ESTRATÉGIA DE SELEÇÃO DE CANAL EM REDE DE RÁDIOS COGNITIVOS

André Chaves Mendes

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientador: José Ferreira de Rezende

Rio de Janeiro
Dezembro de 2015

ESTRATÉGIA DE SELEÇÃO DE CANAL EM REDE DE RÁDIOS COGNITIVOS

André Chaves Mendes

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. José Ferreira de Rezende, Dr.

Prof. José Marcos Silva Nogueira, D.Sc.

Prof. Carlos Alberto Vieira Campos, D.Sc.

Prof^{a}. Luci Pirmez, D.Sc.

Prof. Marcello Luiz Rodrigues de Campos, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
DEZEMBRO DE 2015

Mendes, André Chaves

Estratégia de Seleção de Canal em Rede de Rádios Cognitivos/André Chaves Mendes. – Rio de Janeiro: UFRJ/COPPE, 2015.

XV, 121 p.: il.; 29, 7cm.

Orientador: José Ferreira de Rezende

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2015.

Referências Bibliográficas: p. 112 – 121.

1. Redes Sem fio. 2. Rádio Cognitivo. 3. Rede de Rádios Cognitivos. 4. Acesso Oportunístico ao Espectro. 5. Exploração do Espectro. 6. Seleção de Canal. I. Rezende, José Ferreira de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Aos meus pais, Alberto (in memoriam) e Elcy.

Agradecimentos

Primeiramente, gostaria de agradecer a Deus, que tudo permitiu, doando-me saúde e os meios necessários para buscar alcançar os meus objetivos.

Aos meus pais, Alberto (*in memoriam*) e Elcy; e a minha irmã, Eliane, por toda dedicação e esforço ao longo da vida, que me permitiram encontrar meu caminho e alcançar mais este objetivo.

A minha família, pelo amor e carinho, e por suportarem minhas ausências, totalmente empregadas na consecução deste trabalho de pesquisa.

Ao Professor José Ferreira de Rezende, pela amizade e realismo com que me orientou, desde o mestrado até aqui, sabendo compreender minhas dificuldades, inclusive pessoais, e ajudando-me a superá-las.

Obrigado também aos amigos do LAND, professores, alunos e a nossa prestimosa Carolina, pelas discussões sobre os assuntos acadêmicos e também pelos momentos de descontração, que muito facilitaram o passo além no trabalho.

A equipe da secretaria do PEE: Dani, Maurício, Rosa e todos os demais, pelo suporte operacional; e ao PESC, pelas instalações e equipamentos utilizados.

A todos os meus familiares e amigos, que sempre me deram apoio e força na minha caminhada, e a todos aqueles que em algum momento me ajudaram a crescer e colaboraram para o meu aprendizado.

E, finalmente, agradeço ao Instituto de Pesquisas da Marinha, pelo apoio a realização deste trabalho.

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

ESTRATÉGIA DE SELEÇÃO DE CANAL EM REDE DE RÁDIOS COGNITIVOS

André Chaves Mendes

Dezembro/2015

Orientador: José Ferreira de Rezende

Programa: Engenharia Elétrica

O rádio cognitivo têm o potencial para solucionar o problema da utilização ineficiente do espectro de RF devido a sua capacidade de acessar de modo oportunista um canal de comunicações, sem comprometer a Qualidade de Serviço, durante os períodos de inatividade dos usuários licenciados. Na realidade, para a realização eficiente do acesso dinâmico ao espectro, os processos de investigação (*spectrum exploration*) e exploração (*spectrum exploitation*) do espectro de RF são mandatórios. No primeiro processo, o rádio cognitivo precisa determinar que o canal escolhido para sua comunicação esteja livre de usuários licenciados o mais rapidamente possível, enquanto que o segundo processo está relacionado com a eficiência do mecanismo empregado por ele na descoberta e utilização do canal livre.

Em um cenário de múltiplos canais e de rádios cognitivos dotados de um único transceptor, apenas um canal pode ser sensoreado por vez para detectar possíveis "oportunidades" de uso (também chamadas *white spaces*) e, o mais rapidamente possível, utilizar-se o canal.

