(i)$, pode ser facilmente obtida através da variável *forward*, $\alpha_t(i)$, conforme a definição em (2.5a), porém medida apenas no conjunto

de observações $X_{t-h:t-1}$:

$$\begin{aligned}
\pi_{t,h}(i) &= \frac{P(Y_t = i, X_{t-h:t-1})}{P(X_{t-h:t-1})} \\
&= \frac{\sum_{\forall y_{t-1}} P(Y_{t-1}, X_{t-h:t-1}, Y_t = i)}{P(X_{t-h:t-1})} \\
&= \frac{\sum_{\forall y_{t-1}} P(Y_{t-1}, X_{t-h:t-1}) P(Y_t = i | Y_{t-1}, X_{t-h:t-1})}{P(X_{t-h:t-1})} \\
&= \frac{\sum_{\forall y_{t-1}} P(Y_{t-1}, X_{t-h:t-1}) P(Y_t = i | Y_{t-1})}{P(X_{t-h:t-1})} \\
&= \frac{\sum_{\forall y_{t-1}} \alpha_{t-1}(y_{t-1}) a_{y_{t-1} i}}{P(X_{t-h:t-1})}.
\end{aligned} \tag{3.5}$$

O procedimento de cálculo da variável *forward* pode ser encontrado em detalhes na Seção A.2.

Se denotarmos por α_t o vetor cujo i-ésimo elemento é $\alpha_t(i)$, então (3.5) pode ser reescrito como:

$$\pi_{t,h} = \frac{\alpha_{t-1}\mathbf{A}}{P(X_{t-h:t-1})}. \tag{3.6}$$

Para desenvolver uma recursão que calcula a matriz $\mathbf{R}^f$, chamamos atenção ao fato que:

$$\begin{aligned}
P(R_t^1 = j | Y_t = i) &= P(X_t = j | Y_t = i) \\
&= b_{ij}.
\end{aligned} \tag{3.7}$$

A matriz $\mathbf{R}^1$ é, portanto, a matriz de observação, $\mathbf{B}$. No caso geral, condicionando no valor da observação subseqüente, X_t, podemos reescrever cada elemento de $\mathbf{R}^f$ como:

$$\begin{aligned}
P(R_t^f = j | Y_t = i) &= \sum_{\forall x_t} P(R_t^f = j, X_t | Y_t = i) \\
&= \sum_{\forall x_t} P(R_t^f = j | X_t, Y_t = i) P(X_t | Y_t = i) \\
&= \sum_{\forall x_t} P(R_{t+1}^{f-1} = j - x_t | X_t, Y_t = i) P(X_t | Y_t = i) \\
&= \sum_{\forall x_t} P(R_{t+1}^{f-1} = j - x_t | Y_t = i) b_{i x_t}.
\end{aligned} \tag{3.8}$$

Onde a penúltima igualdade é evidente pela definição (3.1) e a última igualdade foi obtida usando a independência condicional de R_{t+1}^{f-1} e X_t, dado Y_t, de acordo com o Corolário A.2. Para fechar a relação de recorrência, observamos que:

$$\begin{aligned} P(R_{t+1}^{f-1} = j - x_t | Y_t = i) &= \sum_{\forall y_{t+1}} P(R_{t+1}^{f-1} = j - x_t, Y_{t+1} | Y_t = i) P(Y_{t+1} | Y_t = i) \\ &= \sum_{\forall y_{t+1}} P(R_{t+1}^{f-1} = j - x_t | Y_{t+1}, Y_t = i) P(Y_{t+1} | Y_t = i) \\ &= \sum_{\forall y_{t+1}} P(R_{t+1}^{f-1} = j - x_t | Y_{t+1}) P(Y_{t+1} | Y_t = i) \\ &= \sum_{\forall y_{t+1}} P(R_{t+1}^{f-1} = j - x_t | Y_{t+1}) a_{iy_{t+1}}. \end{aligned} \tag{3.9}$$

Usando (3.9) em (3.8), temos:

$$P(R_t^f = j | Y_t = i) = \sum_{\forall x_t} \left[\sum_{\forall y_{t+1}} P(R_{t+1}^{f-1} = j - x_t | Y_{t+1}) a_{iy_{t+1}} \right] b_{ix_t} \tag{3.10}$$

Podemos reescrever nossa recursão em forma matricial como:

$$\mathbf{R}^k = \begin{cases} \mathbf{B} & , \quad k = 1 \\ \sum_{\forall x_t} \mathbf{B}(x_t) \mathbf{A} \mathbf{R}^{k-1} \mathbf{I}_k(x_t) & , \quad 2 \leq k \leq f \end{cases} \tag{3.11}$$

Onde $\mathbf{B}(x_t) = \text{diag}\{b_{ix_t}\}$ e $\mathbf{I}_k(x_t)$ é uma matriz $(1 + rk - r) \times (1 + rk)$, composta por uma matriz identidade $(1 + rk - r) \times (1 + rk - r)$, deslocada de x_t colunas para a direita, e com zeros em todos os elementos restantes. Em retrospectiva, os passos de nosso algoritmo são:

$$\begin{aligned} &\text{Iniciação} \begin{cases} \pi_{t,h} \leftarrow \alpha_{t-1} \mathbf{A} \\ \mathbf{R}^1 \leftarrow \mathbf{B} \end{cases} \\ &\text{Laço principal} \begin{cases} \text{Para } 2 \leq k \leq f \text{ faça:} \\ \quad \mathbf{R}^k \leftarrow \sum_{\forall x_t} \mathbf{B}(x_t) \mathbf{A} \mathbf{R}^{k-1} \mathbf{I}_k(x_t) \end{cases} \\ &\text{Resultado} \begin{cases} \mathbf{r}_t^{f,h} \leftarrow \pi_{t,h} \mathbf{R}^f \end{cases} \end{aligned} \tag{3.12}$$

3.3 Estratégia Ótima de Previsão Linear

O algoritmo da Seção 3.2 pode avaliar a distribuição de R_t^f assumindo que o processo de perdas pode ser modelado por HMM. No entanto, há mais de uma maneira de usar essa informação para obter uma medida que pode ser usada como valor de previsão. Por exemplo, podemos escolher o *preditor de maior probabilidade*, i.e., $\arg\max_i P(R_t^f = i)$, como a estatística de previsão. Apresentamos um teorema que justifica a nossa escolha de preditor.

Teorema 3.1. *Se um modelo pode estimar corretamente as correlações* $E[(R_{t-f}^f)^2]$ *e* $E[R_{t-f}^f R_t^f]$, *condicionando nas perdas anteriores* $X_{t-h:t-1}$, *onde* $h \geq f$, *então o preditor linear de primeira ordem ótimo para* R_t^f *é equivalente a calcular o valor esperado* $E[R_t^f | X_{t-h:t-1}]$.

Demonstração. Pela aplicação dos resultados da Seção 2.3, o preditor linear de primeira ordem de R_t^f é simplesmente:

$$\hat{R}_t^f = a^* R_{t-f}^f, \tag{3.13}$$

onde a^*, o parâmetro ótimo em termos do MSE, condicionado no conhecimento prévio das medições em $X_{t-h:t-1}$, é dado por:

$$a^* = \frac{E[R_{t-f}^f R_t^f | X_{t-h:t-1}]}{E[(R_{t-f}^f)^2 | X_{t-h:t-1}]}. \tag{3.14}$$

Nos resta notar que, se $h \geq f$, então, R_{t-f}^f pode ser determinado somando os valores $X_{t-f:t-1}$. Logo, o preditor ótimo tem a forma:

$$\hat{R}_t^f = \frac{R_{t-f}^f E[R_t^f | X_{t-h:t-1}]}{(R_{t-f}^f)^2} R_{t-f}^f = E[R_t^f | X_{t-h:t-1}]. \tag{3.15}$$

□

O Teorema 3.1 nos dá a indicação que entre todas as estatísticas que podemos avaliar da distribuição prevista para R_t^f, o valor esperado tem a propriedade atraente

de que seu MSE se aproximará daquele de um preditor linear de primeira ordem ótimo.

O algoritmo em (3.12) pode ser simplificado se a aplicação de interesse precisar apenas do valor esperado da taxa de perda ao invés de sua distribuição. Apesar disto, outras métricas da distribuição de R_t^f podem ser usadas com diferentes propósitos. Por exemplo, para algumas aplicações pode ser mais importante determinar a probabilidade de que a taxa de perda esteja acima de um dado limiar. De maneira similar, podemos estar interessados em determinar intervalos de previsão que contenham as medições reais com uma probabilidade p. Logo, nosso algoritmo pode ser usado em um escopo de aplicações muito mais amplo do que aquele considerado nesta dissertação.

Capítulo 4

Modelos para Previsão de Perdas

Neste capitulo, analisaremos três HMMs que podem ser usados em conjunto com o algoritmo da Seção 3.2. Mais especificamente, dois desses modelos são propostas da literatura enquanto outro é original a esta dissertação.

4.1 Modelo de Pacotes Individuais

A alternativa mais natural para modelar eventos de perdas de pacotes é através de símbolos binários, i.e., o sucesso na transmissão ou a ocorrência de uma perda. Em [Salamatian e Vaton 2001], um modelo com essas características foi proposto pela primeira vez para simular processos de perdas em uma rede como a Internet. A definição formal deste modelo é totalmente consistente com aquela apresentada na Seção 2.2, fazendo o número de observações M igual a 2.

4.2 Modelo de Perdas Agregadas

Em [Duarte 2003], foi proposta uma outra abordagem para modelagem de perdas com HMMs. Ao invés de modelar cada perda de maneira individual, as observações

deste modelo representam apenas a quantidade de perdas em um conjunto de pacotes.

Seja S o número de pacotes agrupados em cada observação do modelo de perdas agregadas. As observações do modelo de Markov oculto utilizado variam então entre os inteiros de 0 até S. Os parâmetros do modelo são aqueles de um HMM com M igual a $S + 1$ observações, conforme as definições da Seção 2.2.

Para que sejamos mais precisos, é importante ressaltar que o modelo apresentado em [Duarte 2003] utilizava uma cadeia oculta com uma estrutura especial de transições. Entretanto, em nossos experimentos, *não* iremos utilizar essa estrutura especial. Uma vez que a estrutura da cadeia em [Duarte 2003] foi particularmente motivada por padrões periódicos no processo de perdas, em nome da generalidade, consideramos uma cadeia com transições entre *todos* os pares de estados ocultos.

Em termos de complexidade, o modelo de perdas agregadas oferece uma vantagem em relação ao modelo de pacotes individuais. Para o treinamento, é preciso registrar apenas a quantidade de perdas em conjuntos de S transmissões, e não a informação de cada pacote. Este ganho é sensível em dois aspectos. Primeiro, cada *trace* de pacotes pode ser descrito de forma compacta por outro *trace* que registra apenas as perdas agregadas, e portanto, é S vezes menor que a amostra original. Uma vez que a complexidade de tempo do algoritmo *forward-backward* é linear no tamanho da amostra (veja a Seção A.2), a computação é S vezes mais rápida no modelo agregado do que no modelo de pacotes da Seção 4.1. Em segundo lugar, um protocolo de rede implementando o modelo agregado só precisa comunicar as perdas agregadas para permitir a estimação de parâmetros.

Em contrapartida, o modelo agregado possui uma quantidade de parâmetros de observação proporcional a S. Além disso, este modelo não pode ser usado para avaliar outras estatísticas além da quantidade de perdas por intervalo de tempo. Na próxima seção, apresentaremos uma proposta de modelo que possui as vantagens do modelo de pacotes individuais e do modelo de perdas agregadas.

4.3 Modelo de Observações em Lote

Nesta seção, propomos um modelo hierárquico cuja complexidade da estimação de parâmetros é menor que a dos modelos descritos nas seções anteriores. Nossa proposta é baseada na hipótese de que as mudanças nas estatísticas do canal, em uma escala de tempo curta, podem ser aproximadas com um modelo simples. Nós também discutimos como o algoritmo de previsão da Seção 3.2 pode ser adaptado ao modelo proposto.

4.3.1 Definição do Modelo

Suponhamos que as transições entre estados ocultos ocorram apenas a cada S observações. Outra forma de interpretar este modelo é assumir que, uma vez que o processo entra em um estado oculto, ele emite um grupo de S resultados de transmissão de pacotes. Por essa razão, nos referimos a este como um modelo de observações em lote. Claramente, o caso em que $S = 1$ é equivalente ao modelo de pacotes individuais visto na Seção 4.1.

Esse processo pode ser modelado por um HMM, no qual o estado pode emitir um dentre 2^S possíveis símbolos de observação, i.e., um para cada caminho amostral da série de S pacotes. Entretanto, esse modelo seria computacionalmente inviável até para valores moderados de S. Em nossa abordagem, restringimos a distribuição das observações dentro de um estado oculto assumindo que essas são geradas por um modelo de Gilbert simplificado, i.e., uma cadeia de Markov de 2 estados. O raciocínio por trás de nosso modelo é que as correlações de curto prazo podem ser capturadas por um processo simples, enquanto a dinâmica em escalas de tempo maiores é governada pela cadeia de Markov oculta. Ganhos de complexidade são alcançados ao considerar um grupo de S medições como uma única observação, e ao computar a probabilidade conjunta desse grupo a partir da distribuição do processo gerador em cada estado.

Consideramos que as medições de pacotes estão segmentadas em conjuntos de tamanho S. Mais especificamente, o símbolo x_t denota um vetor de medições, $[x_{t,1}, \ldots, x_{t,S}]$, representando o resultado para cada um dos pacotes no t-ésimo grupo. De forma análoga, redefinimos as variáveis das observações, X_t, como vetores das variáveis, $[X_{t,1}, \ldots, X_{t,S}]$.

