MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2016/05.16.02.36-TDI

APLICAÇÃO DA APRENDIZAGEM POR REFORÇO PARA O PROBLEMA DE ALOCAÇÃO DE ESPECTRO EM REDES ÓPTICAS ELÁSTICAS

Luis Fernando Amorim França

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Solon Venâncio de Carvalho, e Rita de Cássia Meneses Rodrigues, aprovada em 16 de maio de 2016.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34P/3LMJSEL>

INPE
São José dos Campos
2016

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@inpe.br

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2016/05.16.02.36-TDI

APLICAÇÃO DA APRENDIZAGEM POR REFORÇO PARA O PROBLEMA DE ALOCAÇÃO DE ESPECTRO EM REDES ÓPTICAS ELÁSTICAS

Luis Fernando Amorim França

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Solon Venâncio de Carvalho, e Rita de Cássia Meneses Rodrigues, aprovada em 16 de maio de 2016.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34P/3LMJSEL>

INPE
São José dos Campos
2016

Dados Internacionais de Catalogação na Publicação (CIP)

França, Luis Fernando Amorim.

F844a Aplicação da aprendizagem por reforço para o problema de alocação de espectro em redes ópticas elásticas / Luis Fernando Amorim França. – São José dos Campos : INPE, 2016.

xxii + 83 p. ; (sid.inpe.br/mtc-m21b/2016/05.16.02.36-TDI)

Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2016.

Orientadores : Drs. Solon Venâncio de Carvalho, e Rita de Cássia Meneses Rodrigues.

1. Redes ópticas elásticas. 2. Alocação de espectro. 3. Processo Markoviano de decisão. 4. Aprendizagem por reforço. I.Título.

CDU 519.87:519.863

Aluno (a): *Luis Fernando Amorim França*

Título: "APLICAÇÃO DA APRENDIZAGEM POR REFORÇO PARA O PROBLEMA DE ALOCAÇÃO DE ESPECTRO EM REDES ÓPTICAS ELÁSTICAS".

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de *Doutor(a)* em *Computação Aplicada*

Dr. Haroldo Fraga de Campos Velho

Presidente / INPE / São José dos Campos - SP

Dr. Solon Venâncio de Carvalho

Orientador(a) / INPE / SJCampos - SP

Dra. Rita de Cássia Meneses Rodrigues

Orientador(a) / INPE / SJCampos - SP

Dr. Horacio Hideki Yanasse

Membro da Banca / UNIFESP / SJCampos - SP

Dr. Carlos Henrique Costa Ribeiro

Convidado(a) / ITA / São José dos Campos - SP

Dr. Armando Zeferino Milioni

Convidado(a) / ITA / São José dos Campos - SP

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

São José dos Campos, 16 de Maio de 2016

"Experiência não é o que acontece com um homem; é o que um homem faz com o que lhe acontece".

Aldous Huxley

A meus pais Carlos e Maria Luiza, e a meus irmãos Paula, César e Carlinhos

AGRADECIMENTOS

Agradeço a Deus pela dádiva da vida, pelas oportunidades a mim oferecidas, e pela saúde para poder aproveitá-las.

À minha família que, mesmo à distância, me apoiou e incentivou incondicionalmente durante todos esses anos.

Aos meus orientadores Dr Solon Venâncio de Carvalho e Dra Rita de Cássia Meneses Rodrigues pela dedicação, paciência, incentivo e conhecimento compartilhado desde os estudos do mestrado.

Ao Instituto Nacional de Pesquisas Espaciais (INPE), em especial o Curso de Computação Aplicada (CAP), pela oportunidade de desenvolver este trabalho.

Aos meus orientadores durante o período de estágio sanduíche, Lena Wosinska e Paolo Monti, pela dedicação e pelas valiosas contribuições.

A todos os colegas do Laboratório de Redes Ópticas (ONLab) do Instituto Real de Tecnologia da Suécia pela oportunidade de colaboração durante o estágio sanduíche.

Aos meus amigos de longa data, que mesmo à distância permanecem presentes na minha vida.

Aos amigos e colegas do INPE, especialmente do CCS, pelo convívio e amizade durante todos esses anos.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro concedido para realização deste trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro para realização do estágio sanduíche.

Por fim, a todos que colaboraram de maneira direta ou indireta para a realização deste trabalho.

RESUMO

As Redes ópticas elásticas vêm sendo desenvolvidas recentemente com o intuito de prover maior flexibilidade em relação às redes ópticas tradicionais. Nessas redes, recursos, denominados *slots*, são alocados de acordo com a demanda de tráfego. Torna-se possível, então, a geração de caminhos ópticos para estabelecer conexões para diferentes classes de serviços com requerimentos de banda heterogêneos. Ao se estabelecer um caminho óptico deve-se selecionar quais enlaces serão utilizados para rotear a conexão e, para cada enlace dessa rota, quais *slots* serão alocados. Neste trabalho focamos em um enlace de uma rede óptica elástica sob tráfego dinâmico, e, portanto, o roteamento não precisa ser realizado. Nós propomos um modelo analítico, por meio de um processo markoviano de decisão a tempo contínuo, para encontrar uma política ótima de alocação de espectro. Uma vez que essa política é aplicada, nós utilizamos uma cadeia de Markov para calcular suas medidas de desempenho. Para instâncias mais realistas do problema, no entanto, o modelo analítico torna-se inviável de ser resolvido, seja por restrições de memória ou tempo de processamento. Dessa forma, propomos também a utilização de um algoritmo de aprendizagem por reforço para encontrar políticas de alocação de espectro nos casos em que o modelo analítico não pode ser aplicado. Resultados numéricos são apresentados para ilustrar as medidas de desempenho da política de alocação de espectro derivada do nosso modelo em relação a duas políticas comumente utilizadas na literatura, *First-Fit* e *Best-Fit*.

Palavras-chave: Redes ópticas elásticas. Alocação de espectro. Processo markoviano de decisão. Aprendizagem por reforço.

APPLICATION OF REINFORCEMENT LEARNING FOR THE SPECTRUM ALLOCATION PROBLEM IN ELASTIC OPTICAL NETWORKS

ABSTRACT

Elastic Optical Networks (EONs) have been recently proposed to provide flexibility over traditional optical networks. In these networks, resources, called slots, are allocated according to traffic demands, providing the possibility of generating optical paths to establish connection requests for different classes of services with heterogeneous bandwidth requirements. In order to establish the optical paths one must select which links will be used to route each connection and, for each link of the route, which slots will be allocated. In this work we focus in one link of an EON under dynamic traffic, and thus no routing needs to be done. We propose an analytical model, by means of a continuous-time Markov decision process, to find an optimal Spectrum Allocation (SA) policy. Once a SA policy is applied, we use a Markov chain to compute its performance metrics. For more realistic instances of the problem, however, the analytical model is computationally infeasible. Therefore, we also propose the use of a reinforcement learning algorithm in order to find SA policies for the cases where the analytical model cannot be applied. Numerical results are provided to illustrate the performance metrics of the SA policy derived from our model over two SA myopic policies commonly used in the literature, namely First-Fit and Best-Fit.

Keywords: Elastic optical networks. Spectrum allocation. Markov decision process. Reinforcement learning.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE ABREVIATURAS E SIGLAS

EON	–	*Elastic Optical Network*
FDM	–	*Frequency Division Multiplexing*
ITU-T	–	*International Telecommunication Union - Telecommunication Standardization Sector*
OFDM	–	*Orthogonal Frequency Division Multiplexing*
PLI	–	Programação Linear Inteira
PMD	–	Processo Markoviano de Decisão
Relaxed-SMART	–	*Relaxed Semi-Markov Average Reward Technique*
RSA	–	*Routing and Spectrum Allocaiton*
RWA	–	*Routing and Wavelength Assignment*
WDM	–	*Wavelength Division Multiplexing*

SUMÁRIO

Pág.

1 INTRODUÇÃO

Neste capítulo, apresentamos inicialmente o problema de alocação de espectro abordado nesta tese. Em seguida, descrevemos uma maneira de modelá-lo como um problema de tomada de decisão sequencial sob incerteza. Introduzimos, também, o paradigma de aprendizagem por reforço utilizado para resolver instâncias mais realistas desse problema. Por fim, apresentamos o objetivo desta tese e as nossas contribuições.

1.1 Alocação de Espectro em Redes Ópticas Elásticas

Redes ópticas WDM (*Wavelength Division Multiplexing*) vêm sendo amplamente utilizadas nos últimos anos para lidar com o crescimento rápido e contínuo das demandas de tráfego devido à possibilidade de transmitir dados de múltiplas fontes em apenas uma fibra óptica. Operadores de rede esperam que esse crescimento no tráfego continue a aumentar nos próximos anos (CISCO, 2014), impulsionado pela popularização de aplicações que requerem alta largura de banda, como video sob demanda, TV de alta definição e serviços baseados na nuvem, com velocidades variando de Gb/s a Tb/s.

As redes WDM operam sob a grade de 50 GHz sugerida pelo setor de padronização de telecomunicações da União Internacional de Telecomunicação (*International Telecommunication Union - Telecommunication Standardization Sector* - ITU-T), que divide o espectro óptico em porções fixas de 50 GHz comumente referidas apenas como comprimentos de onda[1] (*wavelengths*) na literatura. Nessas redes, cada conexão, também denominada chamada, ocupa a mesma porção de espectro independentemente da velocidade dos dados transmitidos, ou seja, um comprimento de onda é alocado para cada conexão transmitida mesmo quando a sua largura de banda requerida não é suficiente para utilizar toda a capacidade dessa porção do espectro. Essa granularidade rígida, combinada com a heterogeneidade crescente das demandas, pode levar à utilização ineficiente do espectro óptico disponível (CHRISTODOULOPOULOS et al., 2011). Para resolver este problema, as redes devem ser projetadas de modo que sejam capazes de lidar com tráfego heterogêneo enquanto melhoram a eficiência espectral.

Com o intuito de prover flexibilidade às redes ópticas tradicionais, foram propostas

[1]Embora seja um intervalo entre dois comprimentos de onda no espectro, o termo comprimento de onda é largamente utilizado na literatura de redes ópticas para denominar uma porção do espectro. Neste trabalho nós também utilizaremos este termo.

as redes ópticas elásticas (*Elastic Optical Networks* - EONs) (GERSTEL et al., 2012), nas quais o termo elástico se refere à possibilidade de gerar caminhos ópticos capazes de transmitir conexões com demandas variáveis. Um caminho óptico é o conjunto de enlaces (*links*) da rede utilizados para transmissão de conexões entre um par de nós origem-destino. Nas redes elásticas, o espectro é dividido em porções, denominadas *slots*, com uma granularidade menor que a das redes WDM tradicionais. Além disso, *slots* contíguos podem ser alocados para transmitir conexões cuja demanda de tráfego não pode ser satisfeita com apenas um *slot*.

Para se estabelecer um caminho óptico em uma rede óptica elástica precisa-se escolher uma rota e um ou mais *slots* contíguos utilizados para transmitir cada requisição de conexão. Esse problema, denominado roteamento e alocação de espectro (*Routing and Spectrum Allocation* - RSA), é uma generalização do largamente estudado problema de roteamento e atribuição de comprimento de onda (*Routing and Wavelength Assignment* - RWA) nas redes WDM. Ao se resolver o problema RSA deve-se garantir que o mesmo conjunto de *slots* contíguos seja alocado para uma dada conexão ao longo de todos os enlaces da sua rota.

Estudos para resolver o problema RSA vêm sendo desenvolvidos para os cenários de tráfego estático e dinâmico (OZDAGLAR; BERTSEKAS, 2003). Nos cenários estáticos, todas as requisições de conexão são conhecidas antecipadamente, portanto busca-se uma solução para o problema durante o estágio de planejamento da rede. Formulações de programação linear inteira foram propostas para resolver esse problema com o objetivo de minimizar o espectro utilizado para alocar todas as demandas, assim como algoritmos heurísticos desenvolvidos para prover soluções para redes com configurações mais realistas, dado que soluções ótimas puderam ser encontradas apenas para topologias pequenas (CHRISTODOULOPOULOS et al., 2010), (WANG et al., 2011), (KLINKOWSKI; WALKOWIAK, 2011). Sob cenários de tráfego dinâmico, por outro lado, caminhos ópticos devem ser estabelecidos dinamicamente de acordo com a necessidade, ou seja, para cada requisição de conexão um algoritmo é executado em tempo real para determinar qual rota e *slots* serão utilizados na transmissão; se não houver recursos disponíveis, a conexão é bloqueada. Devido à complexidade do problema dinâmico e sua natureza em tempo real, algoritmos heurísticos foram propostos para resolvê-lo, tais como os apresentados por Christodoulopoulos et al. (2013), Shirazipourazad et al. (2013), Wan et al. (2012), e Castro et al. (2012).

Para ambos os cenários, estático e dinâmico, em alguns estudos o problema RSA é decomposto em dois subproblemas: (I) roteamento e (II) alocação de espectro

(*Spectrum Allocation* - SA); que são resolvidos separado e sequencialmente. Dessa forma, uma vez que uma rota é definida para transmitir conexões entre um par origem-destino, uma política de alocação de espectro é aplicada para designar um conjunto de *slots* contíguos disponíveis ao longo dos enlaces da rota. Neste trabalho nós focamos no subproblema de alocação de espectro em redes ópticas elásticas sob tráfego dinâmico, o qual, na maioria dos estudos, é resolvido por meio da aplicação de políticas míopes de alocação de espectro. As políticas míopes são as políticas mais elementares, visto que elas visam tomar boas ações imediatas mas não usam informações ou alguma representação direta das possíveis decisões no futuro. Essas políticas, no entanto, são mais simples e rápidas de se implementar em tempo real.

Um problema relacionado ao tráfego dinâmico é a fragmentação do espectro ao longo do tempo, dada a aleatoriedade em relação às chegadas de conexões e términos de transmissões. Um espectro altamente fragmentado leva ao bloqueio de requisições de conexão que demandam maior largura de banda. Este problema deve ser endereçado para que tais requisições não sejam penalizadas em detrimento das menores. De acordo com Wright et al. (2013), existem duas formas nas quais o problema de fragmentação pode ser abordado: um esquema de desfragmentação, no qual um algoritmo de desfragmentação é aplicado para reconstruir as rotas e realocar o espectro, visando minimizar a quantidade de fragmentos em um ou mais enlaces na rede; ou o desenvolvimento e implementação de um algoritmo de alocação de espectro eficiente, de modo que a alocação de conexões de tamanhos diferentes seja mais justa. Neste trabalho, nosso objetivo é melhorar a utilização do espectro por meio de uma política de alocação de espectro eficiente, sem recorrer a um esquema de desfragmentação.

1.2 Processo Markoviano de Decisão

Considere um sistema cujo comportamento em um tempo t é descrito por um vetor i. Os componentes de i podem representar, por exemplo, a posição de um determinado objeto, volume e temperatura, estoque de um determinado produto, demandas, etc. Supõe-se que no decorrer do tempo tal sistema está sujeito a mudanças de origem determinística ou estocástica, ou seja, os componentes que descrevem seu estado podem sofrer transformações. Caso haja um processo pelo qual pode-se decidir quais transformações podem ser aplicadas ao sistema, este é denominado um processo de decisão, em que uma decisão, ou ação, afeta o comportamento desse sistema. Se mais de uma decisão deve ser tomada no decorrer do tempo, em sequência, temos um processo de decisão sequencial, cujo problema relacionado é denominado problema

de decisão sequencial.

No problema de decisão sequencial deve-se levar em conta a consequência imediata e a longo prazo de cada decisão, em termos de recompensas recebidas ou custos incorridos, de modo que se maximize as recompensas recebidas ao longo do tempo ou se minimize os custos incorridos [2]. A Figura 1.1 ilustra um problema de decisão sequencial no qual, em um determinado instante de tempo, o decisor observa o estado atual do sistema e escolhe uma ação entre um conjunto de ações possíveis para aquele estado. Essa ação gera uma recompensa imediata e pode fazer com que o sistema evolua para um novo estado em um determinado tempo. No próximo instante de decisão o processo se repete.

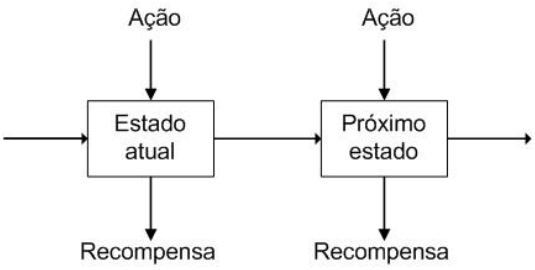

Figura 1.1 - Representação de um problema de decisão sequencial

Fonte: Adaptado de Puterman (2005)

Em alguns sistemas os efeitos das ações tomadas podem não ser previstos de forma determinística, dando origem ao problema de *decisão sequencial sob incerteza*. Dada a natureza estocástica desse problema, a escolha de uma mesma ação para um estado específico pode resultar em uma diferente sequência de estados futuros atingidos pelo sistema, fazendo com que uma sequência fixa de ações não o resolva. Consequentemente, uma solução deve especificar qual ação deve ser escolhida em cada instante do processo de decisão e para qualquer estado que o sistema possa alcançar. Tal solução é denominada *política*, cuja qualidade é medida pela recompensa esperada obtida ao longo do tempo pelos possíveis pares estado-ação gerados por esta. Busca-se, portanto, uma politica ótima que gere a máxima recompensa esperada.

O problema de se tomar uma sequência de decisões em um ambiente que apresenta incerteza surge em diversas áreas como economia, inteligência artificial, teoria de

[2]Para simplicidade de apresentação, neste trabalho nós abordamos apenas a perspectiva de recompensas recebidas a cada ação realizada.

controle e pesquisa operacional (HU; YUE, 2007). Uma maneira de se modelar e resolver tais problemas é por meio dos Processos Markovianos de Decisão (PMDs), nos quais os sistemas são modelados de modo que a propriedade de Markov seja garantida, ou seja, a possibilidade de se prever o "futuro" do processo dado o "presente", independentemente do "passado" (TIJMS, 2003).

Como exemplo de um PMD, retirado de Russell e Norvig (2009), suponha que um agente está situado em um ambiente de grade 4x3 simples como mostra a Figura 1.2(a). Cada estado do sistema é representado pelas coordenadas de uma célula da grade em que o agente pode se situar. A partir do estado inicial indicado deve-se escolher uma ação, a cada instante de tempo, dentre as ações possíveis: movimentar o agente para cima, baixo, esquerda, ou direita. Uma colisão com a parede resulta em nenhum movimento, ou seja, o agente permanece no mesmo quadrado. A interação com o ambiente acaba quando o agente alcança um dos estados finais indicados, recebendo uma recompensa de +1 ou -1. Uma ilustração do modelo de transição estocástico pode ser vista na Figura 1.2(b), em que cada ação alcança o resultado esperado com probabilidade 0,8, porém há uma chance que o agente se mova para ângulos retos relativos à direção pretendida com probabilidade 0,2. Além disso, define-se que o agente receba uma recompensa de -0,04 ao se posicionar em todos os estados exceto os estados terminais.

A solução para esse problema é uma política ótima cuja qualidade é medida pela soma das recompensas esperadas adquiridas no percurso do agente, ou seja, a política que resulta na maior recompensa total esperada. A Figura 1.2(c) ilustra uma política ótima para esse exemplo, na qual pode-se verificar que o maior caminho é escolhido no estado (3,1) em detrimento do menor para não arriscar entrar no estado terminal (4,2), o que resultaria em uma recompensa negativa.

1.3 Aprendizagem por Reforço

Aplicações reais modeladas por PMDs vêm crescendo em vários domínios, como gerenciamento de processos industriais, robótica, e telecomunicações ((SIMÃO et al., 2009), (MOSHARAF et al., 2003), (TACHIBANA et al., 2007)). Buscam-se maneiras eficientes de solucionar problemas para sistemas cada vez mais complexos, cujo espaço de estados pode ser muito grande (SIGAUD; BUFFET, 2010); para esses PMDs de grande porte, computar uma política ótima é geralmente impraticável. Nesse contexto, novas abordagens foram desenvolvidas para prover soluções próximas às ótimas em que não é preciso avaliar exaustivamente o espaço de estados e as ações possíveis. Tais abordagens, de acordo com Powell (2007), são estudadas por comunidades de

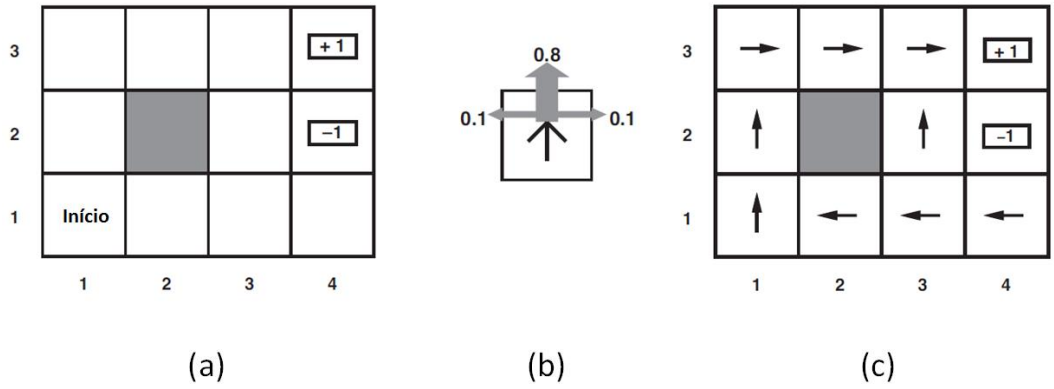

Figura 1.2 - Exemplo simples de um processo markoviano de decisão (a) Representação do ambiente em que o agente interage, a partir do qual define-se os estados e ações disponíveis. (b) Ilustração do modelo de transição estocástico do sistema. (c) Um exemplo de política ótima para o agente.

Fonte: Adaptado de Russell e Norvig (2009)

diferentes áreas do conhecimento e sob diferentes nomes, dentre eles: programação dinâmica adaptativa (*Adaptive Dynamic Programming*), aprendizagem por reforço (*Reinforcement Learning*), programação neuro-dinâmica (*Neuro-Dynamic Programming*) e programação dinâmica aproximada (*Approximate Dynamic Programming*).

A aprendizagem por reforço é simultaneamente um problema e o campo que estuda esse problema e seus métodos de solução (SUTTON; BARTO, 1998). Problemas de aprendizagem por reforço envolvem aprender o que fazer - como mapear situações em ações - a fim de maximizar uma recompensa numérica; o seu objetivo é utilizar recompensas observadas para aprender uma política próxima à ótima para o problema modelado. Em outras palavras, é uma abordagem computacional para o aprendizado a partir da interação entre um agente e o seu ambiente.

Utilizando o exemplo da Figura 1.2, no âmbito da aprendizagem por reforço o agente executa vários experimentos compostos por uma série de movimentos no ambiente. Em cada experimento o agente começa no estado inicial e avança, seguindo uma política pré-definida, por uma sequencia de transições de estados até que um dos estados terminais seja alcançado. O objetivo é utilizar as informações sobre as recompensas obtidas para aprender a recompensa esperada associada com cada estado não-terminal da grade. Nesse caso o agente da Figura 1.2 não teria o conhecimento prévio do PMD completo, ou seja, informações sobre todos os estados, ações, transições e recompensas possíveis. Em algumas aplicações tal conhecimento não pode

ser obtido, seja por restrições do tamanho ou natureza do problema abordado.

A aprendizagem por reforço é diferente do aprendizado supervisionado, no qual a aprendizagem é feita a partir de um conjunto de treinamento e exemplos providos por um supervisor externo e o objetivo é que o sistema extrapole, ou generalize, suas respostas para que aja corretamente em situações não presentes no conjunto de treinamento. Em problemas interativos é geralmente impraticável obter exemplos de comportamentos esperados que são corretos e representativos de todas as situações nas quais um agente deve atuar. Para isso, um agente deve ser capaz de aprender por sua própria experiência. A aprendizagem por reforço difere também do aprendizado não-supervisionado, que geralmente remete à se encontrar estruturas escondidas em coleções de dados não classificados. Embora pode-se pensar que a aprendizagem por reforço seja um tipo de aprendizado não-supervisionado por não se basear em exemplos de comportamento correto, nela tenta-se maximizar uma recompensa ao invés de encontrar estruturas escondidas. A aprendizagem por reforço é considerada, então, como um outro paradigma de aprendizado de máquina, estudado em conjunto com o aprendizado supervisionado e não-supervisionado (SUTTON; BARTO, 1998).

1.4 Objetivo

No contexto das redes ópticas elásticas sob tráfego dinâmico, os *slots* disponíveis no espectro óptico podem ficar divididos em fragmentos devido a requisições de conexão demandando quantidades diferentes de *slots*, e à aleatoriedade inerente das chegadas das requisições e o tempo de transmissão de cada uma. Caso o espectro fique fragmentado, requisições de conexão que demandam um maior número de *slots* contíguos podem ser penalizadas em relação às menores e bloqueadas pela falta de recursos disponíveis. Desse modo, é importante desenvolver e implementar políticas de alocação de espectro que levem em consideração o desempenho de todos os tipos de requisição de conexão da rede.

Nosso objetivo é encontrar uma política de alocação de espectro que leve em conta a dinâmica intrínseca da rede na escolha das ações, ou seja, uma política que não seja míope. Com essa finalidade nós modelamos um enlace de uma rede óptica elástica como um PMD, a partir do qual podemos encontrar uma política de alocação de espectro ótima. Vale ressaltar que nosso modelo abrange apenas um enlace da rede, que se torna um pré-requisito para modelar redes flexíveis com topologias arbitrárias em trabalhos futuros.

Uma vez que uma política de alocação de espectro é encontrada e implementada,

podemos avaliar sua performance de acordo com alguma métrica, como, por exemplo, a probabilidade de bloqueio de conexões na rede. Dado que a nossa política é obtida por meio da solução de um PMD, ao aplicarmos essa política fixando uma ação para cada estado do sistema, obtemos uma cadeia de Markov a tempo discreto (*Discrete-Time Markov Chain* - DTMC), pela qual podemos computar analiticamente as medidas de desempenho correspondentes. Além disso, nosso modelo permite que outras políticas sejam implementadas ao se enumerar todos os estados e aplicar, para cada um deles, a regra correspondente; da mesma forma uma cadeia de Markov pode ser derivada e as medidas de desempenho calculadas.

Para instâncias mais realistas do problema, nas quais o espectro pode ser discretizado em até centenas de *slots*, torna-se impraticável computacionalmente encontrar uma política ótima para o PMD, assim como calcular as medidas de desempenho analiticamente via cadeia de Markov. Para esses casos, nós propomos a utilização de um algoritmo de aprendizagem por reforço, denominado *Relaxed Semi-Markov Average Reward Technique* (Relaxed-SMART), aliado a um esquema de aproximação de função para armazenamento e aplicação das políticas de maneira eficiente. O *Relaxed-SMART* foi inicialmente proposto por Gosavi (2004), e uma modificação foi apresentada posteriormente por Gosavi (2011). Além disso, nós utilizamos simulações para computar as medidas de desempenho das políticas obtidas.

As questões centrais abordadas nessa tese são: podemos obter políticas de alocação de espectro, levando em conta a dinâmica do sistema, que resultem em um melhor desempenho da rede em relação às políticas míopes comumente utilizadas? Podemos aplicar essa mesma abordagem para instâncias mais realistas do problema? A partir dessas perguntas, nossa tese é a seguinte: *Dado um enlace de redes ópticas elásticas sob tráfego dinâmico, políticas de alocação de espectro obtidas a partir de um PMD podem obter desempenho superior às políticas míopes.*

1.5 Estrutura e Contribuições da Tese

Segue abaixo a organização deste documento em conjunto com as nossas contribuições:

- Capítulo 2 - Descrevemos brevemente a arquitetura das redes ópticas elásticas e o problema intrínseco de roteamento e alocação de espectro. Focamos no subproblema de alocação de espectro e apresentamos alguns métodos encontrados na literatura utilizados para resolvê-lo, tais como *First-Fit* e *Best-Fit*;

- Capítulo 3 - Formalizamos brevemente a teoria dos PMDs a Tempo Discreto e a Tempo Contínuo bem como os os principais métodos de resolução para encontrar políticas ótimas para esses modelos: algoritmo de iteração de políticas, algoritmo de iteração de valores e algoritmo de iteração de valores relativo. Apresentamos também o conceito das probabilidades limite, que permitem o cálculo de medidas de desempenho das políticas de controle aplicadas no sistema estudado. O conteúdo desse capítulo é baseado nos principais livros da área, tais como Puterman (2005) e Tijms (1995);

- Capítulo 4 - Nesse capítulo nós propomos um PMD a Tempo Contínuo para modelar o problema de alocação de espectro em um enlace de uma rede óptica elástica sob tráfego dinâmico. Além disso, cadeias de Markov são utilizadas para calcular analiticamente medidas de desempenho tanto da política ótima quanto das políticas míopes;

- Capítulo 5 - Uma descrição mais detalhada da aprendizagem por reforço é feita nesse capítulo. Além disso, apresentamos um algoritmo de aprendizagem por reforço, denominado *Relaxed-SMART*, aliado a um esquema de aproximação de função para encontrar políticas para instâncias realistas do problema de alocação de espectro;

- Capítulo 6 - Nesse capítulo descrevemos primeiramente os parâmetros de configuração das redes utilizadas nos experimentos numéricos e os seus padrões de tráfego. Dois cenários são definidos: um cenário no qual é possível comparar analiticamente as medidas de desempenho das políticas implementadas e outro mais realista e em concordância com outros estudos realizados na área. Resultados para ambos os cenários são então apresentados e discutidos;

- Capítulo 7 - As considerações finais deste trabalho são apresentadas nesse capítulo em conjunto com sugestões para trabalhos futuros.

Por fim, no Apêndice A mostramos detalhes de implementação do modelo analítico do capítulo 4, enquanto que no Apêndice B descrevemos o esquema de atualização do vetor de pesos do modelo de regressão linear utilizado no algoritmo proposto no capítulo 5.

2 O PROBLEMA DE ALOCAÇÃO DE ESPECTRO EM REDES ÓPTICAS ELÁSTICAS

A transmissão de dados pode ser realizada em diferentes meios, como fios de cobre (par trançado e cabo coaxial), fibras ópticas e ondas de rádio. Cada meio apresenta características próprias, dentre elas, o volume de informações que este é capaz de transportar em um determinado tempo. Quanto maior a quantidade de informações que se deseja transmitir em um período de tempo, maior a taxa de transmissão requerida.

Há uma relação direta entre a taxa de transmissão, cuja unidade de medida é bps (número de bits transmitidos por segundo), e a faixa das frequências (em Hertz), que podem ser utilizadas para transportar dados pelo meio (HORAK, 2008). Ilustra-se na Figura 2.1, no espectro eletromagnético, a faixa de frequência em que operam alguns dispositivos de comunicação. Nota-se a grande diferença nas frequências de operação da fibra óptica em relação aos demais meios, como, por exemplo, o par trançado e o cabo coaxial.

Neste capítulo nós apresentamos uma técnica para o melhor aproveitamento da capacidade das fibras ópticas, denominada multiplexação por divisão em comprimento de onda (WDM). Descrevemos também a arquitetura das redes ópticas elásticas, que visa tornar as redes mais flexíveis. São considerados então os problemas intrínsecos às redes que fazem uso dessas técnicas, que são, respectivamente, o problema de roteamento e atribuição de comprimento de onda (RWA) e roteamento e alocação de espectro (RSA). Por fim, focamos no subproblema de alocação de espectro e apresentamos alguns métodos encontrados na literatura utilizados para resolvê-lo.

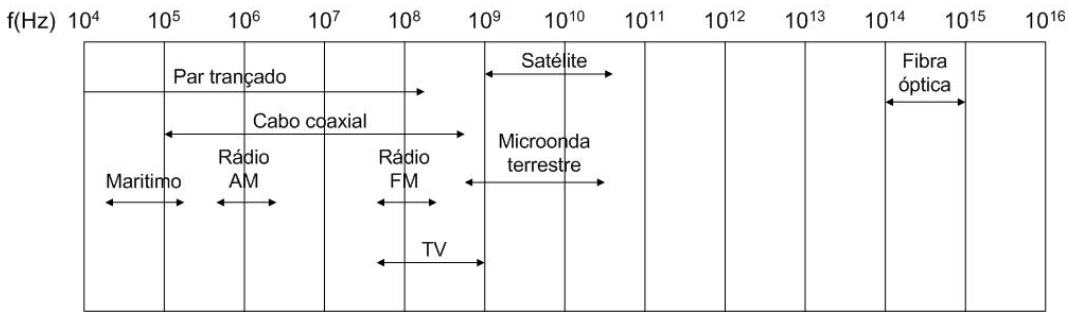

Figura 2.1 - Faixas de frequência do espectro eletromagnético utilizadas na comunicação.

Fonte: Adaptado de Tanenbaum (2002)

2.1 Redes WDM e o Problema RWA

Ao se adotar a tecnologia da fibra óptica, a capacidade de transmissão pode ultrapassar a casa dos 50 Tb/s (terabits por segundo). A transmissão por parte dos clientes e aplicações, no entanto, é limitada pelo uso de equipamentos eletrônicos com taxas na ordem de gigabits por segundo, gerando um gargalo óptico-elétrico (SIVALINGAM; SUBRAMANIAM, 2000). Uma solução para o gargalo óptico-elétrico na transmissão por fibras ópticas é a multiplexação por divisão em comprimento de onda, na qual vários clientes utilizam a mesma fibra óptica por meio de comprimentos de onda (frequências) distintos. As redes que adotam tal solução são denominadas redes WDM. Ao se utilizar canais em paralelo com comprimentos de onda distintos, a taxa de transmissão aumenta de forma linear com o número de canais, tornando a técnica WDM adequada para um melhor aproveitamento da capacidade de transmissão das fibras ópticas. Como exemplo, se 90 canais com 10 Gb/s cada forem multiplexados em uma fibra, a taxa de transmissão chega a 900 Gb/s.

A técnica WDM pode ser vista como uma variação da multiplexação por divisão em frequência (*Frequency Division Multiplexing* - FDM). A técnica FDM é comumente utilizada em circuitos analógicos, nos quais o espectro da frequência é dividido em bandas de frequência, tendo cada usuário a posse exclusiva de uma dessas bandas. Em redes ópticas convencionou-se usar o comprimento de onda no lugar da frequência para denotar a posição de uma onda no espectro eletromagnético.

Para a transmissão de dados em uma rede WDM, uma conexão na camada óptica entre cada par origem-destino deve ser estabelecida. Para tanto, determina-se um caminho, ou rota, na rede para cada um destes pares e aloca-se um comprimento de onda disponível em todos os enlaces deste caminho. Tais caminhos são denominados caminhos ópticos (*lightpaths*), os quais não requerem processamento ou *buffering* (armazenamento do dado em memória temporária para posterior transmissão) em nós intermediários (CHLAMTAC et al., 1992).

A Figura 2.2 ilustra caminhos ópticos criados entre clientes representados por círculos, associados a dispositivos comutadores (*switches*), que utilizam dois comprimentos de onda distintos, representados nas cores vermelha e verde. Ressalta-se que, em cada enlace da rede, os caminhos ópticos que o utilizam devem estar obrigatoriamente em diferentes comprimentos de onda.

Um aspecto na criação dos caminhos ópticos deve ser observado: estes devem ocupar sempre o mesmo comprimento de onda durante todo o seu trajeto, obedecendo a

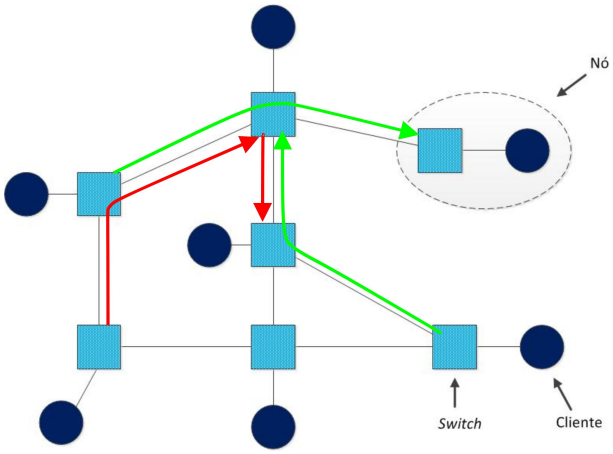

Figura 2.2 - Exemplos de caminhos ópticos em uma rede WDM.

restrição de continuidade de comprimento de onda (MOSHARAF et al., 2005). Na saída de um nó intermediário, por exemplo, o comprimento de onda requerido pode estar sendo utilizado em outra transmissão, gerando um bloqueio no estabelecimento da conexão.

O problema de se escolher a rota e o comprimento de onda que serão utilizados por cada caminho óptico é denominado RWA. Dado um grafo G correspondente a uma rede WDM e um conjunto de requisições de conexão K, deve-se encontrar um caminho óptico (p, w) para cada conexão, em que p é uma rota e w um comprimento de onda, de modo que dois caminhos que compartilhem um arco de G não sejam atribuídos ao mesmo comprimento de onda (JAUMARD et al., 2004). Encontrar uma boa solução para o problema RWA é importante para o aumento da eficiência das redes WDM, já que se tem a garantia de que mais clientes serão atendidos, e, consequentemente, menos chamadas serão bloqueadas.

Para o caso estático, no qual as requisições de conexão são conhecidas *a priori*, Ramaswami e Sivarajan (1995) formulam o problema RWA como Programação Linear Inteira (PLI) cujo objetivo é maximizar o número de conexões que podem ser estabelecidas para um número fixo de comprimentos de onda. Para uma revisão de algumas formulações para esse problema, direcionamos o leitor para o trabalho de Jaumard et al. (2004).

Sob um tráfego dinâmico, decisões devem levar em conta o efeito das futuras requisições de conexão e a disponibilidade dos recursos na rede, e o objetivo torna-se minimizar a probabilidade de bloqueio das chamadas. Para esse caso, heurísticas são

propostas, a exemplo de Bisbal et al. (2004) e Saengudomlert et al. (2006).

2.2 Redes Ópticas Elásticas e o Problema RSA

Nas redes WDM o espectro óptico é discretizado em uma grade com faixas fixas de 50 GHz, alinhadas com a grade de frequência proposta pelo ITU-T. Tal granularidade fixa resulta em algumas desvantagens: um comprimento de onda inteiro é alocado para uma conexão mesmo quando a largura de banda demandada não é grande o bastante para ocupar toda a sua capacidade; e quaisquer dois comprimentos de onda consecutivos devem ser separados por frequências de banda de guarda (*guard-band*), para evitar interferência nos sinais. Caso uma conexão não possa ser transmitida em apenas um comprimento de onda, ou seja, sua largura de banda é maior que a capacidade do comprimento de onda, mais comprimentos de onda devem ser alocados para transmiti-la com as bandas de guarda apropriadas entre eles.

Para superar as desvantagens das redes WDM foi proposta a arquitetura das redes ópticas elásticas, que tem como base a tecnologia de multiplexação por divisão de frequências ortogonais (*Orthogonal Frequency Division Multiplexing* - OFDM). A tecnologia OFDM permite a transmissão de dados em alta velocidade usando múltiplos canais de baixa velocidade, os *slots*, que podem ser sobrepostos no espectro. Dessa forma, frequências de banda de guarda ainda devem ser alocadas entre duas conexões adjacentes, porém não entre dois *slots* utilizados para transmitir uma conexão. Uma pesquisa sobre OFDMs e suas tecnologias relacionadas pode ser encontrada no trabalho de Zhang et al. (2013).

As redes ópticas elásticas podem trabalhar eficientemente com demandas de tráfego misto variando de Gb/s à Tb/s, devido à possibilidade de ajustar dinamicamente os *slots* alocados por cada conexão. Além disso, ao contrário de redes WDM, nas quais frequências de banda de guarda são pré-alocadas e fixas, qualquer *slot* pode ser alocado como banda de guarda no processo de estabelecimento de uma conexão.

A diferença entre as duas arquiteturas de rede pode ser vista no exemplo da Figura 2.3, no qual existem três tipos de conexão com requerimentos de tráfego variando de 10 Gb/s a 400 Gb/s. Considerando a grade fixa da rede WDM, um comprimento de onda é alocada para conexões que demandam 10 Gb/s, embora estas não usem completamente sua capacidade. Cada conexão de 400 Gb/s deve ser transmitida como duas conexões diferentes separadas por uma frequência de banda de guarda. Na grade flexível das redes ópticas elásticas, no entanto, frequências de banda de guarda não são fixas e não há a necessidade de dividir cada conexão de 400 Gb/s

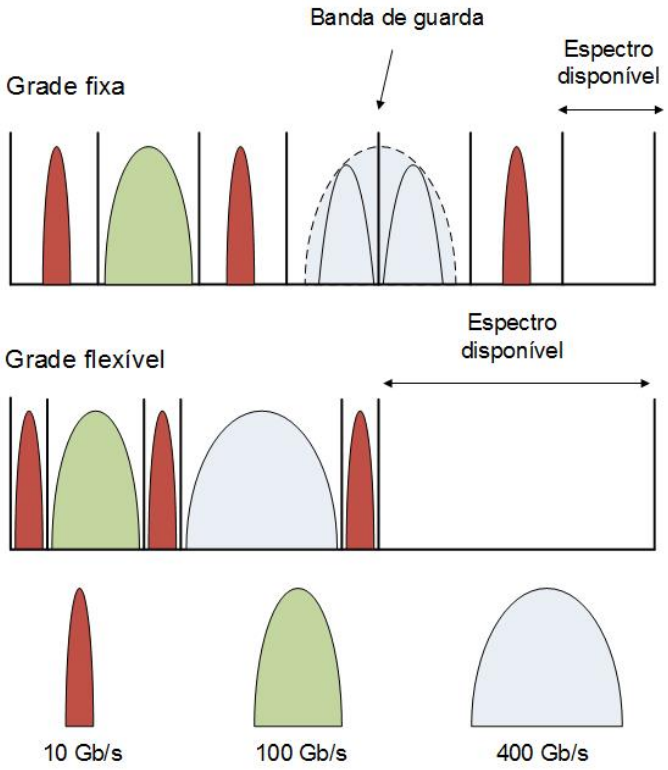

Figura 2.3 - Diferenças entre as grades fixa e flexível.

Fonte: Adaptado de Gerstel et al. (2012)

para transmiti-la pela rede, e, portanto, há uma maior fatia do espectro disponível para a transmissão de futuras conexões.

Mais especificamente, como pode ser visto na Figura 2.4(a), cada conexão é transmitida nas redes elásticas por meio de um canal óptico composto por um ou mais *slots* consecutivos. Cada canal tem uma frequência central (*Central Frequency* - CF) e o espectro é ocupado simetricamente em volta dessa frequência; a largura de cada *slot* é duas vezes o espaçamento entre duas frequências centrais consecutivas. Neste exemplo, o canal é composto por três *slots*. Trabalhos de padronização relacionados às redes ópticas elásticas vêm sendo conduzidos, a exemplo da ITU-T que incluiu a definição de grade flexível em ITU-T (2012); tal grade tem um espaçamento de frequência central de 6,25 GHz e um *slot* de 12,5 GHz.

Para uma dada requisição de conexão, a demanda de tráfego pode ser traduzida no número de *slots* necessários para transmiti-la entre um nó origem e um nó destino na rede, considerando configurações tais como a técnica de modulação aplicada. Desse modo, pode-se representar um enlace de uma rede óptica elástica como no exemplo

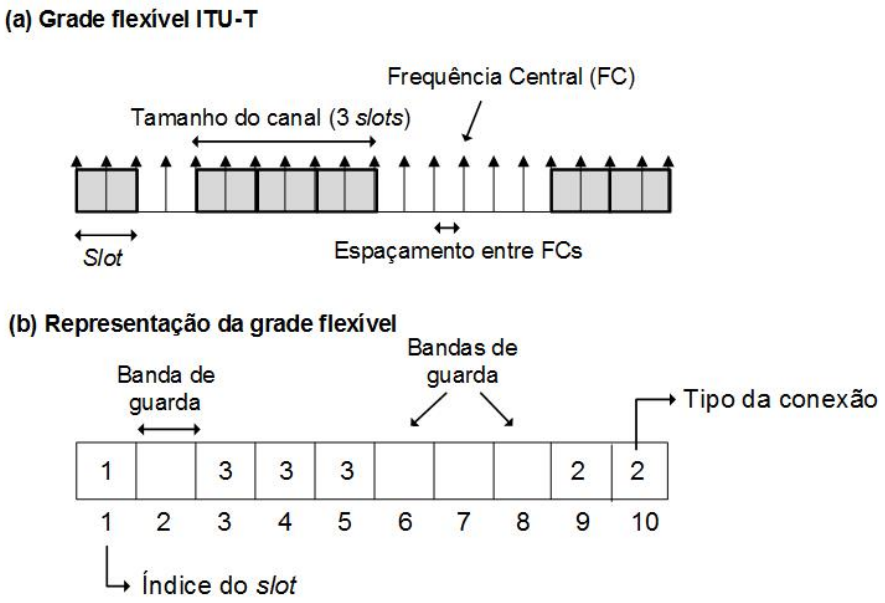

Figura 2.4 - Grade flexível: (a) Exemplo de configuração de *slot* e canal seguindo a grade flexível sugerida pelo ITU-T. (b) Uma representação simplificada dessa configuração.

da Figura 2.4(b), na qual o espectro é discretizado em uma grade de tamanho 10 com requisições de conexão de tipos 1, 2, e 3 que demandam 1, 2 e 3 *slots*, respectivamente. Além disso, um *slot* é usado como banda de guarda.