Para cada estado oculto, i, temos os parâmetros da cadeia de Markov de 2 estados, ilustrados na Figura 4.1, e dados por:

$$r_i = P(X_{t,1} = 1 | Y_t = i); \tag{4.1a}$$

$$p_i = P(X_{t,s} = 1 | X_{t,s-1} = 0,\ Y_t = i), \quad 1 < s \leq S; \tag{4.1b}$$

$$q_i = P(X_{t,s} = 0 | X_{t,s-1} = 1,\ Y_t = i), \quad 1 < s \leq S. \tag{4.1c}$$

Nos referimos ao modelo como a tupla $\lambda = (\pi, \mathbf{A}, \mathbf{r}, \mathbf{p}, \mathbf{q})$, onde $\mathbf{r}$, $\mathbf{p}$, e $\mathbf{q}$ são vetores, contendo os respectivos parâmetros r_i, p_i, q_i, para cada estado, i.

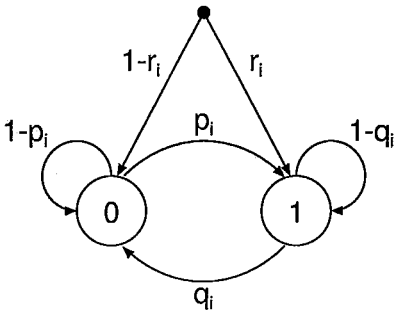


Figura 4.1: Parâmetros da cadeia de Markov de 2 estados para cada estado oculto do modelo proposto.

É importante perceber que, para calcular a probabilidade de uma observação, não é necessário o conhecimento completo da medição de perda para cada pacote. É suficiente manter registro apenas das seguintes estatísticas, em cada grupo de medidas, x_t:

$$x_{t,1} = \text{resultado do primeiro pacote em } x_t; \tag{4.2a}$$

$$S_t^{ij} = \text{número de transições de } i \text{ para } j \text{ em } x_t, \quad i, j \in \{0, 1\}; \tag{4.2b}$$

onde, para S_t^{ij}, é válida a restrição:

$$\sum_{i\in\{0,1\}}\sum_{j\in\{0,1\}} S_t^{ij} = S - 1, \quad \forall t. \tag{4.3}$$

Dada uma instância de x_t, estamos interessados em computar a probabilidade do evento $X_t = x_t$, dado o estado oculto no t-ésimo lote, y_t. Usando as estatísticas definidas acima, temos:

$$b_{y_t,x_t} = \begin{cases} r_{y_t}(p_{y_t})^{S_t^{01}}(1-p_{y_t})^{S_t^{00}}(q_{y_t})^{S_t^{10}}(1-q_{y_t})^{S_t^{11}} & , \quad \text{se } x_{t,1} = 1, \\ (1-r_{y_t})(p_{y_t})^{S_t^{01}}(1-p_{y_t})^{S_t^{00}}(q_{y_t})^{S_t^{10}}(1-q_{y_t})^{S_t^{11}} & , \quad \text{se } x_{t,1} = 0. \end{cases} \tag{4.4}$$

As estatísticas (4.2a–b) podem ser calculadas com um número de operações proporcional ao tamanho da amostra, dado por T. Uma vez calculados, esses valores podem ser usados em conjunto com a Equação (4.4), no procedimento *forward-backward* (veja a Seção A.2). Uma vez que a amostra passa a ser descrita de maneira compacta pelas estatísticas (4.2a–b), o cálculo das variáveis $\alpha_t(i)$ e $\beta_t(i)$ passa a ter complexidade assintótica da ordem de N^2T/S.

4.3.2 Estimação de Parâmetros

De acordo com o Teorema A.2, da Seção A.4, uma vez que restringimos apenas os parâmetros de observação, $\mathbf{B}$, as fórmulas para π e $\mathbf{A}$ permanecerão idênticas àquelas das equações (2.15a) e (2.15b).

Com isso em mente, procedemos nossas derivações da seguinte forma. Aplicando a Equação (4.4) em (2.14c), é possível diferenciar (2.13) em relação a r_i, p_i e q_i para obter as fórmulas correspondentes:

$$r_i = \frac{\sum_{t=1}^{T} \mathbb{I}\{x_{t,1} = 1\}\gamma_t(i)}{\sum_{t=1}^{T}\gamma_t(i)}, \tag{4.5a}$$

$$p_i = \frac{\sum_{t=1}^{T} S_t^{01}\gamma_t(i)}{\sum_{t=1}^{T}(S_t^{01} + S_t^{00})\gamma_t(i)}, \tag{4.5b}$$

$$q_i = \frac{\sum_{t=1}^{T} S_t^{10}\gamma_t(i)}{\sum_{t=1}^{T}(S_t^{10} + S_t^{11})\gamma_t(i)}. \tag{4.5c}$$

Os detalhes intermediários destes resultados podem ser vistos na Seção A.5.

A vantagem computacional de nosso modelo se torna evidente nas equações (4.5a–c), uma vez que o cálculo dessas depende apenas das estatísticas definidas em (4.2a–b). De fato, as métricas em (4.2a–b) são *estatísticas suficientes* para todos os parâmetros do modelo (veja a Seção A.6 para a definição formal). Uma vez que as variáveis *forward* e *backward* dependem apenas de (4.5a–c), uma aplicação que estima parâmetros para o modelo de observações em lote precisa registrar apenas estas estatísticas.

Cada iteração do procedimento de treinamento é mais rápida por um fator de S, como no modelo de perdas agregadas, descrito na seção anterior. Ainda assim, diferentemente do modelo agregado, o número de parâmetros de observação por estado oculto em nosso modelo é independente do tamanho do lote, S. Uma vez que as medições usadas no treinamento são geralmente realizadas no receptor dos pacotes e precisam ser enviadas de volta ao transmissor, também há uma economia no tamanho das mensagens de controle que precisam ser enviadas.

É fácil perceber que as equações (4.5a–c) podem ser interpretadas como razões entre valores esperados, de acordo com as expressões:

$$r_i = \frac{E[\#\text{ de grupos começando com 1 no estado } i]}{E[\#\text{ de grupos começando no estado } i]}, \tag{4.6a}$$

$$p_i = \frac{E[\#\text{ de transições de 0 para 1 no estado } i]}{E[\#\text{ de transições de 0 no estado } i]}, \tag{4.6b}$$

$$q_i = \frac{E[\#\text{ de transições de 1 para 0 no estado } i]}{E[\#\text{ de transições de 1 no estado } i]}. \tag{4.6c}$$

4.3.3 Extensão do Algoritmo de Previsão

Nesta seção, mostramos como calcular a distribuição da variável R_t^f, definida na Seção 3.2, para o modelo de observações em lote. Restringimos nossa análise ao caso mais simples onde f é um múltiplo do tamanho do grupo S, i.e., $f = f'S$

para algum valor inteiro f'. O caso geral é dispendioso em notação e é omitido por brevidade. Separando o número de perdas de cada lote de observações, a Equação (3.1) pode ser reescrita como:

$$R_t^f = \sum_{i=1}^{f'} \sum_{j=1}^{S} X_{t+i-1,j}. \tag{4.7}$$

Note que, entre duas transições do estado oculto, o número de perdas no em um grupo é totalmente determinado pelos parâmetros do estado atual. Podemos utilizar esse fato para construir um modelo de perdas agregadas, como o da Seção 4.2, que conta o número de perdas de pacotes em S transmissões. Um vez que a matriz $\mathbf{B}$ é avaliada para esse modelo agregado, resta apenas aplicar o algoritmo dado pela Equação (3.12) para obter a distribuição de R_t^f.

Cada linha, $\mathbf{b}_i$, da matriz de observação, $\mathbf{B}$, é definida como a distribuição de probabilidade para o número de perdas em um lote de tamanho S, assumindo que o modelo se encontra no estado i. É interessante notar que, para avaliar $\mathbf{b}_i$ neste caso, podemos simplesmente aplicar o mesmo algoritmo dado pela Equação (3.12) sobre os parâmetros da cadeia de Markov de 2 estados contida no estado i.

4.4 Custos Computacionais

Comparamos os custos computacionais da tarefa de previsão adaptativa quando realizada por cada um dos três HMMs que discutimos até então. Especificamente, iremos avaliar a eficiência da metodologia descrita na Seção 3.1, quando aplicada aos modelos das três seções anteriores. Na comparação que segue, consideramos modelos com N estados ocultos.

4.4.1 Estimação de Parâmetros

A respeito do passo de treinamento do modelo, cada iteração do procedimento de para HMMs tem a sua complexidade dominada pela recursão *forward-backward*, conforme pode ser visto na Seção A.2. Para uma amostra de treino de tamanho T, o modelo de pacotes individuais realiza um número de operações aritméticas da ordem de N^2T.

Nos modelos de perdas agregadas e de observações em lote, a amostra de treino pode ser descrita de forma compacta através de estatísticas suficientes para os parâmetros. No caso do modelo de perdas agregadas, essa estatística é o total de perdas em um conjunto de S pacotes, enquanto que para o modelo de observações em lote, é preciso manter registro das estatísticas dadas pelas equações (4.2a–b). Por isso, a complexidade dos passos *forward-backward* para uma amostra de tamanho T é reduzida por um fator de S nesses modelos. Além disso, em um mecanismo de rede implementando esses modelos, apenas essas estatísticas suficientes precisam ser enviadas de volta a fonte, reduzindo assim a quantidade de informação de controle transmitida na rede.

Por último, nota-se que, nos modelos de pacotes individuais e de observações em lote, o número de parâmetros de observação em cada estado oculto é constante, enquanto, no modelo de perdas agregadas, esta quantidade é linear no tamanho do grupo de pacotes, S.

4.4.2 Previsão de Taxas de Perda

Para a tarefa previsão de taxas de perda, é preciso, primeiramente, notar que o passo de inicialização do algoritmo em (3.12) faz uso da recursão *forward* para avaliar $\pi_{t,h}$. Portanto, as mesmas conclusões tecidas a respeito do procedimento de treinamento permanecem válidas. Entretanto, na prática, o laço principal será o gargalo da computação de (3.12).

Para os experimentos que apresentaremos no Capítulo 5, a ordem de magnitude do número de operações realizadas por cada modelo, no laço principal de (3.12), é exibida na Tabela 4.1. Os modelos de perdas agregadas e observações em lote executam três vezes menos operações que o modelo de pacotes individuais. Este é um ganho significativo para um preditor de tempo real.

Modelo	Operações realizadas
modelo de pacotes individuais	9387200
modelo de perdas agregadas modelo de observações em lote	2595800

Tabela 4.1: Número aproximado de operações realizadas por cada modelo no laço principal do algoritmo proposto para previsão de perdas.

Capítulo 5

Resultados Experimentais

Realizamos experimentos para medir a acurácia dos modelos apresentados no Capítulo 4 na previsão de taxas de perdas em *traces* de transmissões na Internet. Na Seção 5.1, apresentamos estes *traces* e assinalamos suas características principais. Em seguida, apresentamos os parâmetros do experimento de previsão, além das métricas utilizadas para comparar os diferentes modelos. Nas seções restantes, apresentamos resultados qualitativos e quantitativos para comparar diferentes modelos e diferentes estratégias de previsão.

5.1 Medições de Perdas de Pacotes

Nos experimentos realizados neste trabalho, tivemos à nossa disposição um conjunto extenso de medições fim-a-fim realizadas entre 4 instituições acadêmicas — duas dessas localizadas no Brasil e outros duas nos Estados Unidos. Mais especificamente, os pontos de geração e coleta se encontram nas Universidades Federais do Rio de Janeiro (UFRJ) e de Minas Gerais (UFMG), além das Universidades de Maryland (UMD) e de Massachssets em Amherst (UMass). As combinações origem-destino utilizadas em nossas medições estão ilustradas na Figura 5.1. Essas medições exibem uma grande variedade de situações da rede, desde horas sem

nenhuma perda até totais interrupções no serviço dos canais intermediários. Em todas as medições realizadas, tráfego de taxa constante foi gerado, utilizando as ferramentas do ambiente Tangram-II, [de Souza e Silva e Leão 2000].

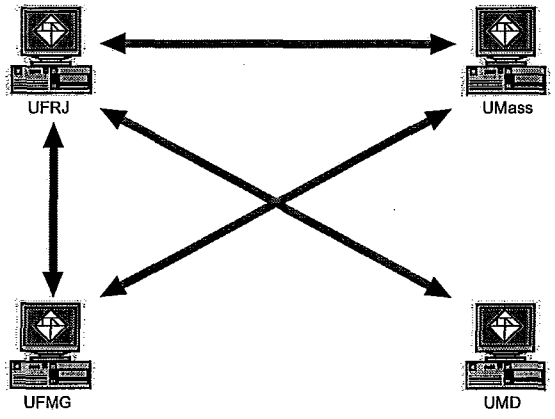


Figura 5.1: Pontos de geração e coleta de tráfego usados em nosso experimento.

Cada sessão de geração de tráfego durou uma hora e um total de 998 sessões foram realizadas em diferentes períodos dos anos de 2001, 2002 e 2004. Em cada dia de experimentos, as sessões foram conduzidas em 3 horários diferentes, geralmente localizados em torno dos períodos de pico de utilização de grande parte dos canais intermediários, levando em consideração as diferenças de relógios entre os pontos extremos.

O padrão de tráfego foi escolhido para emular o comportamento de uma aplicação de Voz sobre IP (VoIP) simplificada, conforme ilustrado na Figura 5.2. A cada 20 milissegundos, a fonte envia um pacote UDP contendo 324 *bytes* de dados de aplicação — 160 amostras de áudio com 2 *bytes* cada — somados a um cabeçalho de controle com 4 *bytes*. Considerando os efeitos dos cabeçalhos das camadas de transporte e rede, a carga total oferecida é de 140.8 kbps.