A granularidade menor das redes ópticas elásticas traz flexibilidade na transmissão de conexões com requerimentos de tráfego distintos. No entanto, assim como no caso das redes WDM, ela também introduz novos desafios que devem ser abordados para planejar e utilizar eficientemente tais redes. Se uma conexão requer mais de um *slot* para ser transmitida, tais *slots* devem ser contíguos, obedecendo a restrição de contiguidade de espectro. A restrição de continuidade de espectro, por outro lado, considera que os mesmos *slots* contíguos devem ser alocados em todos os enlaces que pertencem ao caminho óptico. Enquanto a primeira restrição pode levar a um espectro altamente fragmentado em um enlace, a segunda contribui para a fragmentação em todos os enlaces da rota.

O problema RWA passa a ser substituído pelo problema RSA, no qual, deve-se rotear e alocar os *slots* para acomodar as conexões. Tal problema também emerge tanto para o caso estático quanto para o dinâmico. No tráfego estático, formulações para resolver o problema por meio de PLI visando minimizar o número máximo de *slots* requeridos para atender a demanda foram propostas, a exemplo de Wang et

al. (2011) e Klinkowski e Walkowiak (2011).

Para o tráfego dinâmico, Wan et al. (2011) introduz dois algorimos. O primeiro é uma modificação do algoritmo de *Dijkstra*, no qual toda vez que um enlace é considerado para fazer parte de uma rota, ele é checado para verificar se há *slots* contíguos disponíveis em comum com os outros enlaces já presentes na rota. Caso haja disponibilidade de *slots* o enlace é adicionado, caso contrário o algoritmo considera outros enlaces mesmo que o custo seja maior. O segundo algoritmo constrói uma árvore das rotas distintas com *slots* disponíveis e busca quais dessas rotas resultam em um menor custo. No trabalho de Shirazipourazad et al. (2013) são apresentados resultados analíticos relacionados ao problema RSA para redes com topologia em anel. Dada essa topologia, um dos enlaces é removido aleatoriamente para que o roteamento torne-se trivial e a política de alocação de espectro *First-Fit* é utilizada. Além disso, os autores propõem heurísticas para redes com topologias arbitrárias baseadas na modificação dos algoritmos de *Dijkstra* e dos K-menores caminhos para o roteamento e o *First-Fit* para alocação de espectro. Outra modificação do algoritmo de *Dijkstra* pode ser vista no trabalho de Castro et al. (2012).

Uma revisão de algumas técnicas utilizadas para lidar com o problema RSA, tanto o caso estático quanto o dinâmico, pode ser encontrada no trabalho de Talebi et al. (2014).

2.2.1 Alocação de Espectro

Para ambos os tráfegos, estático e dinâmico, em alguns estudos o problema RSA é decomposto em dois subproblemas: (I) roteamento e (II) alocação de espectro; que são resolvidos separada e sequencialmente. Dessa forma, uma vez que uma rota é definida para transmitir conexões entre um par origem-destino, uma política de alocação de espectro é aplicada para designar um conjunto de *slots* contíguos disponíveis ao longo dos enlaces da rota.

Neste trabalho consideramos apenas um enlace, então não há roteamento algum e só o problema de alocação de espectro é tratado. Nós focamos no cenário de tráfego dinâmico, o qual, na maioria dos estudos, é resolvido pela aplicação de políticas míopes, que são mais simples e rápidas de se executar em tempo real. Como exemplo de políticas míopes de alocação de espectro comumente utilizadas na literatura nós podemos citar:

- *Random-Fit*: na qual um conjunto de *slots* é selecionado aleatoriamente a

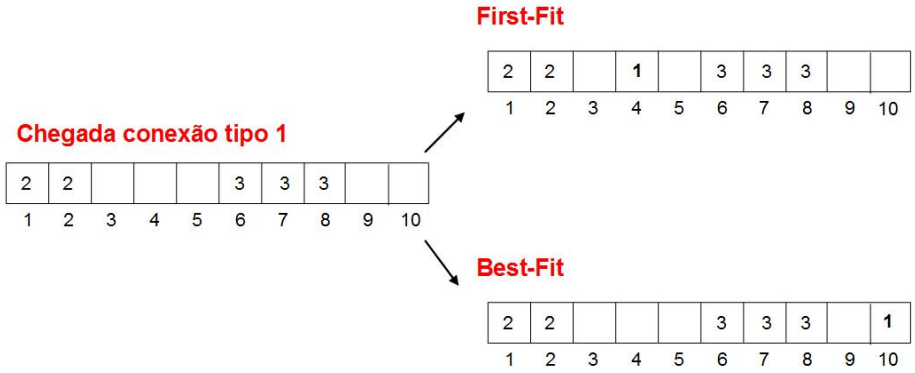

Figura 2.5 - Exemplo de aplicação de regras de alocação de espectro das políticas *First-Fit* e *Best-Fit*.

partir dos disponíveis;

- *First-Fit*: na qual todos os *slots* são numerados, normalmente da esquerda para a direita em ordem ascendente, e o conjunto de *slots* começando com o menor número é escolhido a partir dos disponíveis;
- *Best-Fit*: em que o menor bloco de *slots* contíguos disponíveis que pode satisfazer a requisição é escolhido.

Uma ilustração do funcionamento das políticas *First-Fit* e *Best-Fit* pode ser vista na Figura 2.5, em que tomamos como base o cenário da Figura 2.4.b. Nesse exemplo, duas conexões estão sendo transmitidas, uma do tipo 2 e outra do tipo 3. Com a chegada de requisição de uma conexão do tipo 1, torna-se possível alocá-la nos *slots* de índice 4 ou 10, pois um *slot* deve ser usado como banda de guarda. A política *First-Fit* faz a alocação da conexão no *slot* de índice 4, o primeiro disponível da esquerda para a direita. A política *Best-Fit* aloca no índice 10, pois o bloco contíguo tem tamanho dois (composto pelos índices 9 e 10) enquanto o outro tem tamanho três (índices 3, 4 e 5).

Outras estratégias de alocação de espectro foram estudadas e propostas, como a utilização de algoritmos evolutivos por Almeida et al. (2013), que busca a melhor ordenação de *slots* para o *First-Fit* visando minimizar a probabilidade de bloqueio de requisições. Os autores mostram que essa estratégia supera o algoritmo *First-Fit* convencional, embora ela demande maior esforço computacional. Uma política míope denominada *Exact-Fit* foi proposta por Rosa et al. (2012), a qual procura por um bloco de *slots* contíguos disponíveis que corresponda exatamente ao número

de *slots* da requisição e as bandas de guarda correspondentes. Se tal bloco existe ele é utilizado para alocar a conexão, caso contrário, o maior conjunto livre de *slots* contíguos é escolhido. Os autores mostram que a *Exact-Fit* supera outras políticas míopes em termos de probabilidade de bloqueio de conexão.

3 PROCESSO MARKOVIANO DE DECISÃO

Neste capítulo nós descrevemos brevemente a teoria dos Processos Markovianos de Decisão[1], utilizados para modelar problemas de tomada de decisão sequencial sob incerteza. Apresentamos os PMDs a Tempo Discreto, que modelam sistemas nos quais as decisões devem ser tomadas em um conjunto de pontos fixos no tempo, porém temos como foco principal os PMDs a Tempo Contínuo, onde as decisões são tomadas em instantes aleatórios de tempo, na ocorrência de algum evento.

Apresentamos também dois métodos de resolução dos PMDs: algoritmo de iteração de valores e algoritmo de iteração de políticas[2]. Uma variação do algoritmo de iteração de valores, o algoritmo de iteração de valores relativo, que foi utilizada na implementação deste trabalho também é descrita. Por fim, descrevemos brevemente o conceito das probabilidades limite, que permitem o cálculo de medidas de desempenho das políticas de controle aplicadas no sistema estudado.

Neste trabalho nós modelamos o problema de alocação de espectro como um PMD a Tempo Contínuo cujas políticas são avaliadas pela recompensa média obtida ao longo do tempo sob um horizonte de planejamento infinito. Desse modo, o conteúdo desse capítulo é apresentado para PMDs com essas características: tempo contínuo; horizonte de planejamento infinito; e critério de recompensa média.

3.1 Processo Markoviano de Decisão a Tempo Discreto

O PMD a Tempo discreto é descrito como uma quíntupla

$$\{T,\ S,\ A(s),\ p_t(s,a,s'),\ r_t(s,a)\},$$

em que o conjunto de instantes de decisão é representado por $T = \{1, 2, ...\}$; S é o espaço de estados que o sistema pode assumir; $A(s)$ o conjunto de ações possíveis para o estado $s \in S$; $p_t(s,a,s')$ a probabilidade de transição do estado $s \in S$ para o estado $s' \in S$ ao se escolher a ação $a \in A(s)$ no tempo t; e, por fim, $r_t(s,a)$ é a recompensa ao se tomar a ação $a \in A(s)$ no estado $s \in S$ no tempo t.

No instante t, ao se observar o sistema no estado s, uma ação a deve ser escolhida. Feito isso, tem-se:

[1] Não foram reproduzidos neste capítulo os teoremas e provas associados aos resultados aqui considerados, podendo estes ser encontrados em livros que são referência na área, tais como Puterman (2005) e Tijms (1995).

[2] Informações sobre uma terceira abordagem, baseada em métodos de programação linear, podem ser encontradas em Tijms (1995).

1) uma recompensa imediata $r_t(s, a)$;

2) uma probabilidade $p_t(s, a, s')$ de transição para o estado s' em que:

$$\sum_{s' \in S} p_t(s, a, s') = 1$$

Em um PMD apenas as informações proporcionadas pelo estado e ação atuais são necessários para o cálculo das recompensas, conjunto de ações possíveis e probabilidades de transição do sistema. Busca-se uma boa descrição de cada estado do sistema de modo que seja respeitada a propriedade de Markov, que garante a ausência de memória. Em outras palavras, torna-se possível prever o "futuro" do processo dado o "presente", independentemente do "passado" (TIJMS, 2003).

Uma característica a ser considerada ao se modelar PMDs é se o horizonte de planejamento é finito ou infinito. No primeiro caso, há um instante N fixo em que se acaba o processo de tomada de decisão, já no segundo considera-se que não há um limite fixo para a duração do processo.

3.2 Processo Markoviano de Decisão a Tempo Contínuo

Em alguns casos, o tempo entre os instantes de decisão pode não ser constante, mas sim aleatório com $T = (0, \infty]$. O sistema pode então ser modelado como um Processo Semi-Markoviano de Decisão, generalização que, segundo Puterman (2005), viabiliza a escolha de uma ação cada vez que o estado do sistema muda, e possibilita modelar a evolução do sistema em tempo contínuo. Nesse caso, a recompensa imediata ao se tomar uma decisão pode ser substituída pela recompensa esperada até o próximo instante de decisão. Definimos então as seguintes modificações em relação ao PMD a Tempo Discreto:

- $y(s, a)$ é o tempo esperado até o próximo instante de decisão se a ação a é escolhida no estado s, tal que $y(s, a) > 0$ para todos os estados e ações;
- $\bar{r}(s, a)$ é a recompensa esperada até o próximo instante de decisão se a ação a é escolhida no estado s.

Um caso especial desse processo é o PMD a Tempo Contínuo, em que o intervalo entre dois instantes de decisão sucessivos é exponencialmente distribuído. Tal modelo pode ser utilizado, por exemplo, em sistemas de filas, processos populacionais, e controle de doenças infecciosas (GUO; HERNÁNDEZ-LERMA, 2009).

Em cada estado de um PMD a Tempo Contínuo uma quantidade de eventos pode causar uma transição. A ocorrência de um evento que causa a transição do estado $s \in S$ para o estado $s' \in S$, com $s \neq s'$, ao se escolher a ação $a \in A(s)$, segue uma distribuição exponencial com parâmetro $q(s, a, s')$. Levando-se em conta todas as transições possíveis do estado s, a taxa total de saída desse estado é o mínimo entre exponenciais, cujo parâmetro é a soma das taxas de cada transição, ou seja, $\sum_{\bar{s} \neq s} q(s, a, \bar{s})$, $\bar{s} \in S$.

A probabilidade $p(s, a, s')$ de transição pode ser então calculada pela razão entre a taxa de transição do estado s para o estado s' e a taxa total de saída do estado:

$$p(s, a, s') = \frac{q(s, a, s')}{\sum_{\bar{s} \neq s} q_(s, a, \bar{s})} .$$

Para um estudo mais específico sobre PMDs a Tempo Contínuo, Guo e Hernández-Lerma (2009) pode ser consultado.

3.3 Política de Controle Estacionária

Para ambos os casos, discreto ou contínuo, uma política π é um conjunto de regras que determina qual ação deve ser escolhida a cada instante de decisão. Tal política pode indicar de forma determinística qual ação aplicar ou pode definir uma distribuição de probabilidade sobre as possíveis ações aplicáveis. Neste trabalho são focadas as políticas determinísticas markovianas, caracterizadas pelo mapeamento direto de estados em ações por uma função $d_t : s \to A(s)$. Dessa forma, $\pi = (d_1, d_2, ...)$ especifica qual ação é escolhida quando o sistema ocupa o estado s no instante t, de modo que, para cada $s \in S$, $d_t(s) \in A(s)$.

Dado que o horizonte de planejamento é infinito, consideram-se apenas as políticas estacionárias. Uma politica é dita estacionária quando apresenta a forma $\pi = (d, d, ...)$ em que cada ação é associada a um estado independente do instante t de observação, obtendo uma função $d : s \to A(s)$. Em outras palavras, a política sempre prescreve a mesma ação quando o sistema é observado em um determinado estado.

3.3.1 Critério de Recompensa Média a um Horizonte Infinito

A qualidade de uma política associada a um PMD é avaliada baseando-se na soma cumulativa esperada das recompensas obtidas no decorrer da trajetória do processo,

de acordo com um critério de otimalidade. Os principais critérios estudados, de acordo com Sigaud e Buffet (2010), são:

- Total em horizonte finito: $E[r_0 + r_1 + r_2 + ... + r_{T-1}|s_0]$;
- Descontado com fator γ: $E[r_0 + \gamma r_1 + \gamma^2 r_2 + ... + \gamma^t r_t + ...|s_0]$;
- Médio em horizonte infinito: $\lim_{t\to\infty} \frac{1}{t} E[r_0 + r_1 + r_2 + ... + r_{t-1}|s_0]$;

em que r_t representa a recompensa obtida no instante de decisão t. Percebe-se que os critérios são definidos em relação a um operador de esperança $E[*]$, levando em conta as recompensas acumuladas na trajetória de uma dada política e um estado inicial s_0. Para o critério descontado, o fator de desconto γ, com $0 \leq \gamma < 1$, descreve a preferência em recompensas atuais sobre recompensas futuras. Quanto mais próximo γ está de 0, menos significativas são as recompensas futuras.

Apesar de grande parte das aplicações de PMDs ser focada no critério descontado, em muitos problemas de engenharia as medidas de desempenho não são descritas facilmente em termos econômicos, e, portanto, pode ser preferível comparar políticas com base na sua recompensa média a longo prazo. Alguns exemplos dessas medidas são: tempo médio de espera de um serviço em uma fila, número médio de clientes em um sistema e porcentagem média de demandas atendidas em um sistema de estoque. Neste trabalho, buscamos uma política markoviana estacionária que maximize a recompensa média esperada em um horizonte de planejamento infinito. A escolha do critério de recompensa média esperada deu-se pelo fato deste ser apropriado para problemas relacionados a telecomunicações (TIJMS, 1995), em que o sistema opera por um longo período de tempo e deseja-se que seu desempenho seja consistente a longo prazo.

A recompensa média esperada de um PMD a Tempo Contínuo começando no estado s e seguindo a política estacionária π, denotada por $g^\pi(s)$, pode ser definida como

$$g^\pi(s) = \liminf_{T\to\infty} \frac{E[\sum_{k=1}^{T} r(s_k, \pi(s_k), s_{k+1})|s_1 = s]}{E[\sum_{k=1}^{T} t(s_k, \pi(s_k), s_{k+1})|s_1 = s]} \tag{3.1}$$

em que $\pi(s_k)$ denota a ação escolhida no estado s_k quando a política π é seguida; $r(s_k, \pi(s_k), s_{k+1})$ e $t(s_k, \pi(s_k), s_{k+1})$ denotam, respectivamente, a recompensa recebida na k-ésima transição e o tempo gasto nela. Sendo s_k o estado do sistema no

k-ésimo instante de decisão, o processo estocástico embutido é uma cadeia de Markov a tempo discreto. Ao se trabalhar com a suposição de que essa cadeia de Markov embutida não possui dois conjuntos fechados disjuntos[3], a recompensa média esperada de todas as políticas estacionárias não varia com o estado inicial (PUTERMAN, 2005). Assim, $g^{\pi}(s) = g^{\pi}$, $\forall s \in S$.

3.4 Métodos de Resolução de PMDs

De acordo com Tijms (2003), para um PMD sob o critério de recompensa média e conjunto finito de estados e ações, existe uma política estacionária que é ótima. Uma política π^* é ótima se $g^{\pi^*} \geq g^{\pi}$ para todas as políticas estacionárias π. Além disso, existe um escalar g^* e valores $V^*(s)$, denominados valores relativos associados à política ótima, que satisfazem o seguinte sistema de equações

$$V^*(s) = \bar{r}(s, \pi^*(s)) - g^* y(s, \pi^*(s)) + \sum_{s' \in S} p(s, \pi^*(s), s') V^*(s') \ , \ \forall s \in S \qquad (3.2)$$

em que g^* é a recompensa média máxima obtida pelo PMD a Tempo Contínuo ao se aplicar uma política ótima π^*[4]. A partir desses valores desenvolvem-se as estratégias para obtenção de uma política ótima para um PMD.

3.4.1 Algoritmo de Iteração de Políticas

O algoritmo de iteração de políticas constrói uma sequência de políticas com melhor desempenho até que a política ótima seja encontrada. A ideia é começar com uma política escolhida aleatoriamente, calcular a recompensa média e os valores relativos de cada estado por meio de um sistema de equações lineares, e utilizar tais valores para encontrar uma nova política. O processo continua até que nenhuma melhoria seja possível, como mostra o Algoritmo 1. O algoritmo de iteração de políticas converge após um número finito de iterações para a política com recompensa média

[3] Um conjunto de estados C é fechado se $p(s, \pi(s), s') = 0$ para todo $s \in C$ e $s' \notin C$

[4] Pode-se obter interpretação semelhante para PMDs a Tempo Discreto quando $y(s, a) = 1$

ótima.

Algoritmo 1: Iteração de Políticas

Entrada: Uma política estacionária π e um estado $\bar{s}$ escolhidos arbitrariamente

início

HouveMudança = verdadeiro

enquanto *HouveMudança* **faça**

calcule a solução $[g, V()]$ para o sistema de equações

$V(s) = \bar{r}(s, \pi(s)) - g\ y(s, \pi(s)) + \sum_{s' \in S} p(s, \pi(s), s')V(s')\ ,\ \forall s \in S$

$V(\bar{s}) = 0$

para *cada estado* $s \in S$ **faça**

$$\bar{\pi}(s) = \underset{a \in A}{\operatorname{argmax}} \left\{ \bar{r}(s, a) - g\ y(s, a) + \sum_{s' \in S} p(s, a, s')V(s') \right\}$$

fim

se *política* π *e* $\bar{\pi}$ *são iguais* **então**

HouveMudança = falso

fim

$\pi = \bar{\pi}$

fim

fim

3.4.2 Algoritmo de Iteração de Valores

Nessa seção nós apresentamos um algoritmo para obtenção de políticas ótimas para PMDs que, ao contrário do algoritmo de iteração de políticas, não tem um custo computacional associado à resolução de um sistema de equações a cada iteração. Esse algoritmo, denominado algoritmo de iteração de valores, faz uso de uma solução recursiva advinda do principio da otimalidade de Bellman. De acordo com esse princípio, uma política ótima tem a seguinte propriedade: quaisquer que sejam o estado e decisão iniciais, as decisões restantes devem constituir uma política ótima em relação ao estado resultante da primeira decisão (BELLMAN, 1957). Em outras palavras, pode-se dizer que uma política ótima apresenta subpolíticas ótimas (LEW; MAUCH, 2006).

Levando-se em conta primeiramente o algoritmo para PMDs a Tempo Discreto, define-se, para $n = 1, 2, ...$, a função $V_n(s)$, denominada função de valor do estado s. $V_n(s)$ representa a recompensa máxima esperada nos n períodos até o final do horizonte de planejamento quando o estado atual é s e uma recompensa terminal $V_0(s')$ é recebida quando o sistema atinge o estado terminal s'. O cálculo da função de valor se faz de forma retrógrada no tempo da seguinte forma

$$V_n(s) = \max_{a \in A(s)} \left\{ \bar{r}(s,a) + \sum_{s' \in S} p(s,a,s') V_{n-1}(s') \right\},\ s \in S\ ,$$

de modo que se inicia com um valor arbitrário $V_0(s)$, $\forall s \in S$.

Para cada estado do sistema calcula-se a função $V_n(s)$ e a ação que a maximiza, com o intuito de se obter uma política ótima estacionária. A função de valor fornece limites inferior e superior em relação à diferença $V_n(s) - V_{n-1}(s)$, e o algoritmo termina quando os limites se aproximam da recompensa média ótima levando-se em conta uma tolerância pre-estabelecida ϵ. Segundo Tijms (1995), para um n suficientemente grande, a diferença $V_n(s) - V_{n-1}(s)$ se aproximará da recompensa média máxima por unidade de tempo. Além disso, a política obtida é ótima[5]. O pseudo-código pode ser visto no Algoritmo 2.

Algoritmo 2: Iteração de Valores para PMD a Tempo Discreto

Entrada: $V_0(s)$ com $0 \leq V_0(s) \leq \max_{a \in A(s)} \bar{r}(s,a)$, $\forall s \in S$

início

n = 1

repita

para *cada estado* $s \in S$ **faça**

$V_n(s) = \max_{a \in A(s)} \left\{ \bar{r}(s,a) + \sum_{s' \in S} p(s,a,s') V_{n-1}(s') \right\}$

fim

Seja π_n a política estacionária formada pelas ações resultantes

Calcule os limites:

$m_n = \min_{s \in S} [V_n(s) - V_{n-1}(s)]$

$M_n = \max_{s \in S} [V_n(s) - V_{n-1}(s)]$

$n = n + 1$

até $0 \leq M_n - m_n \leq \epsilon m_n$;

fim

Para o caso contínuo o Algoritmo 2 não se aplica diretamente, já que a recompensa calculada em n instantes de decisão não leva em conta a diferença no tempo entre transições. A solução é aplicar um método de uniformização em que o PMD a Tempo Contínuo é transformado em um PMD a Tempo Discreto equivalente, de modo que, para cada política estacionária, as recompensas médias são as mesmas para ambos

[5] A convergência do algoritmo é provada se ao se fixar uma política a cadeia de Markov resultante seja aperiódica. Caso contrário, o caso da periodicidade pode ser superado por uma perturbação nas probabilidades de transição mostradas em Tijms (1995).