Cada pacote enviado pela rede é marcado com um número de série e aqueles que são entregues com sucesso são registrados pelo destinatário em um arquivo junto com outras informações relevantes. Utilizamos estes *traces* para produzir uma seqüência binária $\{x_i\}_{i=1}^{T}$, onde x_i é definido como 0 se o pacote com o i-ésimo número de série

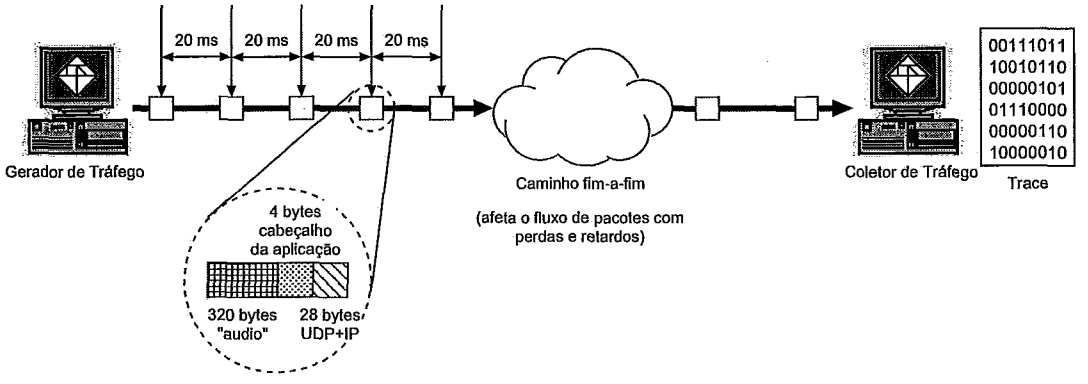


Figura 5.2: Esquema de geração de tráfego emulando uma aplicação VoIP simplificada.

chegou ao destino intacto, ou 1 caso contrário.

Muitos dos 998 *traces* coletados exibem processos de perda que não são interessantes aos nossos propósitos experimentais. Esses incluem estatísticas que são simples demais para previsão, como medições de taxas de perda muito baixas.

A Figura 5.3 mostra as frações de perdas em cada um dos 998 *traces* coletados, ordenados de forma não-decrescente para facilitar a visualização. Embora a média das frações de perdas em todos os *traces* seja de 3.5%, é possível perceber que 67% dos *traces* possui menos que 1% de perdas. Por outro lado, cerca de 4% dos *traces* apresentam mais que 30% dos pacotes perdidos.

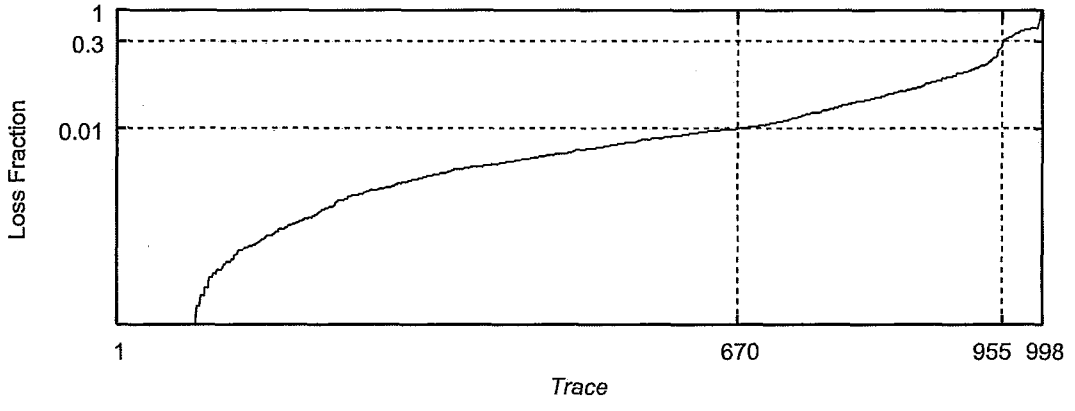


Figura 5.3: Frações de perdas dos 998 *traces* coletados.

Além disso, embora o tamanho médio das rajadas de perda seja de apenas 1.72 pacotes, a Figura 5.4 mostra que cerca de 10% dos *traces* possuem períodos de perdas consecutivas que duram pelo menos 30 segundos.

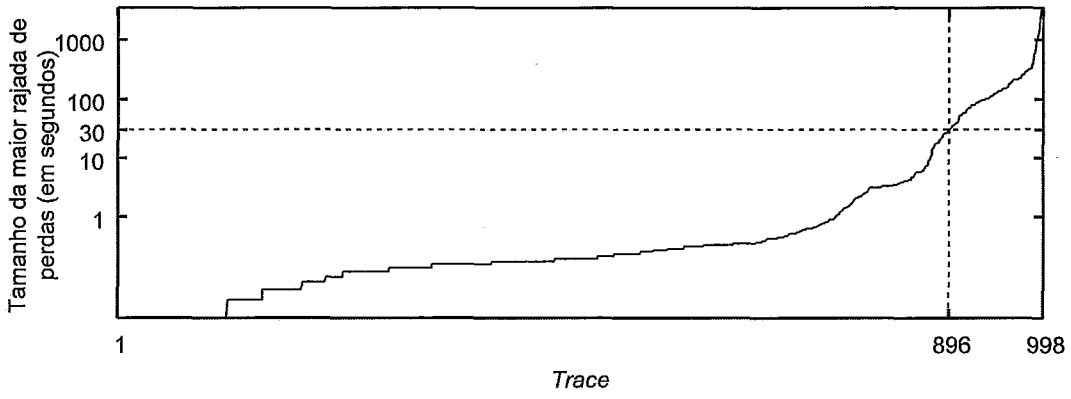


Figura 5.4: Tamanhos das maiores rajadas de perda em cada um dos 998 *traces*.

Em nossos experimentos, selecionamos apenas 194 *traces*, cujas frações de perda estão entre 1% e 30%, e que contém períodos de perdas consecutivas que não ultrapassam 30 segundos. Entre os *traces* selecionados, a fração de perdas média é de 3.8% e o tamanho médio das rajadas é de 1.68 pacotes. A Figura 5.5 mostra as frações de perdas em cada um destes 194 *traces* ordenadas de forma crescente. Nas próximas seções, apresentaremos resultados quantitativos e qualitativos na previsão de taxas de perdas de pacotes para esses 194 *traces*.

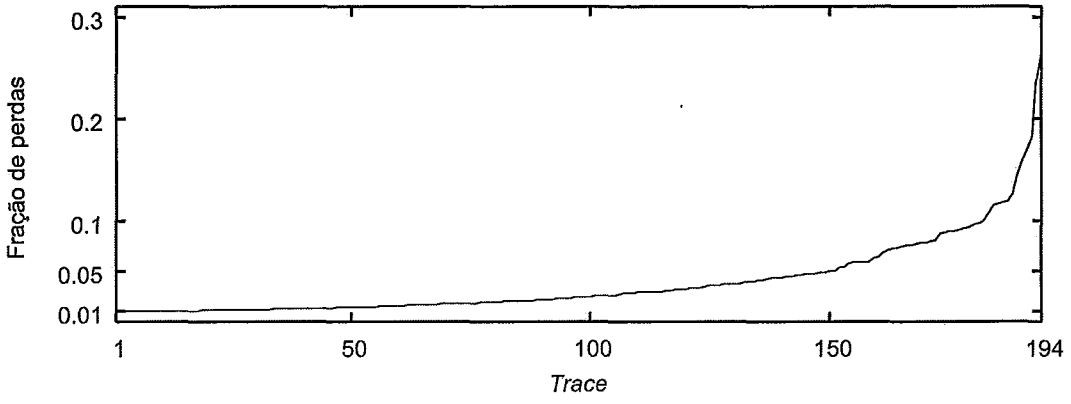


Figura 5.5: Frações de perdas dos 194 *traces* selecionados para análise.

5.2 Modelos e Parâmetros Utilizados

Consideramos três modelos diferentes em nossos experimentos. O primeiro desses é o modelo de pacotes individuais, da Seção 4.1, ao qual nos referimos, por brevidade, como HMM-Pacote. O segundo HMM, que consideramos, é o modelo de perdas agregadas, visto na Seção 4.2, e aqui referenciado como HMM-Agregado. No modelo agregado, agrupamos $S = 50$ pacotes. O terceiro modelo usado é o modelo de observações em lote, visto na Seção 4.3, onde também agrupamos $S = 50$ pacotes por lote. Nos referimos a esse modelo como o HMM-Lote. É importante reparar que, pelas especificações de tráfego, descritas na Seção 5.1, tanto o HMM-Agregado, quanto o HMM-Lote realizam transições entre estados ocultos a cada a 1 segundo de transmissões.

Para cada um dos modelos que temos em consideração, é necessário ainda especificar a quantidade de estados da cadeia de Markov oculta associada. No trabalho de [Salamatian e Vaton 2001], é argumentado que até 4 estados são suficientes para caracterizar *traces* de perdas de pacotes, em canais de comunicação fim-a-fim. Neste trabalho, não estamos interessados em avaliar a sensibilidade da tarefa de previsão ao número de estados no HMM, e por essa razão, todos os modelos considerados em nossos experimentos têm 10 estados ocultos. Acreditamos que seja possível obter resultados qualitativamente razoáveis com quantidades inferiores de estados. Entretanto, deixamos a análise dessa quantidade ótima de estados ocultos para um trabalho futuro, cujo escopo esteja mais próximo da aplicação final.

Para todos os experimentos, aplicamos a metodologia da Seção 3.1. Os parâmetros dos modelos são re-estimados a cada $\tau = 3$ minutos, usando a informação dos últimos $T = 3$ minutos. Justificamos esse intervalo entre treinamentos de maneira empírica, conforme discutiremos na Seção 5.4.3. Para cada modelo, o treinamento, usando as fórmulas do algoritmo Baum-Welch, é realizado por no máximo 1000 iterações, ou até que a diferença de verossimilhança do *trace* em relação ao valor da iteração anterior seja inferior a 10^{-5}.

Seguindo a nossa metodologia, realizamos a previsão da fração de perdas entre os pacotes transmitidos nos próximos $F = 5$ segundos, dados os resultados das perdas nos últimos $H = 10$ segundos. Estas estimativas de previsão são atualizadas a cada $\psi = 5$ segundos. Consideramos esse valor apropriado para o tipo de tráfego de nossos experimentos, que emula uma aplicação para transmissão de voz sobre IP.

Além disso, um intervalo de 5 segundos é o mínimo recomendado entre o envio de pacotes RTCP (*Real-time Transport Control Protocol*), como pode ser visto em [Schulzrinne et al. 2003]. O RTCP foi concebido para permitir o envio de estatísticas de qualidade de serviço (QoS - *quality of service*), do receptor de um fluxo multimídia de volta para o seu transmissor. Nossa metodologia, que depende da coleta de estatísticas para as tarefas de treinamento e previsão, poderia ser implementada, na prática, com o auxílio do RTCP.

5.3 Métricas de Acurácia de Previsão

Para quantificar a precisão das diferentes estratégias de previsão, utilizamos três métricas principais. Reconhecemos que avaliar a acurácia de uma previsão é, por si só, uma tarefa desafiadora. De fato, não é possível definir uma métrica universal, com a qual seja possível observar todas as nuances existentes no que diz respeito a qualidade de previsão.

A primeira métrica que consideramos é o erro médio quadrático (MSE - *mean squared error*), calculado entre os valores previstos e as medições reais. Formalmente, se durante o intervalo de previsão i, a fração de perdas observada é r_i e o valor previsto é p_i, então, para um *trace* com n previsões, o MSE é definido como:

$$e = \frac{\sum_{i=1}^{n} (r_i - p_i)^2}{n} \tag{5.1}$$

Uma outra métrica de nosso interesse é a *correlação cruzada amostral*, calculada entre as medições reais e as previsões. A partir da mesma notação usada na definição anterior do MSE, sejam $\bar{r}$ e $\bar{p}$, as respectivas médias amostrais das medições e das

previsões, isto é:

$$\bar{r} = \frac{\sum_{i=1}^{n} r_i}{n}, \tag{5.2}$$

$$\bar{p} = \frac{\sum_{i=1}^{n} p_i}{n}. \tag{5.3}$$

Então, a correlação cruzada amostral é definida como o coeficiente:

$$c = \frac{\sum_{i=1}^{n} (r_i - \bar{r})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{n} (r_i - \bar{r})^2 \sum_{i=1}^{n} (p_i - \bar{p})^2}} \tag{5.4}$$

Este valor deve ser o maior possível para indicar um boa previsão. Em outras palavras, queremos determinar se as variações na métrica de previsão estão de fato acompanhando as mudanças nas estatísticas do canal, ou se a acurácia obtida é meramente fruto de coincidências aleatórias. Através dessa métrica podemos identificar os *traces* em que uma determinada estratégia de previsão é bem sucedida.

Por outro lado, é fácil perceber, por inspeção da Equação 5.4, que a correlação amostral é normalizada pelo desvio padrão amostral das previsões. Por essa razão, não seria inteiramente correto comparar dois modelos diferentes, prevendo um mesmo *trace* de amostras reais. Para comparar diferentes modelos, damos preferência a medida de *covariância cruzada amostral* que, por sua vez, não é normalizada entre diferentes *traces*. A covariância cruzada amostral é dada pela fórmula:

$$v = \frac{\sum_{i=1}^{n} (r_i - \bar{r})(p_i - \bar{p})}{n-1}, \tag{5.5}$$

onde o denominador é, de fato, $n-1$ ao invés de n, para garantir que o estimador não seja tendencioso, [Ross 1997].

5.4 Resultados Preliminares

Nossa primeira análise tem como objetivo salientar dois resultados, que foram observados a respeito do algoritmo proposto na Seção 3.2. Em primeiro lugar, o algoritmo que propomos é consistentemente melhor em estimar as taxas de perda no curto prazo do que um preditor de estado estacionário. Em segundo, um modelo,

cujas transições entre estados ocultos ocorrem na mesma escala de tempo que uma transmissão individual de pacote, não tem um desempenho tão bom quanto um no qual estas transições ocorrem em uma escala de tempo convenientemente maior.

5.4.1 Medidas Transientes e Estacionárias

Além de aplicar o algoritmo da Equação (3.12), nós também utilizamos a probabilidade de perda em estado estacionário a partir dos parâmetros estimados no modelo. Se denotarmos como π^* o vetor de probabilidades de estado estacionário da cadeia de Markov oculta, o preditor de estado estacionário pode ser avaliado como:

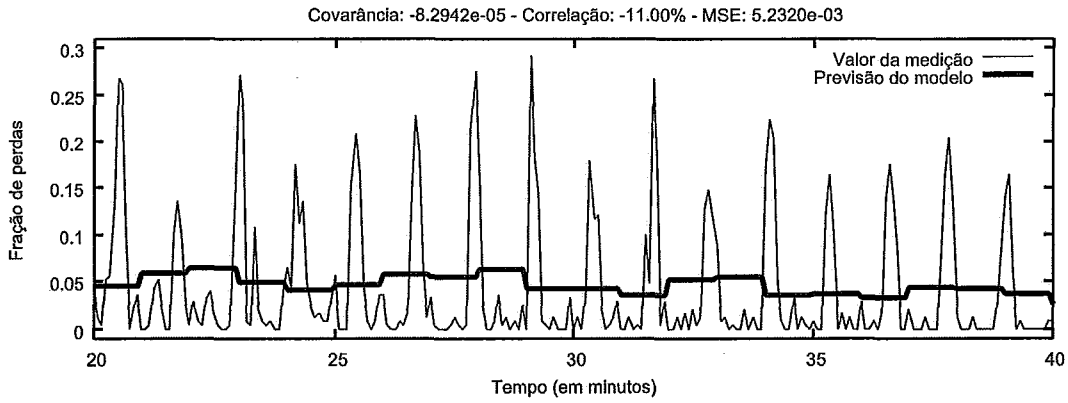
$$l = \sum_{i=1}^{N} \pi_i^* \sum_{s=0}^{S} s b_{is}, \tag{5.6}$$

onde b_{is} é a probabilidade de s perdas entre duas transições da cadeia oculta. Nos casos do HMM-Pacote e do HMM-Agregado, esses são meramente os parâmetros de observação. No HMM-Lote, por outro lado, essa é a distribuição do número de perdas em S observações da cadeia de Markov de 2 estados associada a cada estado oculto, i, conforme discutido na Seção 4.3.3.

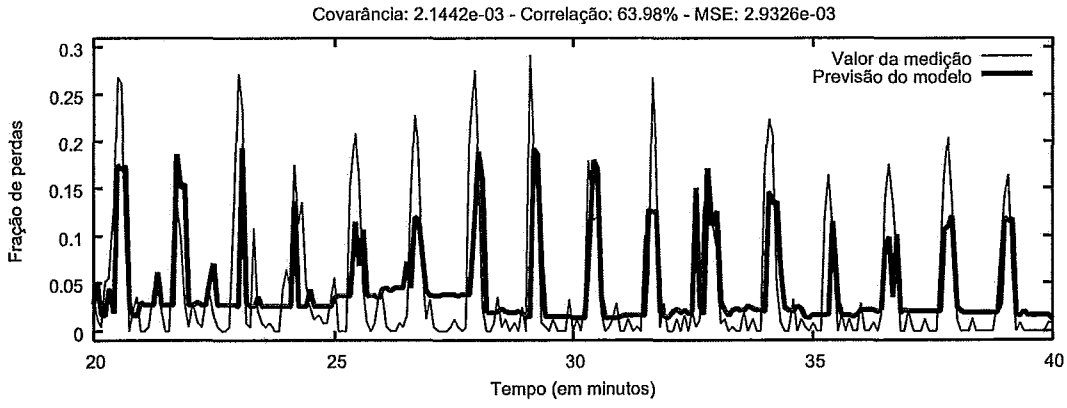
Nós concluímos que, na maioria dos *traces*, nosso algoritmo é consistentemente melhor que a alternativa de estado estacionário, acompanhando as flutuações na fração de perdas de curto prazo. Isso pode ser explicado da seguinte forma: a medida estacionária simplesmente ignora as variações nas estatísticas recentes, e prevê o mesmo resultado para cada época de treinamento. Embora o preditor de estado estacionário possibilite um erro médio quadrático pequeno em muitos *traces*, nosso algoritmo é capaz de capturar vários padrões de curta duração no processo de taxas de perda, tais como picos periódicos, possivelmente causados por mudanças de roteamento, [Zhang et al. 2000].

Para ilustrar essas idéias, a Figura 5.6 mostra os *traces* das previsões da métrica estacionária e do algoritmo de previsão transiente em um segmento de 20 minutos de taxas de perdas reais em um *trace* que exibe periodicidade. É possível observar

que o *trace* da métrica estacionária, na Figura 5.6(a), tem um formato de "degraus", uma vez que as previsões mudam apenas quando os parâmetros do modelo são re-estimados, a cada $\tau = 3$ minutos. Por outro lado, o HMM-Pacote, na Figura 5.6(b), usando nosso algoritmo proposto, pôde prever bem o padrão de perdas periódicas. Repare que a covariância do preditor estacionário é muito próxima de zero, embora o seu MSE seja razoavelmente baixo. O preditor transiente, em contraste, possui um MSE menor, e uma correlação de 63.98%.



(a) HMM-Pacote usando o preditor de estado estacionário



(b) HMM-Pacote usando o preditor transiente do algoritmo proposto

Figura 5.6: Medida de estado estacionário versus o algoritmo proposto para um segmento de 20 minutos de um *trace* com características periódicas.

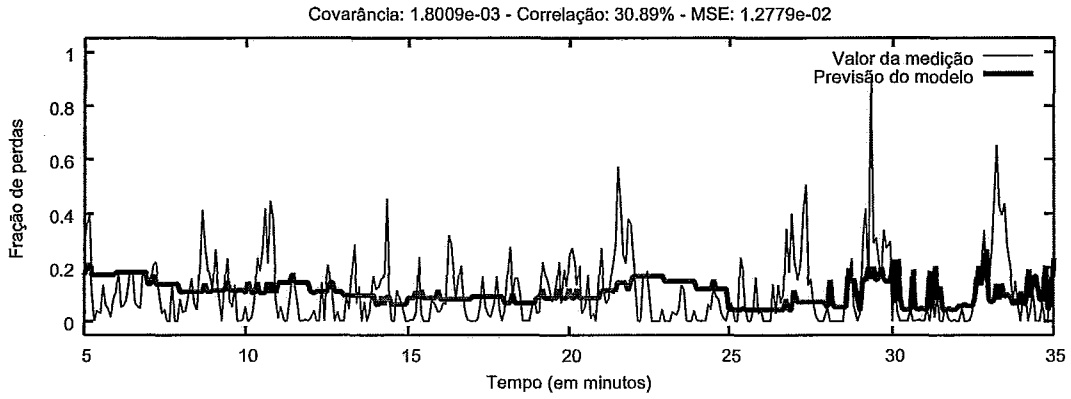
5.4.2 Escala de Tempo das Transições de Estado Oculto

Os gráficos na Figura 5.7 comparam o desempenho das previsões geradas pelos três modelos de Markov ocultos discutidos, usando nosso algoritmo proposto, em 30 minutos de uma outra amostra de *trace*. Embora este *trace* tenha um padrão de rajadas intenso, ele claramente não é periódico como o da Figura 5.6. Pode ser notado, que o HMM-Pacote, na Figura 5.7(a), apresentou um comportamento similar ao do preditor de estado estacionário, na Figura 5.6(a), ignorando muitas das variações na medida, que estão presentes no *trace* real.

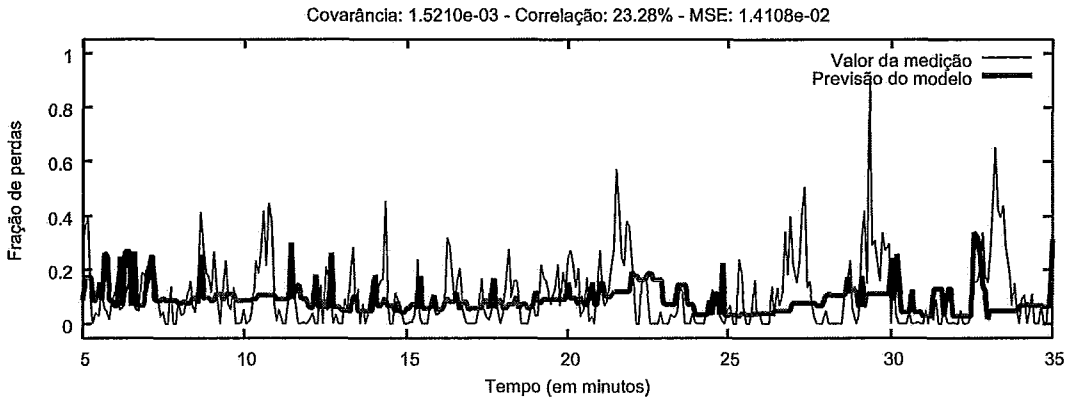
Por outro lado, embora o HMM-Agregado, na Figura 5.7(b), tenha gerado variações mais amplas nas suas previsões, essas erraram mais freqüentemente que as do HMM-Pacote. Finalmente, o HMM-Lote, na Figura 5.7(c), não apenas reproduziu as variações nas medidas, mas também foi mais preciso que os outros modelos, tanto de acordo com o MSE quanto pela covariância cruzada amostral.

Atribuímos a desvantagem do HMM-Pacote ao fato de que, nesse modelo, o estado oculto — junto com as estatísticas do canal — pode mudar a cada pacote transmitido. Como conseqüência, o comportamento de estado estacionário é alcançado muito mais rápido que para os modelos agregado e lote, onde as transições dos estados ocultos ocorrem em uma escala de tempo maior (neste caso, a cada 1 segundo, ou 50 pacotes).

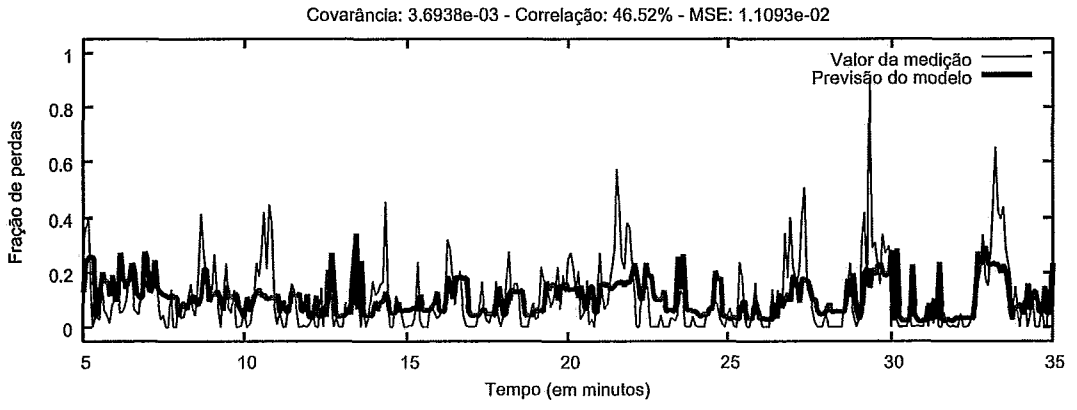
O HMM-Agregado, por outro lado, apesar de produzir variações transientes nas suas previsões, não conseguiu ser mais preciso que o HMM-Pacote, para este *trace* em particular. Uma possível razão para esse fenômeno é o fato de que, neste modelo, a quantidade de parâmetros a serem ajustados, a cada treinamento, é muito maior do que nos modelos pacote e lote. Uma quantidade maior de parâmetros livres, embora possa parecer vantajosa, também torna mais lenta a convergência para o seus valores ótimos. Uma vez que truncamos a execução do algoritmo Baum-Welch a, no máximo, 1000 iterações, um modelo com mais graus de liberdade pode ter dificuldades em ajustar seus parâmetros da maneira ideal.



(a) HMM-Pacote usando o algoritmo proposto



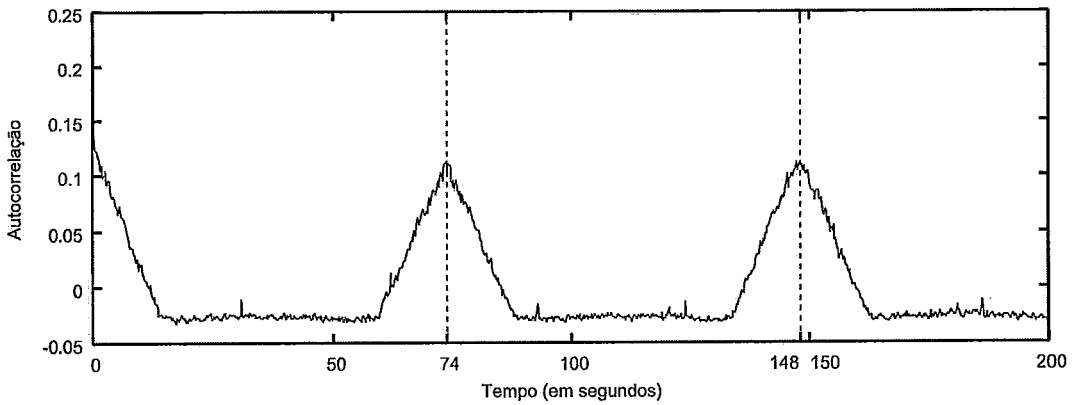
(b) HMM-Agregado usando o algoritmo proposto



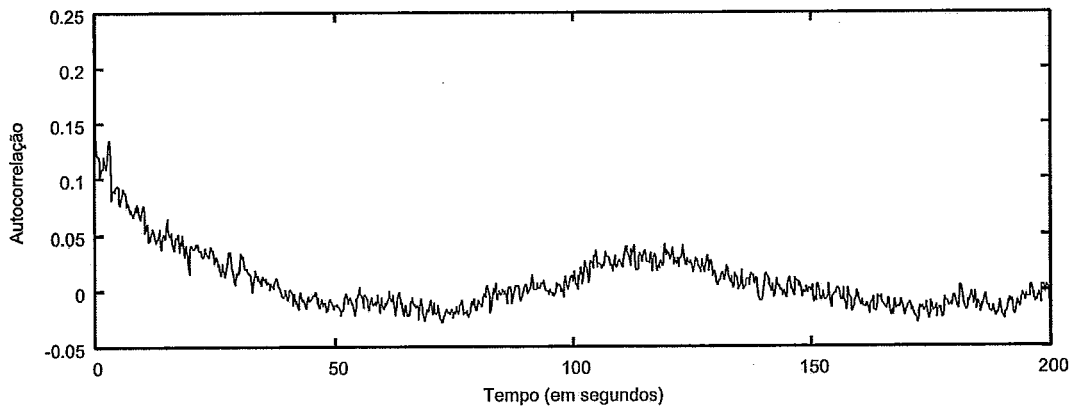
(c) HMM-Lote usando o algoritmo proposto

Figura 5.7: Resultados das previsões em um segmento de 30 minutos de um *trace*.