(TIJMS, 1995). O algoritmo de iteração de valores para o modelo original é então implicado pelo algoritmo para o PMD a Tempo Discreto associado.

O cálculo das funções de valor é feito da seguinte forma

$$V_n(s) = \max_{a \in A} \left\{ \bar{r}^t(s,a) + \sum_{s' \in S} p^t(s,a,s') V_{n-1}(s') \right\} \forall s \in S \tag{3.3}$$

em que

$$\bar{r}^t(s,a) = \frac{\bar{r}(s,a)}{y(s,a)} \tag{3.4}$$

$$p^t(s,a,s') = \eta \frac{p(s,a,s')}{y(s,a)}, \quad \text{se} \quad s \neq s' \tag{3.5}$$

$$p^t(s,a,s') = 1 + \eta \frac{p(s,a,s') - 1}{y(s,a)}, \quad \text{se} \quad s = s' \tag{3.6}$$

com

$$0 < \eta \leq \frac{y(s,a)}{1 - p(s,a,s')} \, . \tag{3.7}$$

Fazendo-se as devidas substituições, o pseudo-código para o caso contínuo é mostrado no Algoritmo 3.

3.4.3 Algoritmo de Iteração de Valores Relativo

Embora o algoritmo de iteração de valores convirja para a solução ótima, ele é instável numericamente (PUTERMAN, 2005). Na prática, uma variação denominada algoritmo de iteração de valores relativo é utilizado para encontrar a politica ótima de PMDs. A diferença para o algoritmo de iteração de valores tradicional reside na subtração de uma constante a cada iteração, em que tal constante é a função de valor de um estado escolhido arbitrariamente. O pseudo-código pra o caso discreto

pode ser visto no Algoritmo 4.

Algoritmo 3: Iteração de Valores para PMD a Tempo Contínuo

Entrada: $V_0(s)$ com $0 \leq V_0(s) \leq \max_{a \in A(s)}[r(s,a)/y(s,a)]$, $\forall s \in S$

início

n = 1

repita

para *cada estado* $s \in S$ **faça**

$$V_n(s) = \max_{a \in A(s)} \left\{ \frac{\bar{r}(s,a)}{y(s,a)} + \frac{\eta}{y(s,a)} \sum_{s' \in S} p(s,a,s') V_{n-1}(s') + \left[1 - \frac{\eta}{y(s,a)} \right] V_{n-1}(s') \right\}$$

fim

Seja π_n a política estacionária formada pelas ações resultantes

Calcule os limites:

$m_n = \min_{i \in S}[V_n(s) - V_{n-1}(s)]$

$M_n = \max_{i \in S}[V_n(s) - V_{n-1}(s)]$

$n = n + 1$

até $0 \leq M_n - m_n \leq \epsilon m_n$;

fim

De forma análoga, no caso contínuo as funções de valor, para cada estado, são calculadas após a aplicação da uniformização do PMD conforme a equação 3.8 .

$$V_n(s) = \max_{a \in A(s)} \left\{ \frac{\bar{r}(s,a)}{y(s,a)} - V_{n-1}(\bar{s}) + \frac{\eta}{y(s,a)} \sum_{s' \in S} p(s,a,s') V_{n-1}(s') + \left[1 - \frac{\eta}{y(s,a)} \right] V_{n-1}(s') \right\} \tag{3.8}$$

3.5 Probabilidades Limite

Dado um PMD, a tempo discreto ou contínuo, e fixada uma política de controle markoviana estacionária π, tem-se a cadeia de Markov embutida no processo com uma matriz de probabilidades de transição entre estados. Por essa cadeia de Markov torna-se possível calcular a probabilidade limite de cada estado. Tais probabilidades podem ser interpretadas como a porção de tempo que o sistema permanece em cada estado. A partir delas, medidas de desempenho podem ser obtidas, de acordo com a finalidade do modelo.

A probabilidade que a cadeia de Markov estará no estado s' no tempo t geralmente converge para um valor limite $p_{s'}$, que é independente do estado inicial (ROSS, 2010). Se as seguintes condições forem atendidas: (a) ao se iniciar no estado s existe uma probabilidade positiva de se estar no estado s', para todo $s \in S$ e $s' \in S$, e (b)

começando em qualquer estado o tempo médio de retorno para esse estado é finito; então as probabilidades limite vão existir e satisfazer as equações

$$\begin{cases} p_{s'} = \sum_{s \in S} p_s \, p(s, \pi(s), s') & , \forall s' \in S \\ \sum_{s' \in S} p_{s'} = 1 \end{cases} .$$

Algoritmo 4: Iteração de Valores Relativo para PMD a Tempo Discreto

Entrada: $V_0(s)$ com $0 \leq V_0(s) \leq \max_{a \in A(s)}[\bar{r}(s,a)/y(s,a)]$, $\forall s \in S$ e um estado $\bar{s}$ escolhido arbitrariamente

início

n = 1

repita

para *cada estado* $s \in S$ **faça**

$$V_n(s) = \max_{a \in A(s)} \left\{ \bar{r}(s,a) - V_{n-1}(\bar{s}) + \sum_{s' \in S} p(s,a,s') V_{n-1}(s') \right\}$$

fim

Seja π_n a política estacionária formada pelas ações resultantes

Calcule os limites:

$m_n = \min_{i \in S}[V_n(s) - V_{n-1}(s)]$

$M_n = \max_{i \in S}[V_n(s) - V_{n-1}(s)]$

$n = n + 1$

até $0 \leq M_n - m_n \leq \epsilon m_n$;

fim

4 MODELO ANALÍTICO PARA ALOCAÇÃO DE ESPECTRO EM UM ENLACE

Sob tráfego dinâmico, os problemas relacionados às redes ópticas descritos no capítulo 2 podem ser vistos como problemas de decisão sequencial sob incerteza e modelados por meio de PMDs. Ao longo do tempo, requisições de conexão chegam à rede e deve-se decidir se estas serão aceitas, e, caso sejam, deve-se selecionar quais recursos na rede serão utilizados para acomodá-las. Ao mesmo tempo, conexões sendo transmitidas podem ser finalizadas e os seus recursos passam a ficar disponíveis.

Alguns trabalhos na literatura modelam o problema RWA por meio de PMDs. Dentre eles podemos citar Hyytia e Virtamo (2000), que propõem uma abordagem na qual se adota uma política de controle inicial criada por uma heurística, como, por exemplo, a *First-Fit*, em conjunto com o roteamento pelo caminho mínimo (*shortest-path*). Definida essa política inicial, realiza-se apenas uma iteração do algoritmo de iteração de políticas, com o objetivo de se obter um política mais eficiente que a anterior.

O subproblema de atribuição de comprimento de onda é investigado no trabalho de Mosharaf et al. (2003) para uma rede WDM composta por três nós e duas classes de tráfego. Os autores utilizam um PMD com o objetivo de encontrar uma política ótima de alocação que maximize a soma ponderada das chamadas utilizadas por cada classe. Mosharaf et al. (2005) também modelam, por meio de PMDs, o problema de atribuição de comprimentos de onda em uma rede WDM com três nós. Os autores supõem três classes de tráfego distintas com suporte ao controle justo, que visa uma distribuição dos recursos da rede de maneira igualitária entre as chamadas de diferentes classes. Baseando-se em Mosharaf et al. (2003), Tachibana et al. (2007) propõem um método de estabelecimento dinâmico de caminhos ópticos em uma rede WDM com conversores de comprimento de onda em todos os nós. Busca-se uma política ótima, obtida a partir de um PMD, que diminua a probabilidade de fracasso ao se tentar estabelecer um caminho óptico.

De acordo com as nossas pesquisas, o problema de alocação de espectro em redes ópticas elásticas ainda não foi investigado por meio de PMDs. Nosso objetivo, então, é encontrar uma política de alocação de espectro não míope modelando um enlace de uma rede óptica elástica como um PMD, a exemplo do que já foi feito para redes WDM. Mais especificamente, apresentamos neste capítulo um PMD a Tempo Contínuo[1] sob horizonte de planejamento infinito levando em conta o critério de

[1]No apêndice A nós apresentamos detalhes de implementação desse PMD, com algumas representações alternativas para alcançar uma maior eficiência computacional.

recompensa média.

Uma vez que uma política de alocação de espectro é escolhida e implementada, deve-se avaliar seu desempenho de acordo com alguma métrica, como, por exemplo, a probabilidade de bloqueio na rede. Com esse intuito, alguns modelos de cálculo de desempenho foram propostos na literatura. Duas cadeias de Markov a tempo contínuo foram propostas por Yu et al. (2013), para fornecer uma análise de desempenho de um único enlace de uma rede óptica elástica sob tráfego dinâmico. Os autores consideram duas políticas de alocação de espectro míopes: *Random-Fit* e *First-Fit*. O modelo proposto considera múltiplas classes de chamadas, em que cada uma requer um número diferente de *slots* contíguos. Além disso, os resultados do modelo analítico são comparados com simulações Monte Carlo, utilizando medidas de desempenho como probabilidade de bloqueio e taxa de utilização de recursos.

No trabalho de Christodoulopoulos et al. (2013), modelos analíticos e aproximados são desenvolvidos para computar a probabilidade de bloqueio de cada conexão e da rede inteira baseado em uma cadeia de Markov. Além disso, um algoritmo de roteamento e alocação de espectro que utiliza esses modelos de probabilidade de bloqueio é proposto para minimizar a probabilidade de bloqueio média da rede. Os autores mostram que os resultados analíticos obtidos estão em concordância com as simulações correspondentes.

Neste capítulo nós descrevemos medidas de desempenho calculadas a partir das probabilidades limite, mostradas na seção 3.5. Tais medidas permitem avaliar as políticas encontradas e compará-las com outras políticas. Nosso modelo permite que cada política seja implementada ao se enumerar todas as configurações de espectro possíveis e aplicar, para cada uma delas, a regra da política de acordo com cada requisição de conexão. Portanto, uma cadeia de Markov pode ser derivada e utilizada para calcular as medidas de desempenho.

4.1 Modelo Markoviano de Decisão a Tempo Contínuo

Considere um enlace de uma rede óptica elástica sob tráfego dinâmico em que o espectro óptico é dividido em $N \in \mathbb{N}^*$ *slots*. Com o objetivo de representar padrões de tráfego misto na rede, define-se $K \in \mathbb{N}^*$ tipos diferentes de requisições de conexão, em que cada tipo k, $1 \leq k \leq K$, é composto por: w_k, o número de *slots* contíguos requeridos para transmiti-la; sua taxa de chegada, λ_k, de acordo com uma distribuição de *Poisson*; e $1/\mu_k$ como o tempo médio de transmissão de uma conexão no enlace de acordo com uma distribuição exponencial. Assume-se que os tipos de

conexão estão ordenados em ordem ascendente de acordo com o número de *slots* que eles demandam, ou seja, $w_k \leq w_{k+1}$ para $1 \leq k \leq K-1$. Além disso, $gb \in \mathbb{N}$ *slots* são usados como banda de guarda.

Quando há uma requisição do tipo k para ser transmitida pelo enlace, deve-se decidir quais *slots* serão alocados para transmiti-la. Caso não existam *slots* suficientes para alocá-la, a requisição é bloqueada. É importante notar que uma requisição de conexão é bloqueada apenas se não há *slots* contíguos disponíveis para transmiti-la e, portanto, rejeição não é uma ação válida no nosso modelo.

A escolha de qual conjunto de *slots* contíguos será utilizado para alocar cada conexão reflete diretamente em quão eficiente o espectro óptico será utilizado ao longo do tempo. Dada a aleatoriedade do sistema, torna-se importante levar em conta as consequências a longo prazo de cada decisão. Com esse intuito nós modelamos o sistema como um PMD a Tempo Contínuo, cuja solução é uma política de alocação de espectro ótima a partir da qual as decisões são tomadas.

Cada estado é definido como uma dupla $s = (c, ev)$. A tupla $c = (c_1, c_2, c_3, ..., c_N)$ descreve a configuração do espectro na qual cada elemento c_i denota se o *slot* no índice i está disponível, $c_i = 0$, se ele está sendo usado para alocar uma conexão do tipo k ($c_i = k$), ou como banda de guarda ($c_i = -1$). O segundo elemento, um evento do sistema ev, compreende um par tipo-valor (ev_t, ev_v), em que ev_t indica se o evento atual é uma requisição de conexão (IN) ou um término de transmissão (OUT). Dado $ev_t = IN$, ev_v é o tipo de requisição que chegou; se $ev_t = OUT$, ev_v representa a j-ésima chamada em progresso, numerada da esquerda para a direita, cuja transmissão foi finalizada e será removida do espectro. Assim, o conjunto de eventos E é dado por

$$
\begin{aligned}
E = \{\ (ev_t, ev_v) \mid\ & ev_t \in \{IN, OUT\}, \\
& \text{se } ev_t = IN \text{ então } ev_v \in \{1, ..., K\}, \\
& \text{se } ev_t = OUT \text{ então } ev_v \in \{1, ..., M\}\ \},
\end{aligned}
$$

em que M é o número máximo de conexões concorrentes, dado pelo maior número de chamadas do tipo 1 concorrentes que o enlace pode transmitir considerando também as bandas de guarda necessárias. Desse modo, M é o maior inteiro para o qual a desigualdade $M(w_1 + gb) \leq N + 1$ se mantém.

Seja $\phi_s(p)$ uma função que retorna o número de *slots* contíguos disponíveis a partir do índice p (incluso) em diante no estado s. Define-se o conjunto de todas as posições

possíveis em que uma requisição do tipo k pode ser alocada no estado s como

$$\psi_s(k) = \{p \in \{1, 2, ..., N\} |\ \phi_s(p) \geq w_k\}\ .$$

Considerando C como o conjunto de todas as tuplas denotando configurações de espectro que são válidas, o espaço de estados S é definido como

$$S = \{\ (c, ev)\ |\ c \in C, ev \in E,$$
$$\text{se } ev_t = OUT\ \wedge\ ev_v = j \text{ então } 1 \leq j \leq n \qquad \},$$

em que n é o número de conexões sendo transmitidas no momento no enlace. Deve existir pelo menos $j \in \mathbb{N}^*$ chamada no sistema para um evento $(ev_t = OUT, ev_v = j)$ ocorrer. Como configurações válidas do espectro entende-se todas aquelas que são compatíveis com os dados de entrada do problema, como os tipos de conexão e o número de *slots* usado como banda de guarda,

O sistema, por hipótese, é observado de modo contínuo ao longo do tempo e uma decisão deve ser tomada imediatamente após a ocorrência de um evento. Desse modo, dado um estado $s = (c, ev)$, o qual denominamos estado pré-decisão, uma ação a é escolhida entre o conjunto de ações factíveis para esse estado. Se há uma chegada de requisição de conexão, $(ev_t = IN, ev_v = k)$, as ações factíveis são: se há ao menos uma posição disponível em que a conexão possa ser alocada, ela é aceita e alocada em uma das posições do conjunto $\psi_s(k)$; caso contrário é bloqueada. Quando há um término de transmissão de uma conexão, a única ação disponível é retirá-la da tupla c, tornando seus *slots* disponíveis.

Depois que uma ação é implementada, o sistema assume um estado pós-decisão $\bar{s} = (c')$ em que permanece até a chegada de um novo evento. O estado pós-decisão é denotado pelo vetor $c' \in C$ que representa a configuração do espectro de acordo com a consequência imediata da ação escolhida a. Se uma conexão do tipo k é aceita na posição p, w_k *slots* contíguos a partir de p são alocados para transmiti-la. Se uma transmissão de conexão é finalizada, seus *slots* e bandas de guarda relacionados são modificados para disponíveis. Por fim, se ela é bloqueada a carga permanece inalterada.

Em um estado pós-decisão $\bar{s}$ a ocorrência de um novo evento é estocástico e determina a evolução do sistema para o próximo estado pré-decisão. Esse novo evento é a chegada de uma nova requisição de conexão seguindo k distribuições de Poisson

independentes, ou o término de uma conexão sendo transmitida de acordo com n distribuições exponenciais independentes. Quando uma requisição do tipo k chega e há *slots* contíguos disponíveis suficientes para acomodá-la, a taxa de transição para o próximo estado $s' = (c', (IN, k))$ é λ_k. Caso contrário, se há um término de transmissão, a taxa de transição para $s' = (c', (OUT, j))$ é μ_{m_j}. m_j é o tipo da j-ésima chamada, da esquerda para a direita, sendo transmitida no enlace, com $1 \leq m_j \leq K$.

Seja $\beta(s, a)$ a taxa total de saída do estado pós-decisão resultante ao se tomar a ação a no estado s, então

$$\beta(s, a) = \sum_{k=1}^{K} \lambda_k + \sum_{j=1}^{n} \mu_{m_j} \ .$$

Desta maneira, as probabilidades de transição de um par estado-ação para o próximo estado pré-decisão s', $p(s, a, s')$ são dadas por

$$p(s, a, s') = \begin{cases} \frac{\lambda_k}{\beta(s,a)} \ , & \forall k \in K \\ \frac{\mu_{m_j}}{\beta(s,a)} \ , & \forall j, \ 1 \leq j \leq n \end{cases} \quad . \tag{4.1}$$

Deve-se notar que a transição para o estado pós-decisão $\bar{s}$ é determinística e depende apenas do estado anterior s e da ação a. Já a transição do estado pós-decisão para o próximo estado pré-decisão s' é estocástica e depende nas taxas de chegada e término de transmissão que são as entradas do modelo. Essa decomposição entre transições determinísticas e estocásticas no modelo tem um importante papel no desenvolvimento e implementação dos algoritmos de aprendizagem por reforço, como o apresentado no capítulo 5.

A dinâmica do sistema é ilustrada na Figura 4.1, em que t_i representa o tempo do i-ésimo instante de decisão e s_i e $\bar{s}_i$ representam, respectivamente, os estados pré-decisão e pós-decisão.

Por fim, a recompensa esperada ao se escolher a ação a no estado pré-decisão s, $r(s, a)$, é proporcional à quantidade de *slots* utilizada para acomodar todas as conexões transmitidas no estado pós-decisão $\bar{s}$. Esse valor é ponderado pelo tempo esperado até a próxima chegada de um novo evento, que é $1/\beta(s, a)$. Portanto,

$$r(s, a) = \frac{\sum_{j=1}^{n} w_{m_j}}{\beta(s, a)} \ ,$$

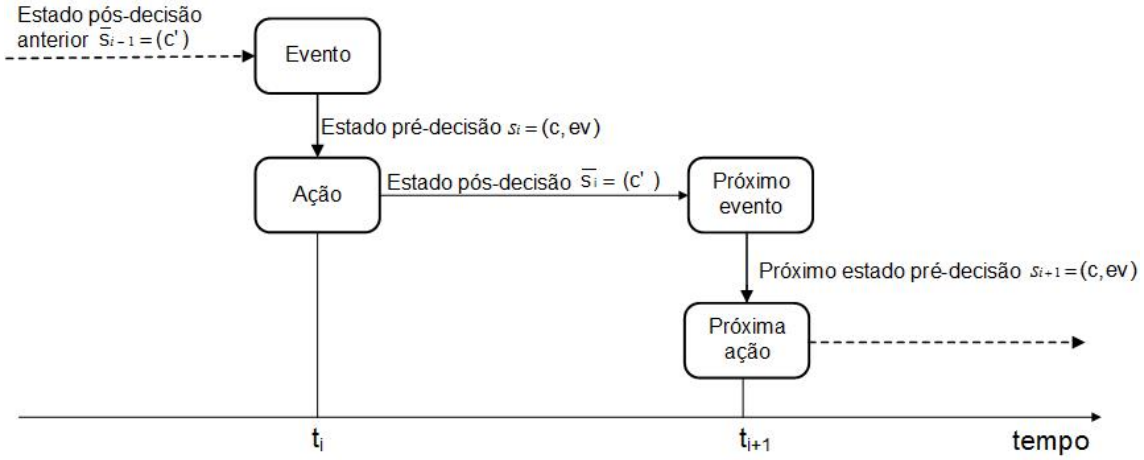

Figura 4.1 - Dinâmica do sistema: após a ocorrência de um evento, uma ação é escolhida entre as ações factíveis e o estado pré-decisão evolui para um estado pós-decisão, no qual o sistema permanece até o próximo instante de decisão. Uma outra ação é escolhida e o processo se repete.

em que w_{m_j} é a quantidade de *slots* alocados para a j-ésima chamada.

4.2 Medidas de Desempenho

A partir do cálculo das probabilidades limite apresentadas na seção 3.5, as seguintes medidas de desempenho podem ser obtidas.

Seja n_{sk} o número de conexões do tipo k no estado s. Calcula-se a taxa de conexões transmitidas com sucesso da classe k baseado na sua taxa de transmissão μ_k e a probabilidade limite p_s. Tal taxa, conhecida também como *throughput* τ_k, é dada por

$$\tau_k = \sum_{s \in S} n_{sk}\, \mu_k\, p_s, 1 \leq k \leq K \ .$$

A probabilidade de bloqueio das conexões do tipo k, PB_k, pode então ser definida como o complemento da razão entre as conexões transmitidas com sucesso por unidade de tempo sobre a quantidade média de requisições de conexão por unidade de tempo. Assim

$$PB_k = 1 - \frac{\tau_k}{\lambda_k}, 1 \leq k \leq K \ .$$

Pode-se calcular também a probabilidade de bloqueio do enlace PB, considerando todas as classes de requisição de conexão, como

$$PB = 1 - \frac{\sum_{k=1}^{K} \tau_k}{\sum_{k=1}^{K} \lambda_k} .$$

Como nas redes ópticas elásticas são empregados cenários nos quais as demandas de conexão podem variar de um a vários *slots*, usar apenas a probabilidade de bloqueio do enlace como medida de desempenho pode não ser apropriado. Nós utilizamos então uma medida de desempenho denominada probabilidade de bloqueio de *slots* no enlace (PBS). Essa é a probabilidade que um *slot* requisitado, pertencente a qualquer conjunto de *slots* relacionados a qualquer requisição, seja bloqueado. Nós calculamos a probabilidade de bloqueio de *slots* como o complemento da razão entre a quantidade de *slots* alocados para transmitir conexões em um enlace por unidade de tempo sobre a quantidade média de *slots* requeridos por unidade de tempo.