É importante enfatizar que o *trace* analisado na Figura 5.7 é consideravelmente mais complexo, para a tarefa de previsão, do que o *trace* da Figura 5.6, que é fortemente periódico. A Figura 5.8 mostra a autocorrelação amostral calculada para ambos os *traces*. É possível perceber que, embora o *trace* na Figura 5.8(b) possua correlações significativas, devido as suas perdas em longas rajadas, essas não são tão marcantes quanto as do *trace* na Figura 5.8(a), onde o padrão periódico, a cada 74 segundos, é claro e evidente.



(a) Autocorrelação do *trace* na Figura 5.6.



(b) Autocorrelação do *trace* na Figura 5.7.

Figura 5.8: Autocorrelações amostrais medidas para os dois *traces* anteriores.

5.4.3 Efeitos do Intervalo de Treinamento do Modelo

Finalmente, estudamos brevemente a sensibilidade da previsão ao parâmetro τ. Por exemplo, se τ é escolhido como 1 minuto, o HMM-Pacote é capaz de atualizar as suas previsões tão rápido quanto os outros modelos com $\tau = 3$ minutos. Entretanto, enquanto os modelos agregado e lote adaptam suas previsões nestes três minutos baseados apenas no algoritmo de previsão, a melhora do HMM-Pacote é devida ao fato de que seus parâmetros estão sendo atualizados mais freqüentemente que antes. Por outro lado, se τ é aumentado para 5 minutos, então o HMM-Agregado e o HMM-Lote se tornam menos precisos que para $\tau = 3$ minutos, sempre que ocorrem mudanças bruscas na taxa de perda.

5.5 Resultados Adicionais

Esta seção tem como objetivo avaliar o desempenho de nosso algoritmo de previsão, e do modelo de observações em lote, para o conjunto completo de 194 *traces*. Como na seção anterior, começaremos por mostrar a vantagem do nosso algoritmo sobre um preditor de estado estacionário. Em seguida, apresentaremos resultados que comparam os três HMMs utilizando nosso algoritmo.

5.5.1 O Algoritmo Proposto e Preditor Estacionário

Assim como na Seção 5.4.1, comparamos o desempenho de nosso algoritmo com a estratégia que utiliza apenas a fração de perda em estado estacionário como valor da previsão. A Figura 5.9 mostra as covariâncias obtidas, usando a Equação (5.5), para os 194 *traces*. Os *traces* foram arbitrariamente ordenados no eixo x de modo que a curva correspondente ao HMM-Pacote usando nosso algoritmo proposto seja não-decrescente.

A figura mostra que para a maioria dos *traces*, o algoritmo proposto obtém uma

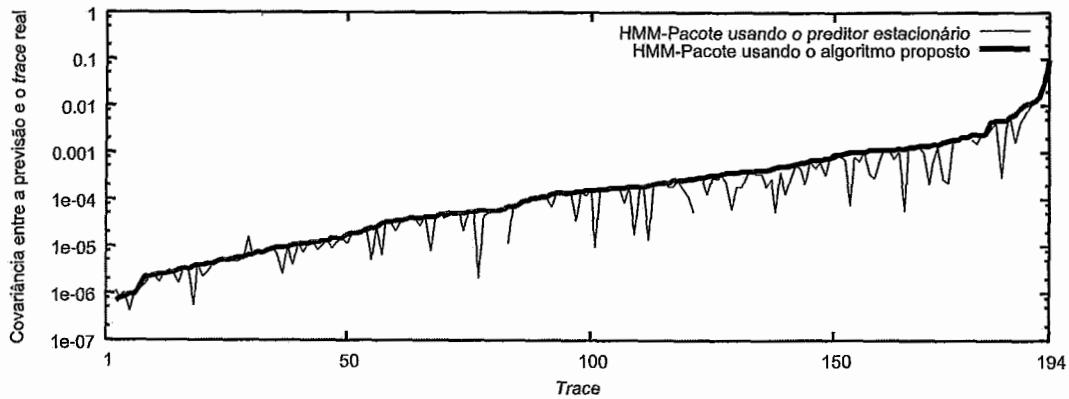


Figura 5.9: Covariâncias amostrais entre previsões e taxas de perda reais nos 194 *traces*, para o algoritmo proposto e o preditor estacionário.

previsão que está mais correlacionada com as amostras reais que aquelas da medida estacionária. Isso ocorre, de fato, em 84% dos *traces*.

5.5.2 Comparação entre HMMs

A Figura 5.10 mostra as covariâncias de cada um dos três HMMs, usando nosso algoritmo, para os 194 *traces*. Cada curva no gráfico corresponde a um dos modelos em comparação. Os *traces* estão ordenados de modo que a curva correspondente ao HMM-Lote seja não-decrescente.

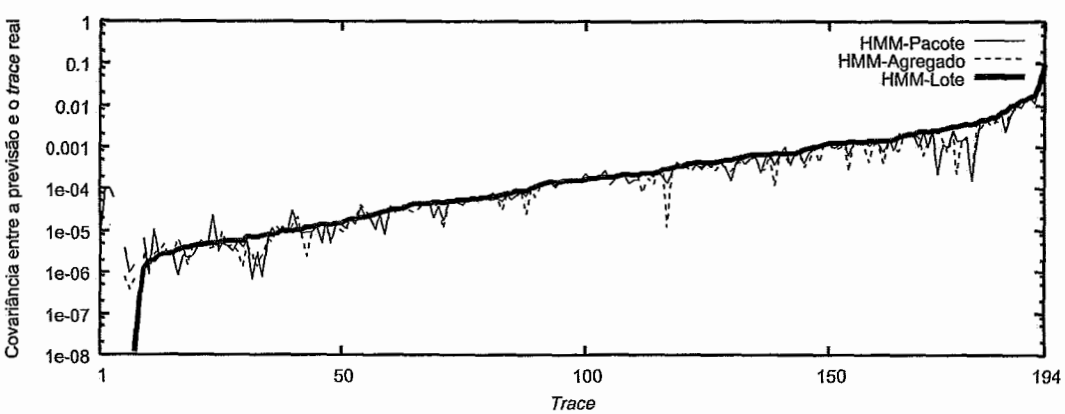


Figura 5.10: Covariâncias amostrais entre previsões e taxas de perda reais nos 194 *traces*, para os três HMMs.

É evidente na Figura 5.10 que, na maior parte do tempo, as covariâncias do HMM-Pacote e do HMM-Agregado estão abaixo da curva de referência do HMM-Lote. De fato, o HMM-Lote provê a covariância mais alta em 71% dos *traces*, enquanto os modelos pacote e agregado o fazem, respectivamente, em 17% e 12% dos casos.

Por outro lado, para comparar o desempenho do HMM-Lote entre os diferentes cenários, a Figura 5.11 mostra que 69% dos *traces* foram previstos com pelo menos 20% de correlação, uma quantia significativa para descartar a hipótese de que os acertos ocorrem por mera coincidência.

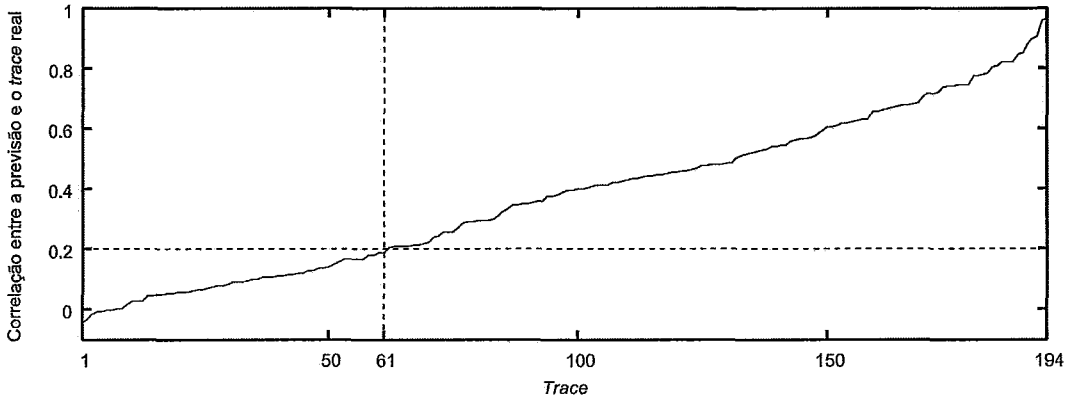


Figura 5.11: Correlações amostrais entre previsões e taxas de perda reais nos 194 *traces* para o modelo de observações em lote.

5.5.3 Relação com o Preditor Linear

Da Seção 2.3, lembramos que o erro do preditor linear de primeira ordem ótimo é limitado entre d e $2d$, onde $2d = 2(E[r_t^2] - E[r_{t-1}r_t])$ é o erro da estratégia replicadora. Identificamos que a diferença d entre as correlações produtos, $E[r_t^2]$ e $E[r_{t-1}r_t]$, tem um impacto direto na acurácia do preditor linear de primeira ordem. Por essa razão, quantificamos o impacto dessa medida nos erros do HMM-Lote usando nosso algoritmo de previsão para taxas de perda.

A Figura 5.12 mostra os erros médios quadráticos obtidos do HMM-Lote, no eixo

vertical, contra a diferença d, no eixo horizontal. Cada ponto no gráfico corresponde a um dos 194 *traces*.

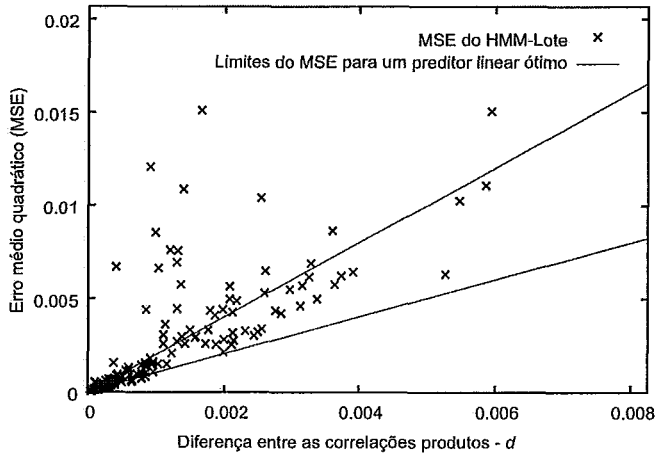


Figura 5.12: MSE das previsões do HMM-Lote para os 194 *traces*.

Conforme mostrado na figura, a grande maioria dos pontos cai entre os limites do preditor ótimo. De fato, isso ocorre em 148 (76.29%) do *traces*. Nesses cenários, o HMM-Lote provê um preditor mais preciso em termos do MSE que a estratégia replicadora, denotada pelo limite superior no gráfico. Nos 46 (23.71%) *traces* restantes, o HMM-Lote não conseguiu estimar propriamente as correlações produtos. É importante lembrar que, de acordo com o Teorema 3.1, essas métricas formam uma condição suficiente para um preditor linear ótimo. Do gráfico, é fácil ver que isso se torna mais comum conforme a métrica $d = E[r_t^2] - E[r_{t-1}r_t]$ aumenta.

Explicamos os erros de previsão do HMM-Lote por duas causas: (*a*) valores grandes de d correspondem a valores pequenos da função de autocorrelação na sua primeira *lag*, indicando que as perdas de pacotes nesses cenários são, em maior parte, não correlacionadas; (*b*) variações altas nas taxas de perdas medidas no curto prazo. Quando esses dois fatores se combinam, é de se esperar que um modelo tenha dificuldades tentando capturar as correlações produtos necessárias para realizar previsões precisas, e portanto, a estratégia replicadora, que é mais conservadora em suas previsões, irá potencialmente ter erros menores.

Modelo	MSE	*Precision-Recall* — limite de 3%			
		Perdas baixas		Perdas altas	
		Precision	*Recall*	*Precision*	*Recall*
HMM-Pacote-EE	0.00271	82.65%	83.96%	72.29%	70.36%
HMM-Pacote	0.00217	84.77%	84.53%	72.77%	73.14%
HMM-Agregado	0.00241	85.43%	84.01%	71.47%	73.65%
HMM-Lote	0.00213	85.95%	84.64%	72.63%	74.66%
Replicador	0.00195	82.11%	82.03%	68.44%	68.55%

Tabela 5.1: Resultados para todos os modelos nos 194 *traces*.

5.6 Resultados Gerais

A Tabela 5.1 reúne, para cada modelo, resultados agregados sobre as previsões contidas nos 194 *traces*. É possível ver que a estratégia replicadora tem um pequeno MSE geral, mas como argumentamos na Seção 5.5.3, isso é causado por cenários de taxas de perdas com variações altas e imprevisíveis.

Outros resultados que exibimos na Tabela 5.1 são os valores *Precision-Recall*. Essas métricas se aplicam a preditores binários, e determinam o quão bom é um dado preditor em determinar se uma condição será ou não satisfeita no futuro. Em suma, o valor *Precision* de uma técnica para prever uma condição x é definido como a fração do tempo em que um preditor pode, corretamente, prever um evento dado que esse de fato ocorre. Por outro lado, a medida de *Recall* é definida como a fração das previsões positivas de um evento, que estão, de fato, corretas.

Em nossos experimentos, consideramos a habilidade de cada modelo em determinar se a fração de perdas, no próximo intervalo de previsão, estará acima de um limite escolhido (neste exemplo, $\theta = 3\%$). Portanto, os valores *Precision* e *Recall* mostrados na Tabela 5.1 se referem às condições de que a fração de perdas seja considerada *baixa* ($\leq 3\%$) ou *alta* ($> 3\%$). Em nosso conjunto de 194 *traces*, 36.3% de todas as medições de taxas de perda, em intervalos de 5 segundos, estão acima

do limite de 3%.

Em nome do didatismo, a Figura 5.13 ilustra as 4 situações nas quais uma previsão pode ser classificada, em nosso experimento. Consideramos um Acerto($\leq$), quando um preditor determina, de forma correta, que a fração de perdas será abaixo do limite de degradação, θ. De maneira correspondente, um Erro($\leq$) está associado às previsões incorretas de que fração de perdas será maior que θ quando, na realidade, ela está abaixo do limiar. De forma complementar às duas medidas anteriores, também definimos o Acerto($>$) e o Erro($>$), em função do evento de que a fração de perdas real é considerada alta.

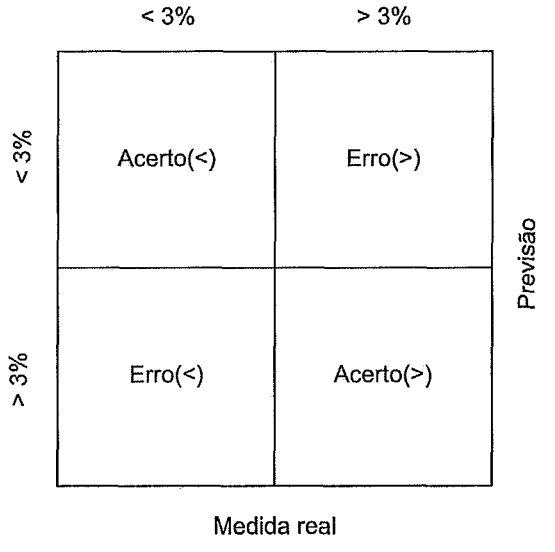


Figura 5.13: Definições de acertos e erros usadas nas métricas de *Precision* e *Recall*.

A partir destas definições, as métricas de *Precision* e *Recall*, para frações de perda baixas, são dadas por:

$$\textit{Precision} \text{ em perdas baixas} = \frac{\text{Acerto}(\leq)}{\text{Acerto}(\leq) + \text{Erro}(\leq)}, \tag{5.7}$$

$$\textit{Recall} \text{ em perdas baixas} = \frac{\text{Acerto}(\leq)}{\text{Acerto}(\leq) + \text{Erro}(>)}; \tag{5.8}$$

e, de forma equivalente, para frações de perda altas:

$$\textit{Precision} \text{ em perdas altas} = \frac{\text{Acerto}(>)}{\text{Acerto}(>) + \text{Erro}(>)}, \tag{5.9}$$

$$\textit{Recall} \text{ em perdas altas} = \frac{\text{Acerto}(>)}{\text{Acerto}(>) + \text{Erro}(\leq)}. \tag{5.10}$$

Este tipo de previsão pode ser usado, por exemplo, na detecção de degradações de desempenho em um canal fim-a-fim para controlar esquemas de *path-switching* (veja, por exemplo, [Bremler-Barr et al. 2003] e [Tao et al. 2004]). Nosso algoritmo aplicado aos HMMs tem resultados melhores nessa métrica comparado à estratégia replicadora.

Também na Tabela 5.1, é possível ver que o HMM-Pacote, usando a medida de perda em estado estacionário como preditor das taxas de perda (denotado na tabela como HMM-Pacote-EE), não tem um desempenho tão bom quanto o do modelo correspondente usando nosso algoritmo da Seção 3.2.

O algoritmo proposto, quando usado com os HMMs, também pode calcular outras métricas de previsão como percentis para a taxa de perdas. A Figura 5.14 mostra um exemplo simples dessa possibilidade. Na figura, um intervalo de previsão de 95% é exibido para três de nossos *traces*. Esse intervalo foi obtido, encontrando os percentis de 2.5 e 97.5 a cada previsão individual. É possível observar que as curva de taxas de perdas reais estão quase que inteiramente contidas entre os limites superiores e inferiores dos intervalos previstos por nosso algoritmo. Este tipo de previsão não pode ser realizado com uma estratégia mais simples como o replicador ou um preditor de estado estacionário.

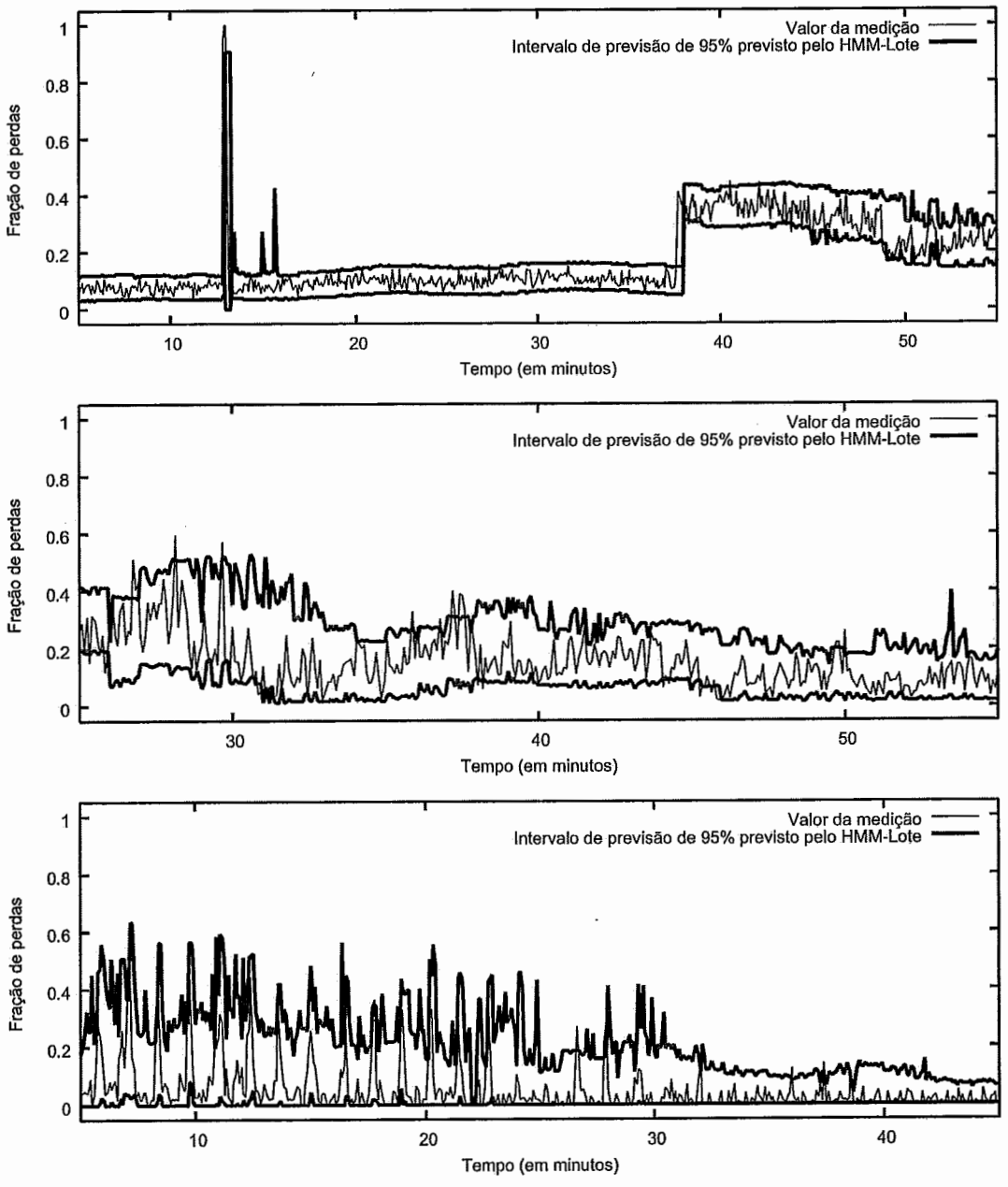


Figura 5.14: Intervalo de previsão de 95% obtido para três exemplos de *traces*.

Capítulo 6

Conclusões e Trabalhos Futuros

Neste trabalho, nós desenvolvemos um novo algoritmo para previsão de perdas de pacotes. Também propomos um modelo hierárquico visando a capturar as variações de curto prazo nas estatísticas de perda. Este modelo, quando usado com o algoritmo de previsão, é suficientemente preciso, prevendo taxas de perdas dentro de uma janela de tempo futuro. Para estimar os parâmetros do modelo proposto, novas equações foram desenvolvidas.

Avaliamos o algoritmo de previsão usando três modelos diferentes e sobre uma série de *traces* coletados na Internet. Nossos resultados mostram que todos os modelos se saem razoavelmente bem em muitos cenários. Entretanto, nosso modelo proposto é muito mais eficiente em termos de estimação de parâmetros e também supera em precisão os outros modelos, na maior parte dos casos. Este é, portanto, o modelo de escolha para o nosso algoritmo.

Também concluímos que a taxa de perda de curto prazo não pode ser bem aproximada pela probabilidade de perda em estado estacionário, uma vez que isso pode resultar em previsões pouco precisas. Além disso, nossos resultados também mostram que as previsões de curto prazo do algoritmo quando o HMM-Pacote é empregado não capturam as altas variações na taxa de perda. Isso ocorre pois esse modelo chega ao estado estacionário em um curto espaço de tempo. Mostramos que

ambos HMM-Agregado e HMM-Lote são mais adequados que o HMM-Pacote para modelar esses efeitos transientes.

É claro que a acurácia da previsão é limitada pela quantidade de correlações temporais no processo de perdas de pacotes. *Traces* com dependências temporais relativamente pequenas reduzem a capacidade dos modelos de prever taxas de perdas de curto prazo.

O algoritmo de previsão deve ser de valor quando aplicado à transmissão de mídia contínua em aplicações como VoIP e vídeo-conferência, ou ainda outras aplicações que possam se beneficiar da previsão de taxas de perda. Temos planos de incorporar o algoritmo como parte de um protocolo de controle adaptativo no futuro.

Além disso, nosso HMM de observações em lote pode ser usado como um modelo geral para um canal de comunicações, avaliando não apenas a taxa esperada de perda, mas também outras estatísticas como percentis de taxas de perdas e métricas do tamanho das rajadas de perdas. Outros modelos como o HMM-Agregado ou a estratégia de replicação não podem simultaneamente calcular todas essas medidas.

Apêndice A

Resultados Auxiliares

Neste apêndice, apresentamos resultados auxiliares a um ou mais pontos do desenvolvimento dos capítulos desta dissertação.

A.1 Redes Bayesianas

Uma *rede Bayesiana* é formada a partir de um *grafo direcionado acíclico*, no qual associamos a cada vértice, v_i, uma variável aleatória, e a cada aresta, $v_i \rightarrow v_j$, a noção de causalidade direta da variável v_i sobre a variável v_j. A Figura A.1 mostra um exemplo de rede Bayesiana com 5 variáveis e sua respectiva estrutura de causalidade.

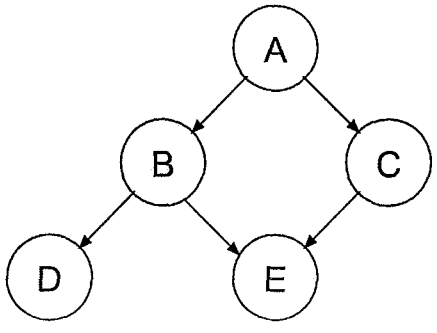


Figura A.1: Uma rede Bayesiana com 5 variáveis (A, B, C, D e E).

Para formalizar o conceito da relação de causalidade entre as variáveis de uma rede Bayesiana, seja $p(v_j)$ o conjunto de variáveis v_i, tais que $v_i \rightarrow v_j$ seja uma das arestas no grafo associado à rede. Isto é, $p(v_j)$ é o conjunto de vértices que são *pais* de v_j. Seja também $a(v_j)$ o fecho transitivo de $p(v_j)$, i.e., $a(v_j)$ é formado pelos elementos de $p(v_j)$ acrescidos de todo elemento v_i, tal que $v_i \rightarrow v_k$ é um aresta e v_k já pertence a $a(v_j)$. Evidentemente, $a(v_j)$ define o conjunto de *antecessores* de um vértice v_j.

Agora enunciamos a seguinte definição que caracteriza a relação de causalidade entre as variáveis de uma rede Bayesiana.

Definição A.1 (Causalidade Direta). *A distribuição de uma variável, v_i, em uma rede Bayesiana, condicionada nos valores de seus antecessores, depende apenas das variáveis que são pais de v_i, i.e.:*

$$P(v_i|a(v_i)) = P(v_i|p(v_i)). \tag{A.1}$$

Dizemos que as variáveis no conjunto $p(v_i)$ possuem causalidade direta sobre v_i.

Um outro conceito importante é o de *d-separação* (*d-separation* ou *dependence separation*), que posteriormente permitirá uma definição formal para a independência condicional em redes Bayesianas.