Seja $w_k \tau_k$ a taxa de *slots* alocados com sucesso para a classe k, então a probabilidade de bloqueio de *slots* é dada por

$$\text{PBS} = 1 - \frac{\sum_{k=1}^{K} w_k \, \tau_k}{\sum_{k=1}^{K} \lambda_k w_k} \quad .$$

5 ALGORITMO DE APRENDIZAGEM POR REFORÇO PARA ALOCAÇÃO DE ESPECTRO EM UM ENLACE

O algoritmo de iteração de valores utilizado para resolver um PMD a Tempo Contínuo, mostrado no capítulo 3, calcula recursivamente uma sequência de funções de valor aproximando a recompensa média máxima por unidade de tempo. Tal cálculo é feito baseando-se nas equações de otimalidade, a exemplo das equações 3.8, ao se iniciar o algoritmo com uma função de valor escolhida arbitrariamente $V_0(s), s \in S$. Tais métodos possuem uma suposição implícita de que as funções de valor podem ser calculadas e armazenadas na memória, ou seja, pode-se mapear todos os estados do sistema (assumindo que este número seja finito) às suas respectivas funções de valor. Tais métodos são denominados exatos pois permitem que seja feita uma computação que convirja para uma solução ótima do problema considerado (SIGAUD; BUFFET, 2010).

Nota-se nos trabalhos que resolvem o problema RWA por meio de PMDs citados no capítulo 3 - a exemplo de Hyytia e Virtamo (2000), Mosharaf et al. (2003) e Mosharaf et al. (2005) - que os autores conseguem encontrar soluções ótimas apenas para modelos de redes de pequeno porte, ou seja, com número reduzido de nós e comprimentos de onda. Além disso, geralmente apenas um dos subproblemas é tratado, o roteamento ou a atribuição dos comprimentos de onda. Mesmo para pequenas instâncias do problema, esses modelos apresentam um grande espaço de estados, a exemplo do que também ocorre ao se trabalhar com o problema RSA em redes ópticas elásticas.

Tomando como base o modelo descrito no capítulo 4, para criar o espaço de estados deve-se enumerar todas as tuplas, o conjunto C, de configurações de espectro possíveis e os eventos relacionados com cada uma dessas configurações. Desse modo, o tamanho do espaço de estados cresce rapidamente de acordo com as dimensões do problema, especialmente quando o tamanho da tupla c cresce. O tamanho da tupla varia em decorrência da quantidade de *slots* na grade; do número de *slos* demandado por cada tipo de conexão; e do tamanho da banda de guarda. Esse crescimento do espaço de estados é conhecido como a primeira maldição da dimensionalidade (*first curse of dimensionality*) (POWELL, 2011) e é a principal razão pela qual é impraticável computacionalmente calcular políticas ótimas para instâncias mais realistas do problema de alocação de espectro. Em algumas instâncias do problema podemos encontrar espectros com até centenas de *slots*.

Além disso, vale notar que neste trabalho estamos lidando com apenas um enlace

da rede e, portanto, torna-se importante encontrar uma boa política de alocação de espectro em um tempo de computação eficiente, para que o modelo possa ser estendido para redes com topologias arbitrárias. As redes elásticas com outras topologias podem ser formadas por dezenas de nós e enlaces, e cada fibra óptica pode conter até centenas de *slots*, tornando computacionalmente inviável encontrar soluções ótimas para os PMDs. Torna-se necessário, então, utilizar técnicas que permitam encontrar boas políticas de maneira aproximada, sem a necessidade de se calcular a função de valor de maneira exata para todos os estados. Para tanto, avança-se no tempo, diferentemente da programação dinâmica, utilizando algoritmos iterativos para estimar a função de valor.

Neste trabalho utilizamos um algoritmo de aprendizagem por reforço, no qual, ao invés de se calcular as funções de valor para todos os estados de maneira exata, avança-se no tempo utilizando-se amostras aleatórias de informações exógenas ao sistema (por meio de uma simulação) e se estima iterativamente as funções de valor para os estados visitados. Evita-se, portanto, a necessidade de calcular e guardar as matrizes de probabilidade de transição e recompensas esperadas desde que se tenha um simulador de eventos do sistema.

Embora o cálculo sobre todos os estados do espaço de estados seja evitado ao se utilizar a aprendizagem por reforço, deve-se guardar as funções de valor para cada estado que possa ser visitado pelo algoritmo, um número que ainda pode ser muito grande. Por isso, nós adaptamos um esquema de aproximação de função no qual uma grande quantidade de funções de valor são guardadas na forma de uma quantidade pequena de escalares.

Mais especificamente, nós propomos neste capítulo a utilização de um algoritmo de aprendizagem por reforço denominado *Relaxed Semi-Markov Average Reward Technique* (Relaxed-SMART), inicialmente proposto por Gosavi (2004) com uma modificação apresentada em Gosavi (2011). Em conjunto com esse algoritmo, empregamos um esquema de aproximação de função de tal modo que nós possamos lidar com instâncias realistas do problema de alocação de espectro. A escolha do algoritmo *Relaxed-SMART* foi motivada dado o seu foco no critério de otimalidade de recompensa média a longo prazo para problemas modelados a tempo contínuo. Esse algoritmo é utilizado também em outras áreas de aplicação como redes sem fio de celular (YU et al., 2008) e controladores de processo (GANESAN et al., 2007).

O problema das maldições de dimensionalidade aparecem também na quantidade de estados e transições da cadeia de Markov embutida no PMD. Torna-se impra-

ticável calcular as medidas de desempenho analiticamente para alguns casos. Nós utilizamos, então, simulações para computar as medidas de desempenho tanto das políticas do algoritmo de aprendizagem por reforço quanto das políticas míopes.

5.1 Algoritmo *Relaxed-SMART*

Ao contrário das equações de otimalidade para os PMDs apresentadas anteriormente, nas quais se utiliza uma função de valor $V(s)$ para cada estado s, considere uma função de valor $Q(s,a)$ para cada par estado-ação do sistema. Alguns algoritmos examinam os valores de $Q(s,a)$ para as ações factíveis relacionadas a cada estado, e a ação que obtiver o maior valor constitui a ação ótima para esse estado. Dessa maneira, a função de valor ótima $Q^*(s,a)$ para um par estado-ação (s,a) relacionada ao critério de otimalidade de recompensa média é definida como

$$Q^*(s,a) = r(s,a) - g^* y(s,a) + \sum_{s' \in S} p(s,a,s') V^*(s') \ , \tag{5.1}$$

e, portanto, $V^*(s) = \max_{a \in A(s)} Q(s,a)$. Nós podemos, então, reescrever a equação 5.1 como

$$Q^*(s,a) = r(s,a) - g^* y(s,a) + \sum_{s' \in S} p(s,a,s') \max_{a \in A(s')} Q(s',a) \ . \tag{5.2}$$

A ideia dos algoritmos de aprendizagem por reforço, como o *Q-Learning* apresentado originalmente para o critério de recompensa descontada em Watkins (1989), é atualizar os valores $Q(s,a)$ iterativamente por meio de simulação dado um valor inicial para cada par estado-ação. A cada estado visitado, uma ação é escolhida e o sistema evolui para um novo estado; tendo como retorno uma recompensa esperada e um tempo de transição que são guardados como um *feedback* do sistema em relação ao agente. Esse retorno é utilizado para atualizar a função de valor para a ação selecionada no estado anterior. O processo é repetido para um grande número de transições de modo que, ao final do algoritmo, as ações que geram os melhores valores para cada estado sejam encontradas e selecionadas como a política de controle.

Com o intuito de evitar que as transições de probabilidade precisem ser calculadas e armazenadas, uma versão da equação de atualização que utiliza uma taxa de aprendizagem é comumente utilizada na prática. Tal algoritmo pode inferir funções de valor $Q(s,a)$ para todos os pares estado-ação a partir das amostras geradas por

um simulador do sistema. A taxa de aprendizagem α, com $\alpha \in (0, 1]$, representa o quanto uma nova experiência do agente influencia a função de valor já calculada. Idealmente, no início do aprendizado, essa taxa dese ser maior e, após algumas amostras da simulação, o peso da experiência pode ficar maior do que o aprendizado. A atualização dos valores $Q(s, a)$ é efetuada então da seguinte forma

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha[r(s,a,s') - g^* t(s,a,s') + \max_{a \in A(s')} Q(s',a)] \,, \qquad (5.3)$$

em que $r(s, a, s')$ é a recompensa recebia ao se escolher a ação a no estado s e evoluir para o estado s', e $t(s, a, s')$ é o tempo entre os instantes de decisão. Essa transformação é análoga ao esquema de aproximação estocástica de Robbins-Monro (ROBBINS; MONRO, 1951), que torna possível substituir a esperança sobre s' por uma amostra gerada via simulação.

O algoritmo *Q-Learning* baseado na equação 5.3 é análogo ao algoritmo de iteração de valores pra o critério de recompensa média, que é numericamente instável (PUTERMAN, 2005). Dessa forma, nós utilizamos o algoritmo proposto por Gosavi (2004), com uma modificação apresentada em (GOSAVI, 2011), que tem suas raízes no algoritmo de iteração de valores relativo. Nesse algoritmo, denominado *Relaxed-SMART*, a recompensa média e as funções de valor estado-ação são aproximados utilizando-se duas taxas de aprendizado em duas escalas de tempo.

No *Relaxed-SMART*, na $(k + 1)$-ésima iteração, a função de valor é atualizada de acordo com a seguinte regra

$$Q^{k+1}(s,a) = (1-\alpha^k)Q^k(s,a) + \alpha^k \left[r(s,a,s') - g^k t(s,a,s') + \eta \max_{a \in A(s')} Q^k(s',a) \right] \,, \qquad (5.4)$$

em que α^k e g^k denotam, respectivamente, a de taxa de aprendizagem principal e a recompensa média na k-ésima iteração do algoritmo, e η é um escalar positivo, $0 < \eta < 1$, cujos valores devem ser próximos de um. Para valores de η suficientemente próximos de 1, como por exemplo $\eta = 0,99$, o algoritmo tem garantia de convergência para a solução ótima (GOSAVI, 2011). Além disso, g^k é estimado ao se tomar a razão da recompensa total recebida e o tempo total de simulação até a k-ésima iteração, da seguinte forma

$$g^{k+1} = (1 - \beta^k)g^k + \beta^k \frac{R(k) + r(s, a, s')}{T(k + 1)} \tag{5.5}$$

em que $R(k)$ e $T(k)$ são respectivamente a soma das recompensas obtidas e o tempo total de simulação até a a k-ésima iteração. β é a taxa de aprendizagem secundária, escolhida de acordo com uma regra de decaimento que faz com que ela tenda a 0 mais rapidamente do que α. O Algoritmo 5 mostra o pseudo-código da fase de treinamento do *Relaxed-SMART*.

5.1.1 Conflito Exploração-Intensificação

Um conflito clássico em várias técnicas, dentre elas a aprendizagem por reforço, é encontrar um equilíbrio entre exploração e intensificação (*exploration-exploitation*). Explorar implica em visitar estados apenas para conhecer os valores $Q(s, a)$ associados, independentemente se é uma boa decisão ou não ir para aquele estado. Já na intensificação, tomam-se decisões que parecem ser as melhores baseadas em informações atuais, ou seja, estados cujas funções de valor são melhores são priorizados. Para obter uma boa recompensa, um agente deve preferir ações que ele experimentou no passado e percebeu que são eficientes; porém, para descobrir tais ações, ele deve experimentar ações que ele não selecionou anteriormente. O dilema é que o agente deve tentar várias ações e progressivamente favorecer as que parecem ser as melhores.

Assumindo que os valores iniciais das funções de valor para os pares estado-ação são menores dos que se espera ao término do algoritmo, a cada vez que um estado é visitado sua função de valor tende a crescer, produzindo uma situação em que seja mais favorável revisitar este estado em detrimento dos estados que continuam com valores iniciais. Rapidamente, tal processo de intensificação pode formar um ciclo entre um pequeno número de estados visitados. Caso os valores iniciais sejam muito altos, o algoritmo tende a apenas explorar os estados, sem intensificá-los. Duas das estratégias comumente empregadas para o conflito exploração-intensificação são apresentadas nas próximas seções[1].

5.1.1.1 Exploração ϵ-gulosa

Nessa estratégia, a escolha de uma ação exploratória na k-ésima iteração depende de uma probabilidade $p(k)$, enquanto que, com uma probabilidade $1 - p(k)$, a ação

[1] Outras estratégias para o conflito exploração-intensificação são discutidas em Powell (2007).

gulosa é selecionada, ou seja, a ação cuja função de valor $Q(s,a)$ tem o maior valor.

Algoritmo 5: *Relaxed-SMART*

Entrada: Valores escolhidos arbitrariamente: $Q(s,a)$, para todos os pares estado-ação; g^1; α^1; β^1; η; NumMaxIter.

início

Seja um estado inicial s

$k = 1$

$R(k) = 0$

$T(k) = 0$

enquanto $k <= NumMaxIter$ **faça**

De acordo com um esquema de exploração-intensificação (seção 5.1.1) escolha uma ação a

Aplique a ação a

Simule um evento e obtenha o próximo estado s'

Atualize a função $Q(s,a)$ da seguinte maneira

$$Q^{k+1}(s,a) = (1-\alpha^k)Q^k(s,a)+\alpha^k\left[r(s,a,s')-g^k t(s,a,s')+\eta \max_{a\in A(s')} Q^k(s',a)\right],$$

se *uma ação gulosa foi escolhida* **então**

$R(k) = R(k) + r(s,a,s')$

$T(k) = T(k) + t(s,a,s')$

$g^{k+1} = (1-\beta^k)g^k + \beta^k \frac{R(k)+r(s,a,s')}{T(k+1)}$

fim

$k = k + 1$

$s \leftarrow s'$

fim

fim

A ideia é manter um certo grau de exploração que decai no decorrer das iterações. Uma maneira de realizar esse decaimento é fazer

$$p(k) = \frac{G1}{G2 + k}$$

em que $G1$ e $G2$ são dois números positivos cujos valores podem ser, por exemplo, 1000 e 2000.

5.1.1.2 Exploração *Boltzmann*

Manter uma exploração em problemas muito grandes, com muitos estados e ações factíveis para cada um deles, pode não ser suficiente para encontrar boas políticas. Além disso, ao se explorar em uma iteração do algoritmo, a ação escolhida pode ter um retorno muito baixo que não ajuda no aprendizado. Uma alternativa para esse caso é utilizar a exploração de *Boltzmann* em que a partir do estado s, uma ação é escolhida com uma probabilidade proporcional à função de valor estimada para esse par estado-ação, $p(s, a)$, dada por

$$p(s,a) = \frac{e^{Q(s,a)/T}}{\sum\limits_{a' \in A} e^{Q(s,a')/T}} \ ,$$

em que T é um parâmetro de temperatura. À medida que o parâmetro T aumenta, a probabilidade de se escolher ações diferentes se torna mais uniforme. Ao passo que T se aproxima de 0, a probabilidade de se escolher a ação gulosa se aproxima de 1. Normalmente, inicia-se T com um valor relativamente grande e tal valor é decaído a cada iteração do algoritmo.

5.1.2 Escolha da taxa de aprendizagem

O algoritmo *Relaxed-SMART* utiliza duas taxas de aprendizado: α, utilizada na atualização das funções de valor para os pares estado-ação; e β, utilizado para atualizar a recompensa média g. Nota-se que β deve convergir para 0 mais rapidamente que α, então, idealmente, quando a iteração k crescer, β^k/α^k deve tender a 0. O decaimento das taxas de aprendizagem podem ser em função da iteração k, a exemplo da regra de decaimento $\alpha^k = \frac{1}{k}$, comumente utilizada na literatura. Outras regras de decaimento das taxas de aprendizado são apresentadas nas próximas seções.

5.1.2.1 Regra do Log

O decaimento ocorre de acordo com

$$\alpha^k = \frac{\log(k)}{k} \ ,$$

em que a atualização de α deve ser feita a partir da segunda iteração. Caso contrário, se o valor inicial de k for 1, a atualização é dada por

$$\alpha^k = \frac{\log(k+1)}{k+1} \ .$$

5.1.2.2 Regra Harmônica Generalizada

Essa regra, uma generalização da regra $1/k$, é dada por

$$\alpha^k = \frac{c}{c+k-1} \ .$$

Ao se aumentar c, a velocidade com que a taxa de aprendizagem se aproxima de 0 a cada iteração é diminuída. Encontrar o melhor valor de c requer entendimento da taxa de convergência da aplicação, então torna-se mais um parâmetro a ser ajustado.

5.1.2.3 Regra Polinomial

O decaimento é dado por

$$\alpha^k = \frac{1}{k^\delta} \ .\delta \in (1/2, 1] \tag{5.6}$$

no qual valores menores de δ diminuem a velocidade de decaimento. Tal regra torna-se interessante quando um sistema pode apresentar comportamento instável nas transições iniciais. O melhor valor de δ é um parâmetro a ser calibrado.

5.1.3 Simulação de eventos

A aplicação da ação a envolve dois passos: alcançar o estado pós-decisão após a aplicação da ação; simular um novo evento de acordo com a dinâmica do sistema para encontrar o novo estado. Vale ressaltar que o primeiro passo é determinístico enquanto o segundo é estocástico. O estado pós-decisão é determinado de acordo com o que foi apresentado na seção 4.1, enquanto que a simulação de eventos é feita de acordo com as probabilidades apresentadas nas equações 4.1.

5.1.4 Recompensa Esperada e Tempo de Transição

Dada uma transição do estado s para o estado s' quando a ação a é escolhida, a recompensa esperada é dada por

$$r(s, a, s') = t(s, a, s') \sum_{j=1}^{n} w_{m_j} ,$$

em que w_{m_j} é a quantidade de *slos* alocados para a j-ésima conexão transmitida pelo enlace. O tempo de transição $t(s, a, s')$ é obtido por uma distribuição exponencial com parâmetro

$$\sum_{k=1}^{K} \lambda_k + \sum_{j=1}^{n} \mu_{m_j}$$

para um sistema com K tipos de requisições de conexão.

5.2 Algoritmo *Relaxed-SMART* com Aproximação de Função

Embora o algoritmo *Relaxed-SMART* evite a visita de todos os estados do espaço de estados a cada iteração, alguns novos desafios são introduzidos:

- Deve-se armazenar uma função de valor para cada par estado-ação visitado pelo algoritmo, uma quantidade que ainda pode ser muito grande e computacionalmente inviável;
- Apenas as funções de valor dos estados visitados pelo sistema são atualizadas. Assumindo que as aproximações iniciais tenham valor 0, deve-se encontrar uma maneira de estimar valores $Q(s, a)$ que não foram atualizados pelo algoritmo.

Nos algoritmos apresentados até aqui, as funções de valor são representadas por uma tabela de pesquisa, sejam elas funções de valor de estado nos PMDs ou para cada par estado-ação na aprendizagem por reforço. Nessa representação, assume-se que todos os valores $V(s)$ e $Q(s, a)$ podem ser visitados e calculados (ou estimados), uma suposição inviável para problemas de grande porte. Torna-se necessário, então, encontrar uma aproximação desses valores por meio de uma quantidade menor de parâmetros. Uma alternativa é procurar identificar características (*features*) das variáveis de um estado (ou par estado-ação), e construir uma aproximação baseando-se

nelas. Essa técnica exige que se utilizem as propriedades de cada atributo do estado, por meio de funções $\phi_f(s)$, $f \in F$, em que f é uma característica.

Neste trabalho nós fazemos a aproximação da função por meio de um modelo de regressão linear. A partir das características extraídas de um par estado-ação, pode-se aproximar o valor da função $Q(s, a)$ da seguinte forma

$$\bar{Q}(s, a|\theta) = \sum_{f \in F} \theta_f \phi_f(s) \ . \tag{5.7}$$

Em um modelo de regressão toma-se como dados de entrada uma série de observações de características e observações de valores relacionados a tais características. Como saída, o modelo retorna um conjunto de parâmetros θ^n que generalizam as n observações (DRAPER; SMITH, 1998). Dessa forma, ao invés de se armazenar um parâmetro para cada par estado-ação, apenas um vetor de parâmetros θ estimado por um método de regressão linear precisa ser armazenado. As ações gulosas são calculadas por

$$a = \underset{a \in A(s)}{\operatorname{argmax}} \ \bar{Q}(s, a|\theta) \ , \tag{5.8}$$

a medida que o algoritmo é executado.

Diversas técnicas derivadas da estatística e aprendizado de máquina podem ser utilizadas para aproximar funções. Hastie et al. (2003) dissertam sobre vários métodos de aprendizado estatístico, enquanto Bertsekas e Tsitsiklis (1996) discutem a utilização de redes neurais para esse fim. Ao contrário dos casos tradicionais em que se tem as observações e seus resultados antecipadamente, na aprendizagem por reforço esses algoritmos devem ser implementados iterativamente. Neste trabalho, especificamente, nós utilizamos um modelo de regressão linear iterativo baseado no conteúdo do capítulo 9 de Powell (2011). As equações para atualização do vetor de pesos θ são descritas no Apêndice B.

5.2.1 Características Extraídas

No contexto das redes ópticas elásticas sob tráfego dinâmico, os *slots* disponíveis podem ficar divididos em pequenos fragmentos em cada enlace no decorrer do tempo. Alguns estudos foram conduzidos na literatura para mensurar essa fragmentação e o

seu impacto no bloqueio de requisições de conexão. a exemplo de Wang et al. (2012), Zhang et al. (2012) e Shi et al. (2013). No desenvolvimento desde trabalho notamos que as medidas de fragmentação do espectro podem ser utilizadas também para uma descrição sucinta da configuração do espectro. Nós utilizamos como uma das características, então, a adaptação de uma medida de fragmentação proposta inicialmente para análises comparativas na ciência política e fragmentação de partidos políticos (LAAKSU; TAAGEPERA, 1979) [2].

Após experimentação com algumas características, as seguintes características foram extraídas dos pares estado-ação:

- caso o evento seja uma chegada tem-se o valor 1, caso contrário 0;
- a quantidade de conexões no estado pós-decisão;
- o número de *slots* ocupados no estado pós-decisão;
- a medida de fragmentação do estado pós-decisão dada por

$$\frac{d^2}{\sum_{j=1}^{n} d_j^2} \tag{5.9}$$

  em que d é a soma de todos os *slots* disponíveis no espectro e d_j a quantidade de *slots* contíguos disponíveis em um dado fragmento do espectro, assumindo-se que há n fragmentos.

5.2.2 Pseudo-Código

O pseudo-código do *Relaxed-SMART* com regressão linear implementado pode ser visto no Algoritmo 6.