Definição A.2 (d-separação). *Dois vértices no grafo de uma rede Bayesiana, v_x e v_y, são ditos d-separados por um conjunto de vértices, Z, se, e somente se, para cada caminho não direcionado p, entre v_x e v_y, as seguintes condições são satisfeitas simultaneamente:*

(a) *p contém uma cadeia $v_i \rightarrow v_j \rightarrow v_k$, ou uma divergência $v_i \leftarrow v_j \rightarrow v_k$, tal que v_j esteja em Z;*

(b) *p não contém uma convergência $v_i \rightarrow v_j \leftarrow v_k$ tal que v_j ou qualquer um de seus descendentes esteja em Z.*

No exemplo da Figura A.1, as variáveis B e C são d-separadas pelo conjunto $Z = \{A\}$. Entretanto o conjunto $Z = \{A, E\}$ não gera a d-separação entre B e C, visto que a condição (b) da Definição A.2 deixa de ser satisfeita.

Finalmente podemos definir o conceito de independência condicional entre variáveis de uma rede Bayesiana da seguinte maneira:

Definição A.3 (Independência Condicional). *Dois conjuntos de variáveis, X e Y, são ditos condicionalmente independentes, dado um terceiro conjunto de variáveis, Z, se, e somente se, cada par de vértices, v_x em X e v_y em Y, é d-separado por Z.*

Um modelo de Markov oculto pode ser visto como um caso especial de uma rede Bayesiana, com a estrutura ilustrada na Figura A.2. É interessante reparar que a estrutura da rede Bayesiana de um HMM é uma *árvore*, isto é, há apenas um caminho entre cada par de vértices. Além disso, não há convergências entre vértices e, por esta razão, o conceito de d-separação é reduzido a satisfazer apenas a condição (a) na Definição A.2. De maneira mais formal:

Definição A.4 (Independência Condicional em Modelos de Markov Ocultos). *Em um modelo de Markov oculto, dois conjuntos de variáveis, X e Y, são condicionalmente independentes, dado um terceiro conjunto, Z, se, e somente se, não existe um caminho entre uma variável de X e outra de Y, sem passar por um vértice de Z.*

O seguinte corolário pode ser obtido diretamente a partir da Definição A.4:

Corolário A.1. *Para um modelo de Markov oculto, as variáveis que definem as observações no futuro, a partir de um instante t $(X_t, X_{t+1}, \ldots)$, são condicionalmente independentes das observações passadas $(X_1, \ldots, X_{t-1})$, dado o estado oculto no instante t, Y_t. Em outras palavras:*

$$P(X_t, X_{t+1}, \ldots | X_1, \ldots, X_{t-1}, Y_t) = P(X_t, X_{t+1}, \ldots | Y_t). \tag{A.2}$$

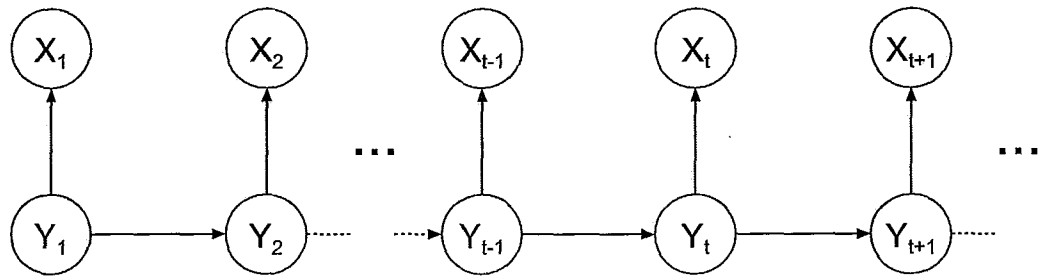


Figura A.2: Representação de um modelo de Markov oculto como uma rede Bayesiana.

De maneira mais geral, utilizamos, na Seção 3.2, o seguinte resultado, que pode ser demonstrado a partir do Corolário A.1:

Corolário A.2. *Seja f uma função das variáveis de observação futuras, a partir de um instante t $(X_t, X_{t+1}, \ldots)$. Então, esta função é condicionalmente independente das observações passadas $(X_1, \ldots, X_{t-1})$, dado o estado oculto no instante t, Y_t. Formalmente:*

$$P(f(X_t, X_{t+1}, \ldots)|X_1, \ldots, X_{t-1}, Y_t) = P(f(X_t, X_{t+1}, \ldots)|Y_t). \tag{A.3}$$

Isto pode ser facilmente demonstrado, condicionando-se a distribuição de f nos valores de cada observação futura, e reduzindo o problema ao enunciado do Corolário A.1.

A.2 O Algoritmo *Forward-Backward*

Seja a variável *forward* definida como:

$$\alpha_t(i) = P(X_{1:t}, Y_t = i|\lambda). \tag{A.4}$$

Isto é, $\alpha_t(i)$ é a probabilidade conjunta da observação de todos os símbolos, $x_1, \ldots, x_t$, e o estado oculto i no tempo t, para uma atribuição fixa de parâmetros, λ. Para t maior que 1, condicionando no valor do estado oculto no tempo $t-1$, podemos

escrever a seguinte expressão recursiva:

$$\begin{aligned}\alpha_t(i) &= \sum_{j=1}^{n} P(X_{1:t-1}, Y_{t-1} = j|\lambda)P(X_t, Y_t = i|X_{1:t-1}, Y_{t-1} = j, \lambda) \\ &= \sum_{j=1}^{n} P(X_{1:t-1}, Y_{t-1} = j|\lambda)P(Y_t = i|Y_{t-1} = j, \lambda)P(X_t|Y_t = i, \lambda) \\ &= \sum_{j=1}^{n} \alpha_{t-1}(j)\, a_{ji}\, b_{ix_t}. \end{aligned} \tag{A.5}$$

Por outro lado, para t igual a 1, é trivial que:

$$\alpha_1(i) = \pi_i b_{ix_1}. \tag{A.6}$$

De forma análoga, seja a variável *backward* definida como:

$$\beta_t(i) = P(X_{t+1:T}|Y_t = i, \lambda). \tag{A.7}$$

O valor de $\beta_t(i)$ deve ser interpretado como a probabilidade conjunta da observação dos símbolos $x_{t+1}, \ldots, x_T$, dados o estado oculto i no tempo t e os parâmetros, λ. Para t menor que T, podemos condicionar o valor de $\beta_t(i)$ no estado oculto no tempo $t+1$:

$$\begin{aligned}\beta_t(i) &= \sum_{j=1}^{n} P(Y_{t+1} = j|Y_t = i, \lambda)P(X_{t+1:T}|Y_t = i, Y_{t+1} = j, \lambda) \\ &= \sum_{j=1}^{n} P(Y_{t+1} = j|Y_t = i, \lambda)P(X_{t+1}|Y_{t+1} = j, \lambda)P(X_{t+2:T}|Y_{t+1} = j, \lambda) \\ &= \sum_{j=1}^{n} a_{ij}\, b_{jx_{t+1}} \beta_{t+1}(j). \end{aligned} \tag{A.8}$$

Finalmente, a base, no caso t igual a T, é dada por:

$$\beta_T(i) = 1. \tag{A.9}$$

O algoritmo *forward-backward* é um ponto central em muitas das computações que envolvem modelos de Markov ocultos. Na Seção 2.2, utilizamos estas definições nas fórmulas do algoritmo Baum-Welch, que estimam parâmetros de máxima verossimilhança para HMMs. Na Seção 3.2, usamos a variável *forward* para computar a distribuição do número de perdas em um intervalo de observações futuras, condicionada nos valores de amostras passadas.

A definição das variáveis *forward* e *backward* pode também ser usada para computar a medida de verossimilhança de uma amostra, uma vez que, para todo t:

$$\begin{aligned} P(\mathbf{X}|\lambda) &= \sum_{i=1}^{n} P(X_{1:t}, Y_t = i, X_{t+1:T}|\lambda) \\ &= \sum_{i=1}^{n} P(X_{1:t}, Y_t = i|\lambda) P(X_{t+1:T}|Y_t = i, \lambda) \\ &= \sum_{i=1}^{n} \alpha_t(i)\beta_t(i). \end{aligned} \tag{A.10}$$

É fácil perceber, por inspeção da Equação (A.5), que a variável *forward*, para t maior que 1 requer um número de operações aritméticas proporcional a N. Uma vez que $N(T-1)$ variáveis precisam ser calculadas através dessa fórmula, a complexidade assintótica da recursão completa é da ordem de N^2T. O mesmo pode ser verificado para a recursão *backward*.

Uma vez que as recursões *forward* e *backward* são tão ubíquas na análise de modelos de Markov ocultos, estas costumam determinar a complexidade de outros procedimentos. De fato, o custo de cada iteração do método Baum-Welch é dominado pelo algoritmo *forward-backward*.

Na prática, o cálculo das variáveis *forward* e *backward*, através das fórmulas que acabamos de apresentar pode ser problemático devido a erros de truncamento nas várias multiplicações envolvendo fatores no intervalo $[0, 1]$. Entretanto, uma solução eficaz e elegante para esse problema é o método de *scaling*. Basicamente, as variáveis *forward* podem ser normalizadas para cada t, mantendo registro dos seus fatores de escala. Por outro lado, as variáveis *backward* podem ser multiplicadas pelos mesmos fatores, de forma que estas também tenham seu erro minimizado.

Muito embora o uso destes fatores de escala seja importante na implementação prática dos algoritmos, os detalhes desta computação não são essenciais ao entendimento dos pontos de interesse desta dissertação. Recomendamos ao leitor interessado neste tema a seção apropriada de [Rabiner 1989]. Ainda assim, enfatizamos que todos os resultados numéricos apresentados nesta dissertação envolvem o cálculo

de variáveis *forward* e *backward* com o procedimento de *scaling*.

A.3 A Desigualdade de Jensen

Antes de apresentar o resultado conhecido como desigualdade de Jensen, é preciso definir o conceito de função côncava. Embora existam diversas formas equivalentes para definir uma função côncava, utilizamos a seguinte:

Definição A.5 (Função Côncava). *Uma função real diferenciável, $f : \mathbb{R}^n \mapsto \mathbb{R}$, é dita côncava em uma região, $\Omega \subseteq \mathbb{R}^n$, se, e somente se, para todos x e y pertencentes a Ω, temos que:*

$$f(y) \geq f(x) + \nabla f(x)(y - x). \tag{A.11}$$

Em outras palavras, para cada ponto, x, de Ω, o hiperplano definido pela aproximação de Taylor de primeira ordem para f em torno de x é um limite superior para todos os pontos da função em Ω.

Por sua vez, a desigualdade de Jensen também é um resultado que pode ser aplicado de maneira bem geral para funções côncavas, mas estamos interessados na sua versão probabilística, que pode ser enunciada da seguinte forma:

Teorema A.1 (Desigualdade de Jensen). *Para uma função côncava, $f : \mathbb{R}^n \mapsto \mathbb{R}$, e um vetor de variáveis aleatórias X, é verdade que:*

$$f(E[X]) \geq E[f(X)]. \tag{A.12}$$

Demonstração. Uma vez que f é côncava, pela Definição A.5, temos que:

$$f(X) \leq f(E[X]) + \nabla f(E[X])(X - E[X]), \tag{A.13}$$

e portanto, podemos escrever:

$$E[f(X)] \leq E[f(E[X]) + \nabla f(E[X])(X - E[X])] = f(E[X]). \tag{A.14}$$

□

A.4 O Algoritmo Baum-Welch

Nesta seção, mostraremos como é possível chegar as fórmulas do algoritmo Baum-Welch a partir do problema de maximização dos termos da função auxiliar, conforme dispostos nas Equações (2.15a–c). Há outras formas de se chegar às mesmas expressões que não levam em consideração a resolução explícita de um problema de otimização, conforme pode ser visto em [Levinson et al. 1983].

A demonstração original de [Baum et al. 1970], que também pode ser encontrada em [Levinson et al. 1983], faz uso do seguinte lema:

Lema A.1. *Sejam os coeficientes $c_1, \ldots, c_N$, maiores que zero. Então o problema de otimização:*

$$\begin{aligned} (P) \quad & \textit{maximizar} \quad \sum_{i=1}^{N} c_i \log x_i, \\ & \textit{sujeito a} \quad \sum_{i=1}^{N} x_i = 1, \end{aligned}$$

possui um único máximo global no ponto em que, para todo i:

$$x_i = \frac{c_i}{\sum_{j=1}^{N} c_j}. \tag{A.15}$$

Demonstração. A derivada parcial do Lagrangeano da função objetivo, em relação a uma variável x_i é dada por:

$$\frac{\partial}{\partial x_i}\left[\sum_{i=1}^{N} c_i \log x_i - \mu \sum_{i=1}^{N} x_i\right] = \frac{c_i}{x_i} - \mu. \tag{A.16}$$

Logo, no ponto ótimo, temos que, para todo i:

$$c_i = \mu x_i. \tag{A.17}$$

Somando (A.17) para todo i, temos:

$$\sum_{i=1}^{N} c_i = \mu, \tag{A.18}$$

e substituindo (A.18) em (A.17), o resultado em (A.15) se torna evidente. □

Uma vez avaliada a função auxiliar, $Q(\lambda|\overline{\lambda})$, as fórmulas de re-estimação são dadas pelo seguinte problema de maximização:

$$\begin{aligned} (P) \quad \text{maximizar} \quad & Q(\lambda|\overline{\lambda}), \\ \text{sujeito a} \quad & \sum_{i=1}^{N} \pi_i = 1, \\ & \sum_{j=1}^{N} a_{ij} = 1, \quad \forall i, \\ & \sum_{j=1}^{M} b_{ij} = 1, \quad \forall i, \end{aligned}$$

Sejam, portanto os três sub-problemas de maximização:

$$\begin{aligned} (P_\pi) \quad \text{maximizar} \quad & Q_1(\pi|\overline{\lambda}), \\ \text{sujeito a} \quad & \sum_{i=1}^{N} \pi_i = 1, \end{aligned}$$

$$\begin{aligned} (P_{\mathbf{A}}) \quad \text{maximizar} \quad & Q_2(\mathbf{A}|\overline{\lambda}), \\ \text{sujeito a} \quad & \sum_{j=1}^{N} a_{ij} = 1, \quad \forall i, \end{aligned}$$

$$\begin{aligned} (P_{\mathbf{B}}) \quad \text{maximizar} \quad & Q_3(\mathbf{B}|\overline{\lambda}), \\ \text{sujeito a} \quad & \sum_{j=1}^{M} b_{ij} = 1, \quad \forall i. \end{aligned}$$

Para mostrar que a solução de (P) pode ser obtida pelas soluções parciais de (P_π), $(P_\mathbf{A})$ e $(P_\mathbf{B})$, apresentamos o seguinte teorema:

Teorema A.2. *Se π^*, $\mathbf{A}^*$ e $\mathbf{B}^*$ são respectivamente máximos locais de (P_π), $(P_\mathbf{A})$ e $(P_\mathbf{B})$, então $\lambda^* = (\pi^*, \mathbf{A}^*, \mathbf{B}^*)$ é um máximo local de (P).*

Demonstração. Sejam π^*, $\mathbf{A}^*$ e $\mathbf{B}^*$ são máximos locais respectivos de (P_π), $(P_\mathbf{A})$ e $(P_\mathbf{B})$, e seja $\lambda' = (\pi', \mathbf{A}', \mathbf{B}')$ um vetor com uma direção viável de acréscimo para (P) em λ^*. Isto é, para algum $\epsilon > 0$ e todo α em $0 < \alpha < \epsilon$, temos que $\lambda + \alpha\lambda'$ ainda satisfaz as restrições de (P).