[2] Outras medidas de fragmentação foram testadas, porém foi por meio desta que foram obtidos os melhores resultados

Algoritmo 6: *Relaxed-SMART* COM REGRESSÃO LINEAR

Entrada: Valores escolhidos arbitrariamente: θ; g^1; α^1; β^1; η; NumMaxIter

início

Seja um estado inicial s

$k = 1$

$R(k) = 0$

$T(k) = 0$

enquanto $k <= NumMaxIter$ **faça**

De acordo com um esquema de exploração-intensificação (seção 5.1.1)

escolha uma ação a

As ações gulosas são escolhidas da seguinte forma:

$a = \underset{a \in A(s)}{\text{argmax}}\ \bar{Q}(s, a|\theta)$

em que

$\bar{Q}(s, a|\theta) = \sum_{f \in F} \theta_f \phi_f(s)$

Aplique a ação a

Simule um evento e obtenha o próximo estado s'

Atualize o vetor de pesos de acordo com as equações do Apêndice B

se *uma ação gulosa foi escolhida* **então**

$R(k) = R(k) + r(s, a, s')$

$T(k) = T(k) + t(s, a, s')$

$g^{k+1} = (1 - \beta^k)g^k + \beta^k \frac{R(k)+r(s,a,s')}{T(k+1)}$

fim

$k = k + 1$

$s \leftarrow s'$

fim

fim

6 EXPERIMENTOS COMPUTACIONAIS

Neste capítulo apresentamos resultados numéricos relativos ao desempenho das políticas de alocação de espectro obtidas tanto pelo modelo proposto no capítulo 4 quanto pelo algoritmo proposto no capítulo 5. Para isso, realizamos os experimentos para dois cenários: o primeiro cenário consiste em um enlace com pequeno número de *slots*, o que possibilita o cálculo de uma política ótima; no segundo cenário, os experimentos são realizados em um enlace de maior tamanho, no qual não é factível se aplicar o modelo analítico. Para ambos os cenários os resultados foram comparados com duas políticas míopes largamente utilizadas na literatura, *First-Fit* e *Best-Fit*.

Para a realização dos experimentos, tem-se como dado de entrada a configuração do enlace: o tamanho do espectro em quantidade de *slots*; número de *slots* contíguos requeridos para cada classe de conexão; taxas de chegada e tempo de transmissão médio de cada tipo de conexão; além da quantidade de *slots* utilizada como banda de guarda. Considerando o primeiro cenário, a partir dos dados de entrada nós construímos o PMD a Tempo Contínuo relacionado, encontramos uma política ótima utilizando o algoritmo de iteração de valores relativo, e geramos a cadeia de Markov relacionada. Para as políticas míopes, nós geramos a cadeia de Markov de acordo com a regra da política para se escolher uma ação. Dessa forma, como saída nós temos as medidas de desempenho de cada política.

No segundo cenário, a partir dos dados de entrada nós aplicamos o algoritmo *Relaxed-SMART* para que uma política seja aprendida. Uma vez que uma política foi aprendida e armazenada na forma de um vetor θ, no modelo de regressão linear, nós aplicamos novamente uma simulação para verificar o desempenho dessa política. As mesmas simulações são aplicadas para as políticas míopes. Os resultados da simulação foram calculados com um intervalo de confiança de 5% e um nível de confiança de 95% Os experimentos foram todos conduzidos em um computador pessoal com processador Intel Core i7-3630QM, CPU a 2,4GHz e 8 GB de memória.

Ao se avaliar uma política de alocação de espectro, além de se observar o desempenho da política para diferentes classes de conexão, deve-se avaliar se ela é eficiente para diferentes padrões de tráfego (*Traffic Patterns* - TPs) que possam surgir na rede. Portanto, para cada cenário utilizamos diferentes padrões de tráfego.

Um dos desafios ao se desenvolver políticas de alocação de espectro é melhorar a utilização do espectro ao mesmo tempo em que se provê uma utilização de recursos igualitária para diferentes classes de conexão. Uma política de alocação de espec-

tro pode, por exemplo, bloquear conexões que requerem mais *slots* contíguos em detrimento das menores, para reduzir a probabilidade de bloqueio geral. Esse comportamento pode levar a uma ineficiência no desempenho da rede. Um indicador de justiça pode ser usado para investigar e medir esse equilíbrio e mostrar quão justa é a rede, considerando a política de alocação de espectro implementada.

Existem diversas medidas de justiça na literatura, porém neste trabalho nós utilizamos o método sugerido por Wang et al. (2014). Esse método é simples porém suficiente para fornecer uma visão de quão justa é a nossa política comparada com as míopes. Define-se o indicador de justiça como a razão $F = PB_b/PB_s$, em que PB_b é a probabilidade de bloqueio do tipo de conexão b, com a maior demanda em termo de quantidade de *slots* contíguos. PB_s, por sua vez, é a probabilidade de bloqueio da classe que demanda menos *slots*. Dessa forma, $F = 1$ representa um sistema justo ideal em que ambos os tipos de chamadas têm a mesma probabilidade de bloqueio. Além da medida de justiça, nós também comparamos os resultados pela probabilidade de bloqueio de *slots* descrita na seção 4.2.

Para ambos os cenários, a demanda de tráfego de cada classe é transformada no número de *slots* correspondentes de acordo com a metodologia utilizada por Castro et al. (2012). O número de *slots* contíguos necessários para alocar uma requisição de conexão do tipo k, w_k, demandando uma largura de banda de $B(k)$ Gb/s é calculado de acordo com: a eficiência espectral do formato de modulação escolhido, denotada por B_{mod} e medida em b/s/Hz, e o tamanho do *slot* $F(S)$ in GHz. Assim,

$$w_k = \left\lceil \frac{B(k)}{B_{mod}\, F(S)} \right\rceil .$$

Consideramos uma granularidade de *slot* de 12,5 GHz, seguindo a definição ITU-T, e o formato de modulação QPSK com $B_{mod} = 2b/s/Hz$. Vale notar que outras metodologias também podem ser utilizadas, visto que o nosso modelo toma como entrada o número de *slots* contíguos, independente de como foi calculado.

6.1 Parâmetros do *Relaxed-SMART*

Os parâmetros utilizados nos experimentos para o algoritmo *Relaxed-SMART* são apresentados nesta seção. Vale notar que, para chegar nessa configuração, foram realizados experimentos com outros parâmetros, e os que contribuíram para uma política com melhor desempenho foram mantidos.

Para escolher uma ação levando em conta o conflito exploração-intensificação testamos três esquemas: a exploração ϵ-gulosa; a exploração *Boltzmann* cujo decaimento da temperatura no decorrer das iterações é dado por

$$T(0) = T_{MAX}$$
$$T(k+1) = T_{MIN} + \kappa(T(k) - T_{MIN})$$

com $\kappa = 0.00002$, $T_{MAX} = 100$ e $T_{MIN} = 0.0001$; e um algoritmo puramente guloso. Os melhores resultados foram obtidos pelo esquema puramente guloso, ou seja, sempre se escolhe a ação relacionada à maior função de valor. De acordo com Powell (2011), caso a representação dos estados, ou pares estado-ação, seja feita por características que capturem aspectos importantes do sistema, muitas vezes a política gulosa apresenta um bom desempenho. Mesmo o esquema de *Boltzmann* pode muitas vezes explorar ações que não são interessantes para o agente.

A taxa de aprendizagem principal na iteração k, α^k, é calculada de acordo com a regra polinomial mostrada na seção 5.1.2.3, com $\delta = 0,55$. Já a segunda taxa de aprendizagem, β^k, é calculada de acordo com a tradicional regra $1/k$. O parâmetro de contração η é definido como $\eta = 0,99$.

Além disso, definimos os seguintes parâmetros: cada elemento do vetor θ é inicializado com um valor aleatório no intervalo $(-1, 1)$; o aprendizado se faz em 10^6 iterações; e o estado inicial s é representado pela grade vazia em conjunto com o evento de uma chegada de conexão do tipo (IN, 1).

6.2 Cenário 1

No cenário 1, conduzimos experimentos para um enlace com o maior número possível de *slots* que nossa implementação pôde resolver, considerando as restrições de memória. Nesses experimentos tem-se como entrada um enlace com espectro de 275 GHz, equivalente a 22 *slots*, com duas classes de conexão: classe 1, demandando 10 Gb/s (1 *slot*), e classe 2 demandando 100 Gb/s (4 *slots*). Além disso, um *slot* é utilizado como banda de guarda e todos os tipos de requisição de conexão têm o mesmo tempo médio de transmissão, $1/\mu = 1$. Dadas essas configurações, variamos a taxa de chegada das requisições de conexão, com o intuito de observar o desempenho de cada política ao se aumentar a carga no enlace.

Os resultados são apresentados com a probabilidade de bloqueio de *slots* variando

de 10^{-4} a 1%, uma faixa de bloqueio de conexões aceitável para as operadoras. Vale mencionar que, mesmo para esse cenário, o tamanho do espaço de estados é aproximadamente 936.000 com mais de 10 milhões de transições de probabilidade não-nulas entre esses estados.

Neste cenário fornecemos resultados numéricos relacionados a três padrões de tráfego distintos: TP 1, um padrão no qual 80% das requisições demandam uma largura de banda de 100 Gb/s; TP 2, um tráfego uniforme, em que cada classe tem a mesma taxa de chegada; e TP 3, no qual há uma maior carga de conexões demandando menos *slots*. Os parâmetros desses três padrões de tráfego são reunidos na Tabela 6.1.

Tabela 6.1 - Cenário 1: demanda de cada classe por padrão de tráfego.

Padrão de Tráfego	Classe 1 (10 GB/s)	Classe 2 (100 Gb/s)
TP 1	20%	80%
TP 2	50%	50%
TP 3	80%	20%

6.2.1 Avaliação de Desempenho

Na Figura 6.1 podem ser observadas as probabilidade de bloqueio de *slots* (PBS) da política resultante do PMD para cada padrão de tráfego. Essa probabilidade varia de acordo com o eixo y secundário. Para cada padrão de tráfego, apresentamos também o *gap* entre a PBS das política do algoritmo *Relaxed-SMART* (denotado apenas por R-SMART), *First-Fit* e *Best-Fit* em relação à política do PMD. Esse *gap* é denotado pelas barras seguindo os valores do eixo y primário. Além disso, o *gap* é calculado como o erro percentual das PBS das demais políticas para a política do PMD e, portanto, quando o *gap* é positivo, a política tem PBS maior que a política ótima. Os valores são plotados no gráfico em função da carga oferecida no enlace, ou seja, a taxa total de chegada de requisições de conexão no enlace multiplicada pelo tempo médio de serviço.

Pode-se notar que a PBS da política do PMD cresce exponencialmente ao passo que a carga oferecida cresce, como esperado. A política *First-Fit* apresenta a maior PBS para todos os padrões de tráfego, seguida pela *Best-Fit*. No TP 1, o *gap* da R-SMART varia em torno de 20% a 25%, enquanto que a *Best-Fit* e *First-Fit* apresentam, respectivamente, um *gap* em torno de 30%-35% e 35%-40%. Ao se observar a medida de justiça para o TP 1 na Figura 6.2, nota-se que a política do PMD é a mais justa de

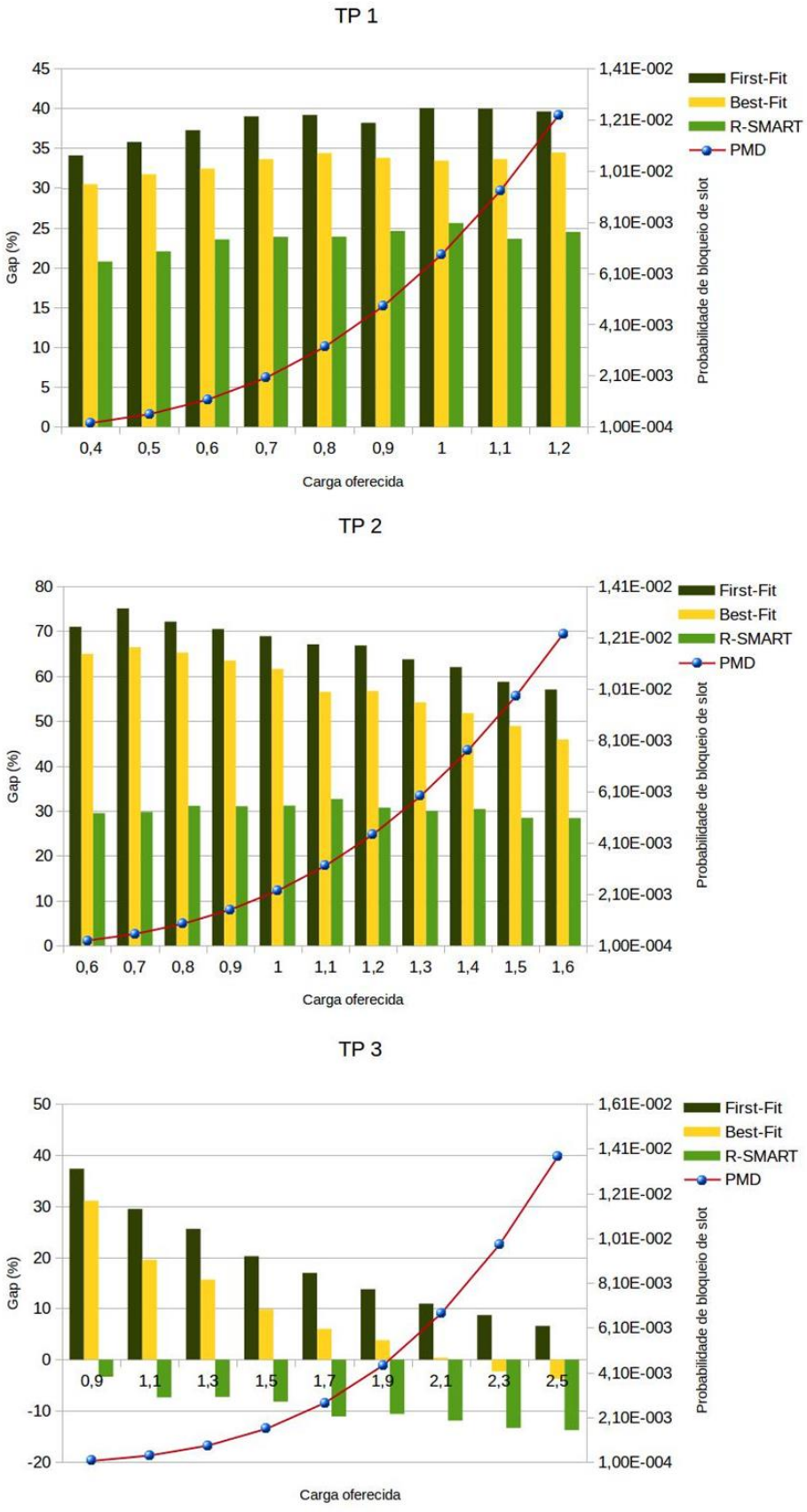

Figura 6.1 - Probabilide de bloqueio de *slot* em função da carga oferecida.

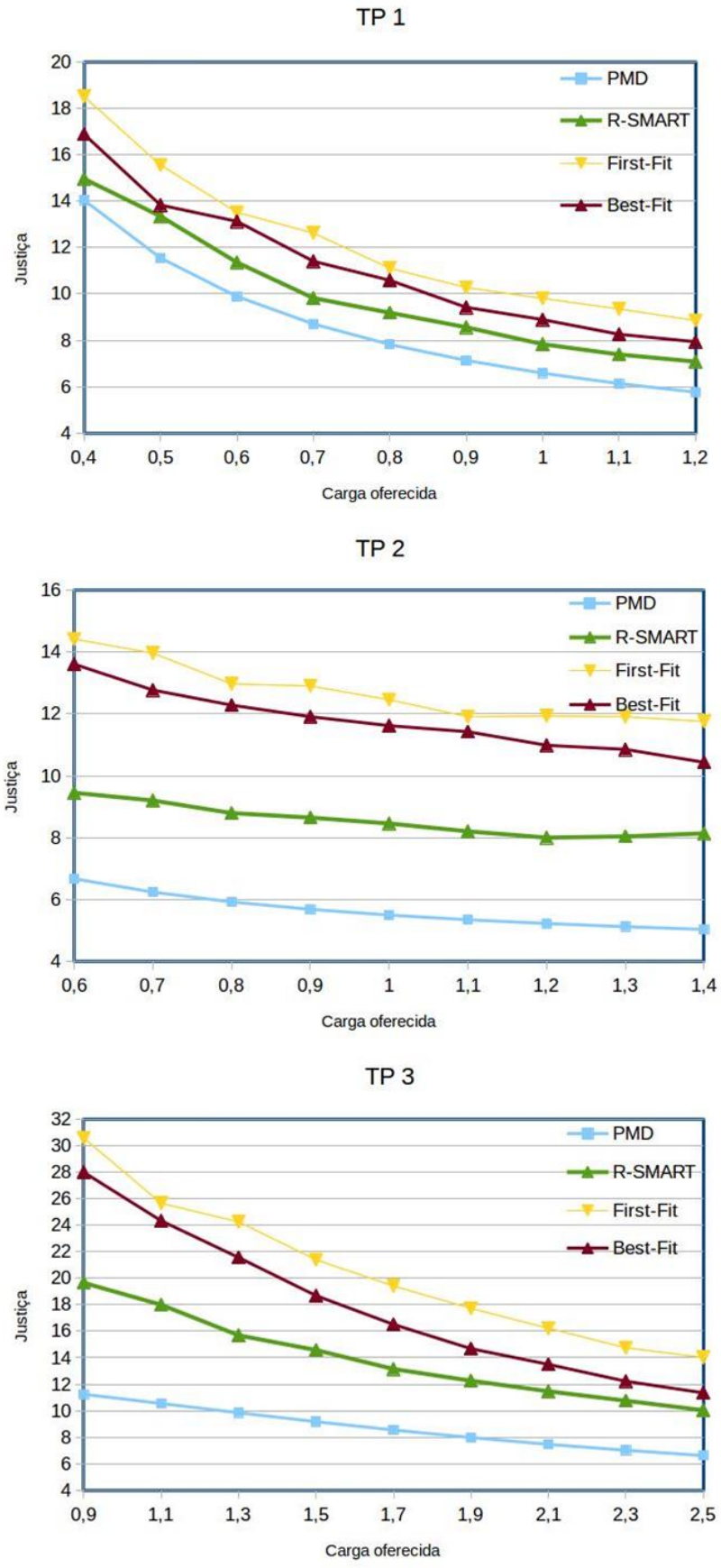

Figura 6.2 - Medida de justiça em função da carga oferecida.

acordo com a métrica, seguida pela R-SMART, *First-Fit* e *Best-Fit*. Podemos notar esse mesmo comportamento também para os outros TPs. Por fim, outra maneira de se verificar a justiça é examinando os valores das Tabelas 6.2 e 6.3, nas quais são apresentadas, respectivamente, os *gaps* relacionados às probabilidades de bloqueio da classe 1 e da classe 2. Analisando as probabilidades de bloqueio da classe 1, os *gaps* das políticas são negativos para todas as cargas. As probabilidades de bloqueio da R-SMART são as mais próximas da política ótima e, portanto, as políticas míopes bloqueiam menos conexões que exigem apenas 1 *slot*. Elas são, no entanto, as que mais bloqueiam conexões da classe 2 (Tabela 6.3).

Para o TP 2, no qual ambas as classes têm a mesma carga, os resultados apresentados se assemelham ao TP 1. No entanto, observa-se na Figura 6.1 que o *gap* da PBS aumenta principalmente para as políticas míopes. Enquanto o *gap* da R-SMART se mantém em torno dos 30%, o do *First-Fit* varia aproximadamente de 55% à 75%, e o *Best-Fit* de 45% à 65%. Ao se analisar a métrica de justiça, nota-se que os valores não estão tão próximos como no TP 1. A política do PMD continua sendo a mais justa, seguida pela R-SMART e pelas políticas míopes. Podemos notar também uma maior diferença nos *gaps* na probabilidade de bloqueio das requisições em relação à política do PMD nas Tabelas 6.4 e 6.5. Uma diferença para as tabelas relativas ao TP 1 é que, para a classe 1, o *gap* negativo tem o módulo maior ao se aumentar a carga, para todas as políticas. Ou seja, ao se aumentar a carga, as chamadas do tipo 1 são as menos bloqueadas. Já para a classe 2, a tendência do *gap* das políticas míopes é diminuir em função do aumento da carga, enquanto o *gap* da R-SMART se mostra estabilizado.

Por fim, analisando a PBS do TP 3 notamos um comportamento diferente dos demais. O *gap* das políticas míopes apesar de começar na faixa de 30% a 40% chega próximo a 0% no caso do *First-Fit*, ou em torno de -12% para o *Best-Fit*. Já o R-SMART apresenta um *gap* negativo para todas as cargas plotadas. Esse *gap* não significa, no entanto, que o desempenho dessas políticas foi melhor que da política do PMD, pois temos que levar em conta também a justiça no bloqueio das conexões. Nota-se então, na Figura 6.2, que a política do PMD continua sendo a mais justa, embora a diferença para as demais caia de acordo com o aumento da carga. Isso acontece pelo fato de haver mais chegadas de chamadas da classe 1 no TP 3, e, portanto, como a política do PMD consegue ainda alocar chamadas da classe 2, o bloqueio das chamadas menores cresce muito. As probabilidades de bloqueio de cada classe podem ser consultadas nas Tabelas 6.6 e 6.7.

Vale ressaltar que a política *First-Fit* é a mais simples de ser implementada e demanda um menor esforço computacional, porém ela causa uma pior performance. Embora um esforço computacional consideravelmente maior seja necessário para encontrar a política ótima do PMD, essa política é calculada *offline* e é implementada em uma tabela. Portanto, o esforço computacional em tempo real é reduzido apesar da grande quantidade de memória requerida para armazenar tal tabela.

Tabela 6.2 - TP 1: *Gap* da probabilidade de bloqueio da classe 1 em relação à política ótima.

Carga	R-SMART	*First-Fit*	*Best-Fit*
0,4	-2,37	-3,88	-14,64
0,5	-8,77	-12,83	-4,93
0,6	-6,78	-12,90	-13,48
0,7	-4,60	-16,61	-11,31
0,8	-8,15	-14,51	-13,43
0,9	-9,34	-16,11	-11,47
1	-7,59	-17,59	-13,43
1,1	-9,95	-19,29	-12,82
1,2	-10,93	-19,95	-14,08

Tabela 6.3 - TP 1: *Gap* da probabilidade de bloqueio da classe 2 em relação à política ótima.

Carga	R-SMART	*First-Fit*	*Best-Fit*
0,4	4,04	15,56	12,50
0,5	5,47	17,43	13,87
0,6	7,07	19,06	14,87
0,7	7,61	20,93	16,20
0,8	7,93	21,38	17,18
0,9	8,82	20,83	16,92
1	9,93	22,75	16,90
1,1	8,44	22,99	17,33
1,2	9,42	22,94	18,30

Tabela 6.4 - TP 2: *Gap* da probabilidade de bloqueio da classe 1 em relação à política ótima.

Carga	R-SMART	*First-Fit*	*Best-Fit*
0,6	-7,57	-19,27	-17,57
0,7	-10,89	-20,02	-17,01
0,8	-10,47	-19,56	-18,54
0,9	-12,59	-23,03	-20,16
1	-13,39	-23,57	-21,70
1,1	-12,13	-23,03	-24,93
1,2	-13,24	-25,00	-23,62
1,3	-15,74	-27,60	-25,36
1,4	-17,78	-28,63	-24,90
1,5	-16,99	-30,38	-26,47
1,6	-17,58	-30,80	-26,98

Tabela 6.5 - TP 2: *Gap* da probabilidade de bloqueio da classe 2 em relação à política ótima.

Carga	R-SMART	*First-Fit*	*Best-Fit*
0,6	30,86	74,36	68,03
0,7	31,34	78,85	69,74
0,8	32,86	75,97	68,78
0,9	32,91	74,59	67,17
1	33,17	73,09	65,39
1,1	34,73	71,28	60,27
1,2	32,81	71,17	60,49
1,3	32,20	68,22	58,04
1,4	32,80	66,50	55,53
1,5	30,70	63,21	52,70
1,6	30,73	61,51	49,63

Tabela 6.6 - TP 3: *Gap* da probabilidade de bloqueio da classe 1 em relação à política ótima.

Carga	R-SMART	*First-Fit*	*Best-Fit*
0,9	-13,93	-19,99	-16,87
1,1	-15,75	-16,08	-18,47
1,3	-10,23	-19,61	-17,16
1,5	-11,12	-19,03	-15,83
1,7	-11,51	-19,34	-14,76
1,9	-11,26	-20,00	-12,85
2,1	-12,68	-20,39	-14,43
2,3	-14,26	-19,66	-14,07
2,5	-14,11	-21,95	-14,28

Tabela 6.7 - TP 3: *Gap* da probabilidade de bloqueio da classe 2 em relação à política ótima.

Carga	R-SMART	*First-Fit*	*Best-Fit*
0,9	50,42	117,45	106,89
1,1	43,69	104,05	87,99
1,3	42,95	97,80	81,27
1,5	41,18	88,71	71,35
1,7	36,06	83,09	64,62
1,9	36,39	77,71	60,34
2,1	33,99	72,78	54,55
2,3	31,35	68,73	49,62
2,5	30,17	65,16	46,89

6.3 Cenário 2

Neste cenário consideramos um enlace com largura de espectro de 4000 GHz, ou 320 *slots*. Esse parâmetro foi baseado nos experimentos de Wan et al. (2012). Consideramos quatro tipos de requisição de conexão que demandam, respectivamente, 100 Gb/s, 200 Gb/s, 400 Gb/s e 1 Tb/s (4, 8, 16 e 40 *slots*) baseado em Amar et al. (2014). Além disso, adotamos 1 *slot* para frequência de guarda e todos os tipos de conexão apresentam o mesmo tempo de transmissão médio ($1/\mu = 1$). Dadas essas configurações, variamos a carga oferecida ao enlace.

Assim como no primeiro cenário, os resultados são apresentados com a probabilidade de bloqueio de *slots* variando de 10^{-4} a 1%. Além disso, três padrões de tráfego distintos foram utilizados, como mostrado na Tabela 6.8. No TP 1, o intuito é

verificar o comportamento das políticas para uma rede com tráfego maior para as chamadas que demandam mais *slots*; no TP 2 temos um tráfego uniforme; e o TP 3 apresenta uma demanda maior para chamadas que utilizam menos *slots*.

Tabela 6.8 - Cenário 2: demanda de cada classe por padrão de tráfego.

Padrão de Tráfego	Classe 1	Classe 2	Classe 3	Classe 4
TP 1	10%	20%	30%	40%
TP 2	25%	25%	25%	25%
TP 3	40%	30%	20%	10%

6.3.1 Avaliação de Desempenho

Nesse cenário as probabilidade de bloqueio de *slots* da política resultante do R-SMART são mostradas para cada padrão de tráfego, variando de acordo com o eixo y secundário. Tomando como base a política R-SMART, são apresentados os *gaps* para as políticas míopes, denotados pelas barras seguindo os valores do eixo y primário. Portanto, quando o *gap* é positivo, a política tem PBS maior que a política R-SMART. Nesse cenário a medida de justiça foi calculada levando em conta as probabilidades de bloqueio da classe 2 e 4, respectivamente, pois as probabilidades de bloqueio da classe 1 em alguns experimentos era muito baixa, o que atrapalhava a visualização nos gráficos.

Pode-se notar pela Figura 6.3 que, a exemplo do primeiro cenário, a política *First-Fit* apresenta o maior PBS para todos os padrões de tráfego, seguida pela *Best-Fit*. No TP 1, o *gap* da *First-Fit* varia em torno de 28% a 35%, enquanto que a *Best-Fit* apresenta um *gap* variando de 5% a 15%. Por meio da medida de justiça, apresentada na Figura 6.4, observa-se que a política R-SMART é a mais justa, seguida pela *Best-Fit*. A política *First-Fit*, por sua vez, apresenta uma maior diferença na medida de justiça em relação as outras.

Para o TP 2, os resultados apresentados se assemelham ao TP 1. Os *gaps* da PBS da *First-Fit* e *Best-Fit* variam, respectivamente, de 30% a 43% e 4% a 20%. Como nesse padrão de tráfego não se tem o predomínio de nenhum tipo de conexão, o espectro pode apresentar uma maior fragmentação ao longo do tempo. Assim, é interessante notar que nesse caso a PBS do R-SMART manteve-se ainda mais baixo em relação às demais políticas. A política *First-Fit* continua sendo a mais injusta, de acordo com a métrica adotada. Para as maiores cargas, a justiça do R-SMART e *Best-Fit* estão mais próximas.

Por fim, para o TP 3, o *gap* da *First-Fit* em relação ao R-SMART é o maior observado entre os padrões de tráfego, no intervalo de 35% à 62%. Como a carga de chamadas menores é maior nesse padrão, mais conexões simultâneas podem ser alocadas, e esse *gap* maior aponta a falha da *First-Fit* em fazer isso em relação às demais políticas. Já o *gap* do *Best-Fit* se manteve semelhante ao do segundo padrão de tráfego. Pode-se notar, a exemplo dos demais padrões de tráfego, que a política R-SMART se manteve com o melhor índice de justiça, próxima à *Best-Fit* nas cargas mais elevadas. A *First-Fit*, como esperado, apresenta uma medida mais distante das demais.

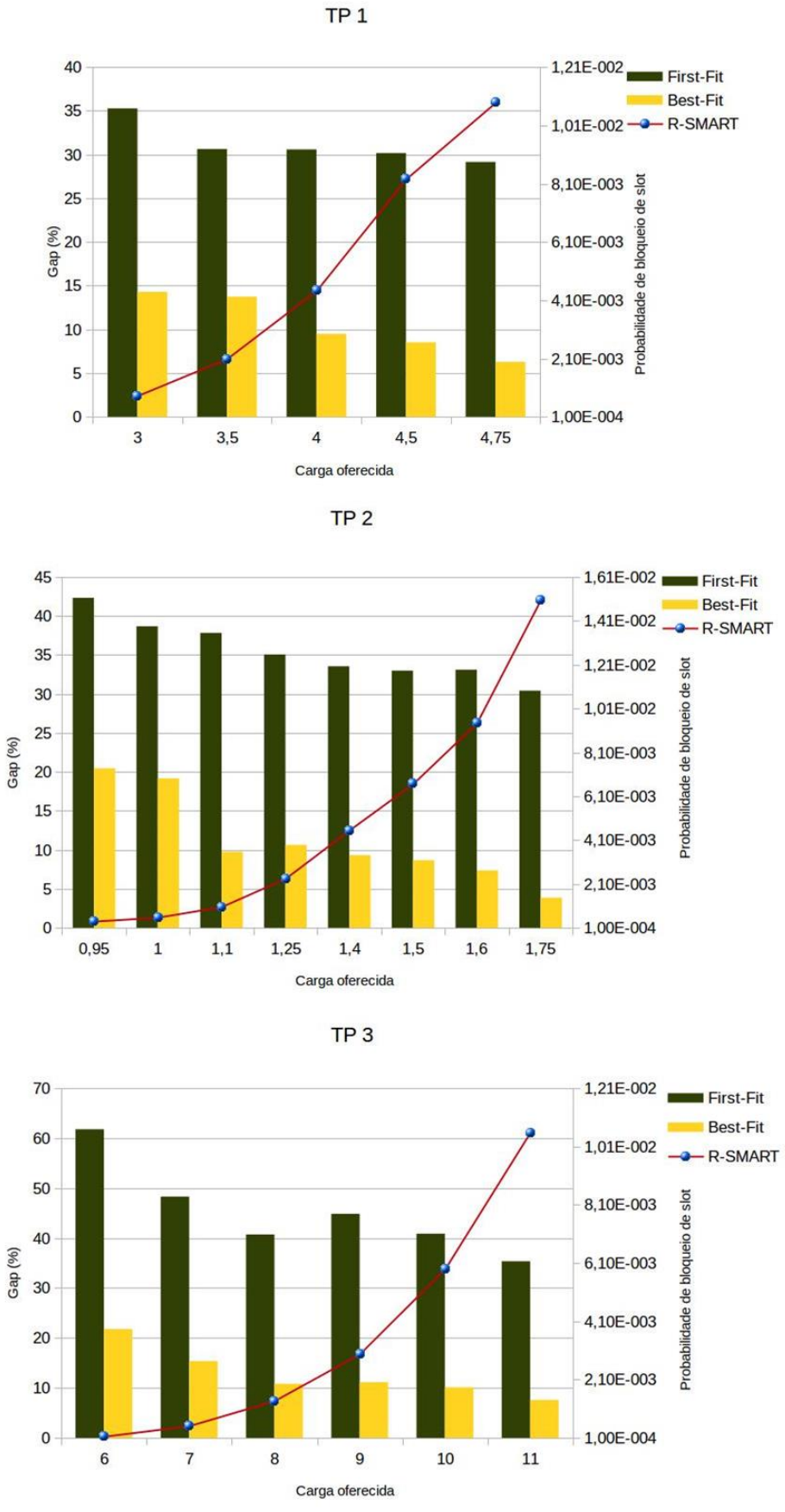

Figura 6.3 - Probabilide de bloqueio de *slot* em função da carga oferecida.

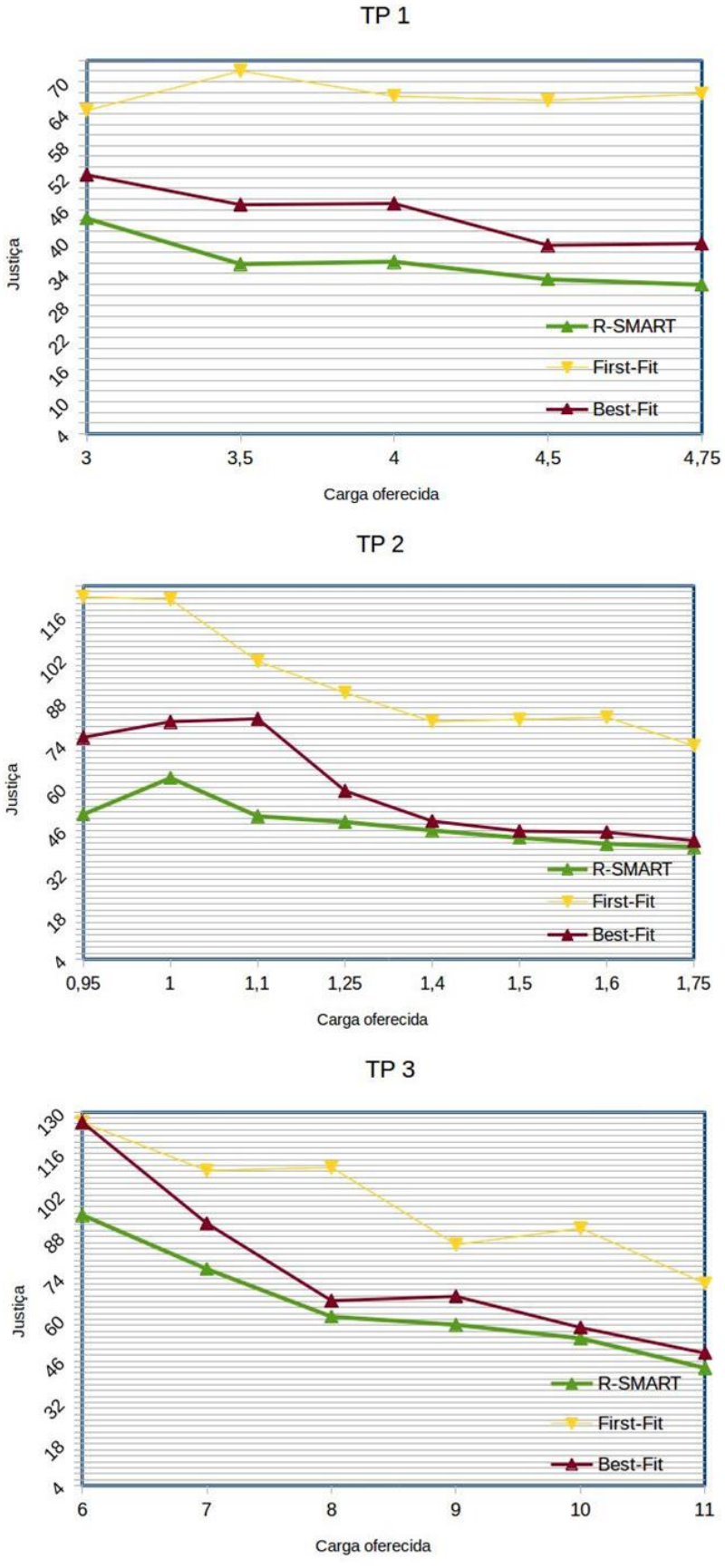

Figura 6.4 - Medida de justiça em função da carga oferecida.

7 CONSIDERAÇÕES FINAIS

As redes ópticas elásticas vêm sendo estudadas nos últimos anos e estima-se que em pouco tempo sejam implantadas em larga escala. Desenvolver métodos que resolvam os problemas relacionadas a essas redes, tal como o problema RSA, torna-se uma parte importante desse processo de implantação. Nesse contexto, essa tese contribui com uma nova maneira de abordar o problema de alocação de espectro. Ao modelar um enlace de uma rede óptica elástica como um PMD, nós conseguimos encontrar políticas de alocação de espectro que superam as políticas míopes. Apesar das políticas míopes serem atrativas pela fácil implementação e pouco esforço computacional exigido, os equipamentos das operadoras possuem cada vez mais capacidade de processamento, tornando-se viável a implementação de algoritmos que demandam mais recursos computacionais. Cabe as operadoras, então, escolher qual política utilizar, levando em conta suas necessidades e recursos disponíveis. Além disso, a abordagem do problema de alocação de espectro como um PMD pode ser ainda adaptada para outros problemas de alocação de recursos em telecomunicações.

Neste trabalho nós propusemos o modelo de um único enlace, o qual consideramos um pré-requisito para trabalhos futuros com topologias de rede arbitrárias. Desenvolver uma política de alocação de espectro eficiente para um enlace é importante, dado que em outras topologias a alocação dos *slots* deve ser feita em todos os enlaces de cada rota. Se tal alocação não for eficiente, mais conexões serão bloqueadas e a rede pode apresentar uma baixa performance.

Ao desenvolver o modelo analítico, fornecemos uma ferramenta por meio da qual outras políticas podem ser implementadas e validadas analiticamente para pequenas instâncias do problema. Para instâncias mais realistas, nós propusemos a utilização de um algoritmo de aprendizagem por reforço que apresentou resultados satisfatórios em relação às políticas míopes. Dessa maneira, viabilizamos a aplicação da nossa abordagem em circunstâncias reais.

Dentre as vantagens dos algoritmos de aprendizagem por reforço tem-se a possibilidade de se aliar simulação à otimização. Torna-se possível procurar políticas de controle a partir de um simulador do sistema, sem a necessidade de se armazenar em memória todos os dados do problema modelado como um PMD. Além disso, a aproximação de função permite que, mesmo para modelos com grandes quantidades de estados e/ou ações, as políticas sejam armazenadas e implementadas por um conjunto pequeno de parâmetros.

Observamos, por meio dos experimentos realizados, que a utilização de uma medida de justiça torna-se útil ao se trabalhar com redes ópticas elásticas, visto que demandas de largura de banda para diferentes classes de conexão são consideradas. Além disso, nós focamos na ideia de que se uma política de alocação de espectro eficiente é aplicada, então a fragmentação do espectro ao longo do tempo será diminuída. Dessa maneira, diferentes classes de conexão terão mais chances de ser alocadas com sucesso, resultando em uma maior justiça entre as chamadas da rede.

Por fim, outro aspecto deste trabalho foi a aplicação de PMDs a Tempo Contínuo com critério de recompensa média, que não são tão largamente estudados quanto os PMDs sob critério de recompensa descontada. Contribuímos, portanto, para um maior entendimento e disseminação desses modelos específicos que podem ser utilizados em outros problemas de engenharia.

7.1 Trabalhos Futuros

Como sugestão de trabalho futuro, pode-se pesquisar a viabilidade de se encontrar algum conhecimento implícito na política ótima encontrada para o modelo descrito no capítulo 4. Tal conhecimento pode ser representado por regras por meio das quais as decisões são tomadas. Pode-se ainda verificar se tais regras, caso existam, variam de acordo com cada instância do problema. Caso um comportamento seja confirmado, políticas mais simples podem ser derivadas, embora a garantia de otimalidade seja perdida.

Pode-se também investigar a adaptação de outros métodos de aprendizado de máquina em conjunto com o algoritmo *Relaxed-SMART*. Neste trabalho, duas das motivações de se escolher a regressão linear foram: o fato desta ser simples de ser implementada; poucos cálculos são efetuados em tempo real para encontrar a ação para cada estado do sistema. O esforço computacional para se encontrar tais ações tem grande relevância para os operadores de redes, visto que esse cálculo pode ser executado em nós de redes com poucos recursos computacionais. No entanto, métodos como redes neurais artificiais podem prover melhor generalização para a aproximação das funções de valor para os pares estado-ação, resultando em uma política mais eficiente.

O modelo de um enlace proposto neste trabalho pode ser também estendido para outras topologias de redes ópticas e, dessa forma, o problema RSA completo é tratado. A medida que a topologia cresce o problema se torna mais desafiador, de modo que torna-se importante a obtenção de políticas em tempo computacional

factível. Inicialmente a topologia de anel pode ser adotada, pois nessa topologia o roteamento é simples e menos enlaces são necessários. Uma ideia é que o modelo estendido compreenda um caminho óptico completo, e não apenas um enlace.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, R. C.; DELGADO, R. A.; BASTOS-FILHO, C. J. A.; CHAVES, D. A. R.; PEREIRA, H. A.; MARTINS-FILHO, J. F. An evolutionary spectrum assignment algorithm for elastic optical networks. In: 2013 15TH INTERNATIONAL CONFERENCE ON TRANSPARENT OPTICAL NETWORKS (ICTON), 2013, Spain. **Proceedings...** Spain: IEEE, 2013. p. 1 – 3. 18

AMAR, D.; Le Rouzic, E.; BROCHIER, N.; AUGE, J.-L.; LEPERS, C.; PERROT, N.; FAZEL, S. How problematic is spectrum fragmentation in operator's gridless network? In: 2014 INTERNATIONAL CONFERENCE ON OPTICAL NETWORK DESIGN AND MODELING (ONDM), 2014, Stockholm, Sweden. **Proceedings...** Stockholm, Sweden: IEEE, 2014. p. 67–72. 60

BELLMAN, R. **Dynamic programming**. 1. ed. Princeton (NJ): Princeton University Press, 1957. 26

BERTSEKAS, D. P.; TSITSIKLIS, J. N. **Neuro-dynamic programming**: optimization and neural computation series. Nashua, NH: Athena Scientific, 1996. ISBN 1886529108. 48

BISBAL, D.; MIGUEL, I. d.; GONZÁLEZ, F.; BLAS, J.; AGUADO, J. C.; FERNÁNDEZ, P.; DURÁN, J.; DURÁN, R.; LORENZO, R. M.; ABRIL, E. J.; LÓPEZ, M. Dynamic routing and wavelength assignment in optical networks by means of genetic algorithms. **Photonic Network Communications**, Springer Netherlands, v. 7, p. 43–58, 2004. 14

CASTRO, A.; VELASCO, L.; RUIZ, M.; KLINKOWSKI, M.; PALACIOS, J. P. F.; CAREGLIO, D. Dynamic routing and spectrum (re)allocation in future flexgrid optical networks. **Computer Networks**, v. 56, n. 12, p. 2869–2883, 2012. ISSN 13891286. 2, 17, 52

CHLAMTAC, I.; GANZ, A.; KARMI, G. Lightpath communications: an approach to high bandwidth optical wan's. **Communications, IEEE Transactions on**, v. 40, n. 7, p. 1171 –1182, jul 1992. ISSN 0090-6778. 12

CHRISTODOULOPOULOS, K.; TOMKOS, I.; VARVARIGOS, E. Time-varying spectrum allocation policies and blocking analysis in flexible optical networks. **IEEE Journal on Selected Areas in Communications**, v. 31, n. 1, p. 13–25,

jan. 2013. ISSN 0733-8716. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6381740>. 2, 32

CHRISTODOULOPOULOS, K.; TOMKOS, I.; VARVARIGOS, E. A. Routing and spectrum allocation in ofdm-based optical networks with elastic bandwidth allocation. In: 2010 IEEE GLOBAL TELECOMMUNICATIONS CONFERENCE GLOBECOM 2010, 2010, Miami, Florida. **Proceedings...** Miami, Florida: IEEE, 2010. p. 1–6. 2

_____. Elastic bandwidth allocation in flexible ofdm-based optical networks. **Journal of Lightwave Technology**, v. 29, n. 9, p. 1354–1366, 2011. ISSN 0733-8724. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5727897>. 1

CISCO. **Cisco visual networking index**: forecast and methodology, 2013-2018. 06 2014. 1

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 3. ed. Hoboken, NJ: Wiley-Interscience, 1998. ISBN 0471170828. 48

GANESAN, R.; DAS, T. K.; RAMACHANDRAN, K. M. A multiresolution analysis-assisted reinforcement learning approach to run-by-run control. **IEEE Transactions on Automation Science and Engineering**, v. 4, n. 2, p. 182–193, abr. 2007. ISSN 1545-5955. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4147550>. 40

GERSTEL, O.; JINNO, M.; LORD, A.; YOO, S. Elastic optical networking: a new dawn for the optical layer? **IEEE Communications Magazine**, v. 50, n. 2, p. s12–s20, 2012. ISSN 0163-6804. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6146481>. 2, 15

GOSAVI, A. Reinforcement learning for long-run average cost. **European Journal of Operational Research**, v. 155, n. 3, p. 654–674, jun. 2004. ISSN 03772217. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0377221702008743>. 8, 40, 42

_____. Target-sensitive control of Markov and semi-Markov processes. **International Journal of Control, Automation and Systems**, v. 9, n. 5, p.