Usamos a notação $\mathbf{a}_i$ e $\mathbf{b}_i$ para representar a i-ésima linha de $\mathbf{A}$ e $\mathbf{B}$, respectivamente. Também denotamos por ∇_x o gradiente com respeito as variáveis no vetor x. A taxa de variação de $Q(\lambda|\overline{\lambda})$ na direção de λ' é dada por (veja [Luenberger 2003] para detalhes):

$$\begin{aligned}\nabla_\lambda Q(\lambda^*|\overline{\lambda})\lambda' &= \sum_{i=1}^{N} \frac{\partial Q_1(\pi^*|\overline{\lambda})}{\partial \pi_i}\pi_i' + \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial Q_2(\mathbf{A}^*|\overline{\lambda})}{\partial a_{ij}}a_{ij}' \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{\partial Q_3(\mathbf{B}^*|\overline{\lambda})}{\partial b_{ij}}b_{ij}' \qquad \text{(A.19)} \\ &= \nabla_\pi Q_1(\pi^*|\overline{\lambda})\pi' + \sum_{i=1}^{N} \nabla_{\mathbf{a}_i} Q_2(\mathbf{A}^*|\overline{\lambda})\mathbf{a}_i' \\ &\quad + \sum_{i=1}^{N} \nabla_{\mathbf{b}_i} Q_3(\mathbf{B}^*|\overline{\lambda})\mathbf{b}_i' \qquad \text{(A.20)}\end{aligned}$$

Uma vez que λ' é uma direção de acréscimo para $Q(\lambda^*|\overline{\lambda})$, temos que:

$$\nabla_\lambda Q(\lambda^*|\overline{\lambda})\lambda' > 0, \qquad \text{(A.21)}$$

e portanto, pelo menos um dos três termos do lado direto da Equação (A.20) deve ser estritamente positivo. Seja este termo $\nabla_\pi Q_1(\pi^*|\overline{\lambda})\pi'$, sem perda de generalidade. Uma vez que as restrições de (P) estão particionadas em três conjuntos, um para cada um dos problemas menores, π' também deve ser uma direção viável de

acréscimo para (P_π), o que por si só, contradiz a idéia que π^* é um máximo local para esse problema. □

O mesmo argumento usado no Teorema A.2 pode ser usado para mostrar que:

Corolário A.3. *Os problemas* $(P_\mathbf{A})$ *e* $(P_\mathbf{B})$ *podem ser resolvidos a partir das soluções dos respectivos sub-problemas:*

$$(P_\mathbf{A}) \quad \textit{maximizar} \quad \sum_{j=1}^{N} \log a_{ij} \sum_{t=1}^{T-1} \xi_t(i,j),$$
$$\textit{sujeito a} \quad \sum_{j=1}^{N} a_{ij} = 1,$$

$$(P_\mathbf{B}) \quad \textit{maximizar} \quad \sum_{j=1}^{N} \log b_{ij} \sum_{t=1}^{T} \mathbb{I}\{x_t = j\}\gamma_t(i),$$
$$\textit{sujeito a} \quad \sum_{j=1}^{M} b_{ij} = 1,$$

para cada estado oculto, i, do modelo de Markov oculto.

Usando o Teorema A.2, o Corolário A.3 e o Lema A.1, é fácil concluir que as fórmulas de re-estimação Baum-Welch são, de fato, dadas pelas Equações (2.15a–c).

A.5 Extensão do Algoritmo Baum-Welch

Ao propor o modelo de observações em lote, criamos restrições nos parâmetros de observação de modo que o processo de perda em cada lote seja gerado por uma cadeia de Markov de 2 estados.

O Teorema A.2 da Seção A.4 implica que se se adicionarmos outras restrições envolvendo π, $\mathbf{A}$ e $\mathbf{B}$, *uma vez que elas permaneçam independentes*, as fórmulas de

re-estimação mudarão apenas para as variáveis sobre as quais as novas restrições se aplicam.

Desta forma, podemos restringir nossa análise apenas aos parâmetros de observação **r**, **p** e **q**. Aplicando a definição da probabilidade de observação de um lote, na Equação (4.4), no termo da função auxiliar correspondente aos parâmetros de observação, dado pela Equação (2.14c), temos que as expressões que buscamos são os pontos de máximos das seguintes funções, para cada estado oculto, i:

$$Q_4(r_i|\overline{\lambda}) = \log r_i \sum_{t=1}^{T} \mathbb{I}\{x_{t,1} = 1\}\gamma_t(i) + \log(1 - r_i) \sum_{t=1}^{T} \mathbb{I}\{x_{t,1} = 0\}\gamma_t(i), \quad \text{(A.22a)}$$

$$Q_5(p_i|\overline{\lambda}) = \log p_i \sum_{t=1}^{T} S_t^{01}\gamma_t(i) + \log(1 - p_i) \sum_{t=1}^{T} S_t^{00}\gamma_t(i), \quad \text{(A.22b)}$$

$$Q_6(q_i|\overline{\lambda}) = \log q_i \sum_{t=1}^{T} S_t^{10}\gamma_t(i) + \log(1 - q_i) \sum_{t=1}^{T} S_t^{11}\gamma_t(i), \quad \text{(A.22c)}$$

que, através do Lema A.1, são dados pelas expressões (4.5a–c).

A.6 Estatísticas Suficientes

O conceito de estatística suficiente é recorrente no estudo de estimação de parâmetros para modelos probabilísticos. Portanto, seja a seguinte definição:

Definição A.6. *Seja θ o conjunto de parâmetros de um modelo probabilístico, e X uma amostra de valores para o mesmo modelo. Uma estatística $\Gamma(X)$ da amostra é uma* estatística suficiente *se, e somente se, a distribuição da amostra, X, dado o valor de $\Gamma(x)$ não depende de θ.*

Na Seção 4.3, propomos um modelo hierárquico cujos parâmetros podem ser estimados de maneira mais eficiente a partir de um conjunto de estatísticas suficientes que, na prática, pode ser muito menor que a amostra completa do processo de perdas.

Referências Bibliográficas

[Baum et al. 1970] Baum, L. E., Petrie, T., Soules, G., e Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

[Bolot et al. 1999] Bolot, J. C., Parisis, S. F., e Towsley, D. F. (1999). Adaptive FEC-based error control for Internet telephony. In *Proceedings of the IEEE INFOCOM*, pp. 1453–1460.

[Bremler-Barr et al. 2003] Bremler-Barr, A., Cohen, E., Kaplan, H., e Mansour, Y. (2003). Predicting and bypassing end-to-end internet service degradations. *IEEE Journal on Selected Areas in Communications*, 21(6):961–978.

[Brockwell e Davis 2002] Brockwell, P. J. e Davis, R. A. (2002). *Introduction to Time Series and Forecasting.* Springer.

[de Souza e Silva e Leão 2000] de Souza e Silva, E. e Leão, R. M. M. (2000). The TANGRAM-II Environment. In *TOOLS '00: Proceedings of the 11th International Conference on Computer Performance Evaluation: Modelling Techniques and Tools*, pp. 366–369, London, UK. Springer-Verlag.

[de Vielmond e de Souza e Silva 2005] de Vielmond, C. C. L. B. e de Souza e Silva, E. G. (2005). Implementação e avaliação de um algoritmo FEC adaptativo para voz sobre IP. Projeto Final de Curso, Universidade Federal do Rio de Janeiro - IM - Departamento de Ciência da Computação.

[Dempster et al. 1977] Dempster, A. P., Laird, N. M., e Rubin, D. B. (1977). Maximum likelihood from imcomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

[Duarte 2003] Duarte, F. P. (2003). Algoritmo adaptativo para previsão e recuperação de perda de pacotes em aplicações multimídias usando cadeias de Markov ocultas. Tese de Mestrado, Universidade Federal do Rio de Janeiro - COPPE - Programa de Engenharia de Sistemas e Computação.

[Duarte et al. 2003] Duarte, F. P., de Souza e Silva, E. A., e Towsley, D. F. (2003). An adaptive FEC algorithm using hidden Markov chains. *SIGMETRICS Performance Evaluation Review*, 31(2):11–13.

[Elliot et al. 1995] Elliot, R. J., Aggoun, L., e Moore, J. B. (1995). *Hidden Markov Models: Estimation and Control.* Springer-Verlag.

[Elliott 1963] Elliott, E. O. (1963). Estimates of error rates for codes on burst-noise channels. *Bell Systems Technical Journal*, 42:1977–1997.

[Elliott 1965] Elliott, E. O. (1965). A model of the switched telephone network for data communications. *Bell Systems Technical Journal*, 44:89–109.

[Gilbert 1960] Gilbert, E. N. (1960). Capacity of a burst-noise channel. *Bell Systems Technical Journal*, 39:1253–1265.

[Ji et al. 2004] Ji, P., Liu, B., Towsley, D. F., Ge, Z., e Kurose, J. F. (2004). Modeling frame-level errors in GSM wireless channels. *Performance Evaluation*, 55(1-2):165–181.

[Karol et al. 2004] Karol, M., Krishnan, P., e Li, J. J. (2004). Rapid fault detection and recovery for IP telephony. In *Proceedings of IEEE ICC*, volume 3, pp. 1478–1483.

[Kemeny e Snell 1969] Kemeny, J. G. e Snell, J. L. (1969). *Finite Markov Chains.* D. Van Nostrand Company.

[Konrad et al. 2001] Konrad, A., Zhao, B., Joseph, A., e Ludwig, R. (2001). A Markov-based channel model algorithm for wireless networks. In *Proceedings of the Fourth ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*, pp. 28–36.

[Levinson et al. 1983] Levinson, S. E., Rabiner, L. R., e Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell Systems Technical Journal*, 62(4):1035–1074.

[Luenberger 2003] Luenberger, D. G. (2003). *Linear and Nonlinear Programming.* Springer.

[Rabiner 1989] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285.

[Rabiner e Schafer 1978] Rabiner, L. R. e Schafer, R. W. (1978). *Digital Processing of Speech Signals.* Prentice Hall.

[Rappaport 2001] Rappaport, T. S. (2001). *Wireless Communications: Principle and Practice.* Prentice Hall.

[Ross 1997] Ross, S. M. (1997). *Simulation.* Academic Press.

[Salamatian e Vaton 2001] Salamatian, K. e Vaton, S. (2001). Hidden Markov modeling for network communication channels. In *Proceedings of the ACM SIGMETRICS*, pp. 92–101.

[Schulzrinne et al. 2003] Schulzrinne, H., Casner, S., Frederick, R., e Jacobson, V. (2003). RTP: A Transport Protocol for Real-Time Applications. RFC 3550 (Standard).

[Su et al. 2004] Su, Y. C., Yang, C. S., e Lee, C. W. (2004). The analysis of packet loss prediction for Gilbert-model with loss rate uplink. *Information Processing Letters*, 90:155–159.

[Tao e Guerin 2004] Tao, S. e Guerin, R. (2004). On-line estimation of Internet path performance: an application perspective. In *Proceedings of IEEE INFOCOM*, volume 3, pp. 1774–1785.

[Tao et al. 2005] Tao, S., Xu, K., Estepa, A., Fei, T., Gao, L., Guerin, R., Kurose, J. F., Towsley, D. F., e Zhang, Z. (2005). Improving VoIP quality through path switching. In *Proceedings of IEEE INFOCOM*, volume 4, pp. 2268–2278.

[Tao et al. 2004] Tao, S., Xu, K., Xu, Y., Fei, T., Gao, L., Guerin, R., Kurose, J. F., Towsley, D. F., e Zhang, Z. (2004). Exploring the performance benefits of end-to-end path switching. In *Proceedings of IEEE ICNP*, pp. 304–315.

[Wei et al. 2002] Wei, W., Wang, B., e Towsley, D. F. (2002). Continuous-time hidden Markov models for network performance evaluation. *Performance Evaluation*, 49(1-4):129–146.

[Yajnik et al. 1999] Yajnik, M., Moon, S. B., Kurose, J. F., e Towsley, D. F. (1999). Measurement and modeling of the temporal dependence in packet loss. In *Proceedings of the IEEE INFOCOM*, pp. 345–352.

[Zhang et al. 2001] Zhang, Y., Duffield, N. G., Paxson, V., e Shenker, S. (2001). On the constancy of Internet path properties. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, pp. 197–211.

[Zhang et al. 2000] Zhang, Y., Paxson, V., e Shenker, S. (2000). The stationarity of Internet path properties: routing, loss, and throughput. Technical report, ACIRI.