941–951, out. 2011. ISSN 1598-6446. Disponível em: <http://link.springer.com/10.1007/s12555-011-0515-6>. 8, 40, 42

GUO, X.; HERNÁNDEZ-LERMA, O. **Continuous-time markov decision processes**: theory and applications. 1. ed. Berlin: Springer, 2009. 22, 23

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of statistical learning**. New York: Springer, 2003. ISBN 0387952845. 48

HORAK, R. **Telecommunications and data communications handbook**. 2. ed. Hoboken (NJ): Wiley-Interscience, 2008. 11

HU, Q.; YUE, W. **Markov decision processes with their applications**. 1. ed. New York: Springer, 2007. 5

HYYTIA, E.; VIRTAMO, J. Dynamic routing and wavelength assignment using first policy iteration. In: FIFTH IEEE SYMPOSIUM ON COMPUTERS AND COMMUNICATIONS. ISCC 2000, 2000, France. **Proceedings...** France: IEEE Computer Society, 2000. p. 146 – 151. 31, 39

ITU-T. **Spectral grids for WDM applications**: dwdm frequency grid. Feb, 2012. 15

JAUMARD, B.; MEYER, C.; THIONGANE, B.; YU, X. Ilp formulations and optimal solutions for the rwa problem. In: GLOBAL TELECOMMUNICATIONS CONFERENCE, GLOBECOM '04, 2004, Dallas, USA. **Proceedings...** Dallas, USA: IEEE, 2004. p. 1918 – 1924. 13

KLINKOWSKI, M.; WALKOWIAK, K. Routing and spectrum assignment in spectrum sliced elastic optical path network. **IEEE Communications Letters**, v. 15, n. 8, p. 884–886, 2011. ISSN 1089-7798. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5910089>. 2, 17

LAAKSU, M.; TAAGEPERA, R. Effective number of parties: a measure with applications to west europe. **Comparative Political Studies**, Sage Publications, Inc., v. 12, n. 1, p. 3–27, 4 1979. 49

LEW, A.; MAUCH, H. **Dynamic programming**: a computational tool. 1. ed. New York: Springer, 2006. (Studies in Computational Intelligence). 26

MOSHARAF, K.; LAMBADARIS, I.; TALIM, J.; SHOKRANI, A. Fairness control in wavelength-routed wdm ring networks. In: IEEE GLOBAL

TELECOMMUNICATIONS CONFERENCE, 2005, St. Louis, USA. **Proceedings...** St. Louis: IEEE Communications Society, 2005. p. 2091–2095. 13

MOSHARAF, K.; TALIM, J.; LAMBADARIS, I. A markov decision process model for dynamic wavelength allocation in wdm networks. In: IEEE GLOBAL TELECOMMUNICATIONS CONFERENCE, 2003, San Francisco, USA. **Proceedings...** San Francisco: IEEE Communications Society, 2003. p. 2590–2594. 5, 31, 39

_____. Optimal resource allocation and fairness control in all-optical wdm networks. **IEEE Journal on Selected Areas in Communications**, v. 23, n. 8, p. 1496–1507, 2005. 31, 39

OZDAGLAR, A. E.; BERTSEKAS, D. P. Routing and wavelength assignment in optical networks. **IEEE/ACM Transactions on Networking**, v. 11, n. 2, p. 259–272, 2003. 2

POWELL, W. B. **Approximate dynamic programming**: solving the curses of dimensionality (wiley series in probability and statistics). Hoboken, NJ: Wiley-Interscience, 2007. ISBN 0470171553. 5, 43

_____. **Approximate dynamic programming**: solving the curses of dimensionality. 2nd. ed. Hoboken, New Jersey: John Wiley & Sons, Inc, 2011. ISBN 978-0-470-60445-8. Disponível em: <http://www.wiley.com/WileyCDA/WileyTitle/productCd-047060445X.html>. 39, 48, 53

PUTERMAN, M. L. **Markov decision processes**: discrete stochastic dynamic programming. Hoboken, NJ: Wiley-Interscience, 2005. 4, 9, 21, 22, 25, 28, 42

RAMASWAMI, R.; SIVARAJAN, K. N. Routing and wavelength assignment in all-optical networks. **IEEE/ACM Transactions on Networking**, v. 3, n. 5, p. 489–500, 1995. 13

ROBBINS, H.; MONRO, S. A stochastic approximation method. **Ann. Math. Statist.**, The Institute of Mathematical Statistics, v. 22, n. 3, p. 400–407, 09 1951. Disponível em: <http://dx.doi.org/10.1214/aoms/1177729586>. 42

ROSA, A.; CAVDAR, C.; CARVALHO, S.; COSTA, J.; WOSINSKA, L. Spectrum allocation policy modeling for elastic optical networks. In: HIGH CAPACITY OPTICAL NETWORKS AND EMERGING/ENABLING TECHNOLOGIES, 2012, Istanbul. **Proceedings...** Istanbul: IEEE, 2012. p. 242–246. 18

ROSS, S. M. **Introduction to probability models**. Tenth edition. Boston: Academic Press, 2010. ISBN 978-0-12-375686-2. 29

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence**: a modern approach. 3rd. ed. New Jersey: Prentice Hall, 2009. ISBN 0136042597. 5, 6

SAENGUDOMLERT, P.; MODIANO, E.; GALLAGER, R. G. On-line routing and wavelength assignment for dynamic traffic in wdm ring and torus networks. **IEEE/ACM Trans. Netw.**, IEEE Press, v. 14, n. 2, p. 330–340, abr. 2006. 14

SHI, W.; ZHU, Z.; ZHANG, M.; ANSARI, N. On the effect of bandwidth fragmentation on blocking probability in elastic optical networks. **IEEE Transactions on Communications**, v. 61, n. 7, p. 2970–2978, July 2013. ISSN 0090-6778. 49

SHIRAZIPOURAZAD, S.; DERAKHSHANDEH, Z.; SEN, A. Analysis of on-line routing and spectrum allocation in spectrum-sliced optical networks. In: 2013 IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS (ICC), 2013, Hungary. **Proceedings...** Hungary: IEEE, 2013. p. 3899 – 3903. 2, 17

SIGAUD, O.; BUFFET, O. **Markov decision processes in artificial intelligence**. 1. ed. Hoboken, NJ: Wiley-IEEE Press, 2010. 5, 24, 39

SIMÃO, H. P.; DAY, J.; GEORGE, A. P.; GIFFORD, T.; NIENOW, J.; POWELL, W. B. An approximate dynamic programming algorithm for large-scale fleet management: A case application. **Transportation Science**, v. 43, n. 2, p. 178–197, 2009. 5

SIVALINGAM, K. M.; SUBRAMANIAM, S. (Ed.). **Optical WDM networks**: principles and practice. 1. ed. New York: Springer, 2000. 12

SUTTON, R. S.; BARTO, A. G. **Introduction to reinforcement learning**. 1st. ed. Cambridge, MA, USA: MIT Press, 1998. ISBN 0262193981. 6, 7

TACHIBANA, T.; KASAHARAT, S.; SUGIMOTO, K. Dynamic lightpath establishment for service differentiation based on optimal mdp policy in all-optical networks with wavelength conversion. In: IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS, 2007, Glasgow, UK. **Proceedings...** Glasgow: IEEE Communications Society, 2007. p. 2424–2429. 5, 31

TALEBI, S.; ALAM, F.; KATIB, I.; KHAMIS, M.; SALAMA, R.; ROUSKAS, G. N. Spectrum management techniques for elastic optical networks: A survey.

Optical Switching and Networking, v. 13, p. 34–48, jul. 2014. ISSN 15734277. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1573427714000253>. 17

TANENBAUM, A. S. **Computer networks**. 4. ed. Upper Saddle River (NJ): Prentice Hall, 2002. 11

TIJMS, H. C. **Stochastic models**: an algorithmic approach. Chichester (UK): John Wiley & Sons, 1995. 9, 21, 24, 27, 28

_____. **A first course in stochastic models**. 2. ed. Chichester (UK): Wiley, 2003. 5, 22, 25

WAN, X.; HUA, N.; ZHENG, X. Dynamic routing and spectrum assignment in spectrum-flexible transparent optical networks. **Journal of Optical Communications and Networking**, v. 4, n. 8, p. 603, 2012. ISSN 1943-0620. Disponível em: <http://www.opticsinfobase.org/abstract.cfm?URI=jocn-4-8-603>. 2, 60

WAN, X.; WANG, L.; HUA, N.; ZHANG, H.; ZHENG, X. Dynamic routing and spectrum assignment in flexible optical path networks. In: OPTICAL FIBER COMMUNICATION CONFERENCE AND EXPOSITION (OFC/NFOEC), 2011 AND THE NATIONAL FIBER OPTIC ENGINEERS CONFERENCE, 2011, Los Angeles, CA. **Proceedings...** Los Angeles, CA: IEEE, 2011. p. 1–3. 17

WANG, X.; KIM, J.; YAN, S.; RAZO, M.; TACCA, M.; FUMAGALLI, A. Blocking probability and fairness in two-rate elastic optical networks. In: 16TH INTERNATIONAL CONFERENCE ON TRANSPARENT OPTICAL NETWORKS (ICTON), 2014, Austria. **Proceedings...** Austria: IEEE Computer Society, 2014. p. 1–4. 52

WANG, X.; ZHANG, Q.; KIM, I.; PALACHARLA, P.; SEKIYA, M. Utilization entropy for assessing resource fragmentation in optical networks. In: OPTICAL FIBER COMMUNICATION CONFERENCE AND EXPOSITION (OFC/NFOEC), 2012 AND THE NATIONAL FIBER OPTIC ENGINEERS CONFERENCE, 2012, Los Angeles, CA. **Proceedings...** Los Angeles, CA: IEEE, 2012. p. 1–3. 49

WANG, Y.; CAO, X.; PAN, Y. A study of the routing and spectrum allocation in spectrum-sliced elastic optical path networks. In: JOINT CONFERENCE OF THE IEEE COMPUTER AND COMMUNICATIONS SOCIETIES, INFOCOM

2011, 2011, Shanghai, China. **Proceedings...** Shanghai, China: IEEE Computer and Communications Societies, 2011. p. 1503 –1511. 2, 17

WATKINS, C. J. C. H. **Learning from delayed rewards**. Tese (Doutorado) — King's College, Cambridge, UK, May 1989. 41

WRIGHT, P.; LORD, A.; VELASCO, L. The network capacity benefits of flexgrid. In: 2013 INTERNATIONAL CONFERENCE ON OPTICAL NETWORK DESIGN AND MODELING (ONDM), 2013, Brest. **Proceedings...** Brest: IEEE, 2013. p. 7–12. 3

YU, F.; WONG, V.; LEUNG, V. A new qos provisioning method for adaptive multimedia in wireless networks. **IEEE Transactions on Vehicular Technology**, v. 57, n. 3, p. 1899–1909, maio 2008. ISSN 0018-9545. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4357390>. 40

YU, Y.; ZHANG, J.; ZHAO, Y.; CAO, X.; LIN, X.; GU, W. The first single-link exact model for performance analysis of flexible grid wdm networks. In: OPTICAL FIBER COMMUNICATION CONFERENCE/NATIONAL FIBER OPTIC ENGINEERS CONFERENCE 2013, 2013, California, USA. **Proceedings...** California, USA: IEEE, 2013. p. 1 – 3. 32

ZHANG, G.; De Leenheer, M.; MOREA, A.; MUKHERJEE, B. A survey on ofdm-based elastic core optical networking. **IEEE Communications Surveys & Tutorials**, v. 15, n. 1, p. 65–87, jan. 2013. ISSN 1553-877X. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6148192>. 14

ZHANG, M.; LU, W.; ZHU, Z.; YIN, Y.; YOO, S. J. B. Planning and provisioning of elastic o-ofdm networks with fragmentation-aware routing and spectrum assignment (rsa) algorithms. In: COMMUNICATIONS AND PHOTONICS CONFERENCE (ACP), 2012 ASIA, 2012, China. **Proceedings...** China: OSA, 2012. p. 1–3. 49

APÊNDICE A - NOTAS DE IMPLEMENTAÇÃO DO MODELO ANALÍTICO

Neste apêndice descrevemos uma maneira alternativa de se representar alguns elementos do PMD a Tempo Contínuo descrito no capítulo 4. Tal representação foi criada e implementada com o intuito de ter uma maior eficiência computacional e permitir que instâncias maiores do problema fossem resolvidas. Essa eficiência se dá principalmente pela maneira de se enumerar os estados e probabilidades de transição não-nulas e guardá-las na memória. Vale ressaltar que apenas os elementos que diferem dos já descritos no capítulo 4 são considerados.

A.1 Configuração do espectro

Cada estado s é composto por uma carga c e um evento ev. Uma carga descreve a grade de *slots* como n pares, em que cada par (c_i, k_i) indica, respectivamente, a posição do primeiro *slot* alocado para a i-ésima conexão e o seu tipo. Um exemplo é mostrado na tabela A.1, na qual 0 representa um *slot* vazio e k um *slot* utilizado por uma conexão do tipo k. Neste exemplo tem-se uma grade de 5 *slots* e dois tipos de conexão ($k = 1$ e $k = 2$) que demandam, respectivamente, 1 e 2 *slots* O tamanho da banda de guarda é 1, $g = 1$.

Tabela A.1 - Exemplos de carga em uma grade de 5 slots.

Grade de *Slots*	**# de conexões**	**Pares**
(0 0 0 0 0)	0	{}
(0 2 2 0 0)	1	$\{(2,2)\}$
(1 0 1 0 1)	3	$\{(1,1),(3,1),(5,1)\}$

O conjunto de cargas C, contendo todas as cargas factíveis, é definido como

$$C = \{\ (n,(c_1,k_1),...,(c_n,k_n)) \mid n \in \mathbb{N},\ 1 \leq k_i \leq K \text{ para } 1 \leq i \leq n,\ c_1 \geq 1,\ c_n \leq N - w_{k_n},\ c_{i+1} \geq c_i + w_{k_i} + gb \text{ para } 1 \leq i \leq n\ \}.$$

A.2 Função $\phi_s(p)$

Seja a função,

$$\phi_s(p) = \begin{cases} N - p + 1 & \text{se } n = 0 \ \vee \ (n > 0 \ \wedge \ p > c_n + w_{k_n} + gb - 1) \\ c_1 - gb - p & \text{se } n > 0 \ \wedge \ p < s_1 - gb \\ c_i - gb - p & \text{se } n > 0 \ \wedge \ \exists i \in \{2, ..., N\} \\ & \text{de modo que } \ c_{i-1} + w_{k_{i-1}} + gb \leq p < c_i - gb \\ 0 & \text{caso contrário} \end{cases}$$

que retorna o número de *slots* contíguos disponíveis a partir da posição p no estado s. O conjunto de todas as posições em que uma requisição do tipo k pode ser alocada no estado s é dada por

$$\psi_s(k) = \{p \in \{1, 2, ..., N\} |\ \phi_s(p) \geq w_k\} \ .$$

A.3 Espaço de Estados

O espaço de estados S é definido por

$$\begin{aligned} S = \{ \ (c, ev) \mid c \in C, ev \in E, & \\ & \text{se } ev_t = IN \ \wedge \ ev_v = k \text{ então } \#\psi_s(k) > 0, \ 1 \leq k \leq K \\ & \text{se } ev_t = OUT \ \wedge \ ev_v = i \text{ então } n > 0, \ 1 \leq i \leq n \qquad \}, \end{aligned}$$

no qual devem haver ao menos 1 conexão em curso no sistema para um evento OUT ser considerado. Além disso, uma chegada de requisição do tipo k é considerada apenas quando existe um ou mais conjuntos de *slots* contíguos que possam acomodá-la.

A.4 Conjunto de Ações e Taxas de Transição

O sistema, por hipótese, é observado continuamente no tempo e uma decisão deve ser tomada após a ocorrência de um evento. Se o evento for uma chegada, $ev_t = IN$, a requisição pode ser bloqueada ($BLOCK$) ou aceita em qualquer posição disponível p (ACC_p). Nenhuma ação (NOA) é tomada quando o evento é um término de conexão ($ev_t = OUT$). Desse modo, o conjunto de ações para o estado s é

$$A(s) = \begin{cases} \{ACC_p, BLOCK\} & \text{se } ev_t = IN, \ \ \forall \ p \in \psi_s(k) \\ \{NOA\} & \text{se } ev_t = OUT \end{cases}$$

Quando uma ação é escolhida, o estado $s = (c, ev)$ assume uma carga pós-decisão

$c' \in C$ onde ele permanece até a chegada de um novo evento. Essa nova carga varia de acordo com as tabelas A.2 e A.3. Se uma conexão do tipo k é aceita na posição p, w_k *slots* contíguos a partir da posição p são alocados para acomodá-la; se uma conexão é terminada os *slots* relacionados ficam disponíveis; e se um bloqueio acontece a carga não é modificada.

Tabela A.2 - Carga pós-decisão para $ev = IN$

Evento	**Condição**	**Ação**	**Nova carga c'**
(IN, k)	$\#\psi_s(k) > 0$	$ACC_p,\ p < c_1$	$(n+1, (p,k), (c_1,k_1), ..., (c_n,k_n))$
		$ACC_p,\ c_i < p < c_{i+1}$	$\left(\begin{array}{c} n+1, (c_1,k_1), ..., (c_i,k_i), (p,k), \\ (c_{i+1},k_{i+1}), ..., (c_n,k_n) \end{array}\right)$
		$ACC_p,\ p > c_n$	$(n+1, (c_1,k_1), ..., (c_n,k_n), (p,k))$
		$BLOCK$	$(n, (c_1,k_1), ..., (c_n,k_n))$

Tabela A.3 - Cargas pós-decisão para $ev = OUT$

Evento	**Condição**	**Ação**	**Nova carga c'**
(OUT, i)	$i = 1$	NOA	$(n-1, (c_2,k_2), ..., (c_n,k_n))$
	$1 < i < n$		$\left(\begin{array}{c} n-1, (c_1,k_1), ..., (c_{i-1},k_{i-1}), \\ (c_{i+1},k_{i+1}), ..., (c_n,k_n) \end{array}\right)$
	$i = n$		$(n-1, (c_1,k_1), ..., (c_{n-1},k_{n-1}))$

Dada a carga pós-decisão, taxas de transição para o próximo estado s' dependem da ocorrência do próximo evento, como mostrado na Tabela A.4. Quando uma requisição do tipo k chega e ela pode ser acomodada no espectro, a taxa de transição para o próximo estado $s' = (c', (IN, k)$ é λ_k. Se uma conexão é terminada, a taxa de transição para $s' = (c', (OUT, i))$ é μ_{k_i}. Portanto, taxas de transição não-nulas entre quaisquer dois estados $s \in S$ e $s' \in S$ ao se tomar a ação $a \in A(s)$, $q_{ss'}(a)$, são representadas pela Tabela A.4.

Tabela A.4 - Taxas de Transição

Condição	**Novo Evento**	**Taxa**	**Novo Estado**
$\#\psi_s(k) > 0$	(IN, k)	λ_k	$(c', (IN, k))$
$1 \leq i \leq n$	(OUT, i)	μ_{k_i}	$(c', (OUT, i))$

APÊNDICE B - ATUALIZAÇÃO DO VETOR θ EM REGRESSÃO LINEAR ITERATIVA

Neste apêndice descrevemos as equações utilizadas na atualização do vetor θ para o *Relaxed-SMART* com regressão linear apresentado no capítulo 5. As funções de valor $Q(s, a)$ são atualizadas recursivamente por meio de um esquema de mínimos quadrados para obter o vetor θ^k a cada iteração do algoritmo. Essa atualização leva em conta uma taxa de aprendizagem α utilizada para aproximar as funções de valor. Diferentes regras de decaimento dessa taxa podem ser utilizadas.

B.1 Método Recursivo de Mínimos Quadrados para Dados Estacionários

A equação de atualização que utiliza implicitamente a regra de taxa de aprendizagem $\alpha^k = 1/k$ é dada por

$$\theta^k = \theta^{k-1} - H^k \phi^k \bar{\epsilon}^k \ , \tag{B.1}$$

em que H^k é uma matriz computada por

$$H^k = \frac{1}{\gamma^k} B^{k-1} \ . \tag{B.2}$$

O erro $\bar{\epsilon}^k$ é calculada utilizando

$$\bar{\epsilon}^k = \bar{Q}(s, a|\theta) - \bar{q}^k \ , \tag{B.3}$$

em que

$$\bar{q}^k = r(s, a, s') - g^k t(s, a, s') + \eta \max_{b \in A(s')} Q(s', b) \ . \tag{B.4}$$

B^{k-1} é uma matriz $|F|x|F|$, que é atualizada recursivamente de acordo com

$$B^k = B^{k-1} - \frac{1}{\gamma^k}(B^{k-1}\phi^k(\phi^k)^T B^{k-1}) \ , \tag{B.5}$$

e γ^k é um escalar calculado utilizando

$$\gamma^k = 1 + (\phi^k)^T B^{k-1} \phi^k \ . \tag{B.6}$$

A Equação B.1 assemelha-se a um algoritmo de gradiente estocástico, com uma diferença significativa. Ao invés de utilizar uma taxa de aprendizagem típica, tem-se uma matriz H^k que serve como uma matriz de escala.

A estratégia utilizada para iniciar os valores de B^k é utilizar $B^0 = \epsilon I$, em que I é uma matriz identidade e ϵ é uma constante pequena.

B.2 Método Recursivo de Mínimos Quadrados para Dados Não-Estacionários

O método recursivo de mínimos quadrados para dados estacionários leva em conta um peso igual em todas as observações ao longo do tempo, ao passo que pode-ser preferível colocar um peso maior em observações mais recentes. Além disso, intrinsecamente nas equações de atualização considera-se a estratégia de decaimento da taxa de aprendizagem $1/k$. Embora tal regra funcione bem para dados estacionários, ela pode não ser adequada para o aprendizado em certos casos.

Desse modo, ao invés de se minimizar erros totais, minimiza-se uma soma ponderada de erros com um fator de desconto utilizado para descontar observações mais antigas em relação às mais novas. Para aplicar esse princípio nas equações de atualização as únicas mudanças são feitas na fórmula de atualização para B^k, dada por

$$B^k = \frac{1}{\lambda}\left(B^{k-1} - \frac{1}{\gamma^k}(B^{k-1}\phi^k(\phi^k)^T B^{k-1})\right) , \tag{B.7}$$

e a nova expressão para γ^k é dada por

$$\gamma^k = \lambda + (\phi^k)^T B^{k-1} \phi^k \ . \tag{B.8}$$

Nas equações acima, λ funciona de maneira similar a uma taxa de aprendizagem, porém na direção oposta. Ao se definir $\lambda = 1$ os pesos de todas as observações tornam-se os mesmos, o que seria ótimo caso estivéssemos lidando com dados estacionários. Valores menores de λ aumentam o peso das observações mais recentes.

Ao se escolher uma regra de decaimento para a taxa de aprendizagem α^k, λ^k é

definido na iteração k usando

$$\lambda^k = \alpha^{k-1}\left(\frac{1-\alpha^k}{\alpha^k}\right). \tag{B.9}$$

Ao se substituir na equação acima a regra da taxa de aprendizagem para dados estacionários, $1/k$, obtêm-se $\lambda = 1$.