Assim, nesta tese, nos concentramos no processo de exploração do espectro de RF (*spectrum exploitation*) e propomos um mecanismo que estabelece uma ordem dinâmica de sensoreamento de canais para os rádios cognitivos, que considera na sua análise a possibilidade de estacionar em um canal e utilizá-lo, visando maximizar os ganhos de uma métrica de interesse, ou continuar a busca, mesmo se o canal estiver livre de usuários licenciados. Além disso, nossa proposta não exige um conhecimento a priori das capacidades médias e/ou das probabilidades de disponibilidade de cada canal, sendo esta considerada um indicador da atividade do primário.

Em seguida, nossa solução é avaliada através de simulações e comparada com outras propostas da literatura.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

CHANNEL SELECTION STRATEGY IN COGNITIVE RADIO NETWORKS

André Chaves Mendes

December/2015

Advisor: José Ferreira de Rezende

Department: Electrical Engineering

The cognitive radio has the potential to solve the problem of inefficient use of RF spectrum due to their ability to access opportunistically a communication channel, without compromising the Quality of Service during periods of inactivity of licensed users. In fact, for the efficient conduct of the dynamic spectrum access, the processes of spectrum exploration and spectrum exploitation are mandatory. In the first process, the cognitive radio must determine that the chosen channel is free of licensed users as soon as possible, while the second is related to the efficiency of the mechanism employed by it to find and use a free channel.

In a scenario of multiple channels and cognitive radios equipped with a single transceiver, only one channel can be sensed at a time to detect the channel's "opportunity" (also called *white space*) and, as fast as possible, effectively use the channel.

Thus, this thesis focus on RF spectrum exploitation process and propose a mechanism that establishes a dynamic channel sensing order in a multi-channel and multi-user cognitive radio network, which takes into account the possibility to stop at a channel and utilize it, to maximize the gains of a metric of interest, or continue search, even if the channel is free of licensed users. Furthermore, our proposal does not require a priori knowledge of the mean channel capacities and/or probabilities of availability for each channel, this being considered a licensed user activity indicator.

Then, this solution is evaluated by simulations and compared with other proposals in the literature.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

C_{INST}	*Capacidade Instantânea do Canal*, p. 42, 57, 58, 82
$\overline{C}$	*Capacidade Média de Canal*, p. 42, 57, 82
$\overline{C}_{MAX}$	*Capacidade Média Máxima do Canal*, p. 42, 43, 57, 58
ADC	*Analog to Digital Converter*, p. 8
ANATEL	Agência Nacional de Telecomunicações, p. 1
AODV	*Ad hoc On-Demand Distance Vector*, p. 21
CCC	*Common Control Channel, p. 25*
DAC	*Digital to Analog Converter, p. 8*
DCS	*Dynamic Channel Selection, p. 2, 29*
DTN	*Disruption/Delay-Tolerant Network, p. 11*
FCC	*Federal Communications Commission, p. 1, 108*
FHC	*Fator de Homogeneidade dos Canais, p. 42, 57, 58, 82*
FU	*Fator de Utilização do Canal pelo Primário, p. 43, 60*
FVA	*Fator de Variabilidade do Ambiente, p. 42, 57, 58, 82*
IEEE	*Institute of Electrical and Electronics Engineers, p. 1*
LTE	*Long-Term Evolution GSM, p. 22*
MAC	*Media Access Control, p. 108*
MDP	*Markovian Decision Process, p. 13*
PER	*Packet Error Rate, p. 20*
POMDP	*Partialy Observable Markovian Decision Process, p. 23*

QoS	*Quality of Service, p. 2, 3*
RF	*Radiofrequência, p. 1, 2, 6, 23, 50, 52–54, 61, 108*
SC	*Small Cells in 5G Heterogeneous Networks, p. 25*
SDR	*Software Defined Radio, p. 2, 8*
SNR	*Signal to Noise Ratio, p. 34, 36, 48, 52, 109*
Tcl	*Tool Command Language, p. 40, 55*
USRP	*Universal Software Radio Peripheral, p. 21*

Capítulo 1

Introdução

Com o avanço das tecnologias que utilizam transmissões sem fio, o espectro de radiofrequências (RF) vem se tornando um recurso escasso. Uma das razões que contribuem para a atual escassez são as políticas para o controle do acesso ao espectro que são adotadas pelas agências reguladoras, como por exemplo, a FCC (*Federal Communications Commission*) [1], nos Estados Unidos, e a ANATEL [1] (Agência Nacional de Telecomunicações) [2], no Brasil. Nos modelos adotados atualmente, as bandas do espectro de RF podem ser classificadas conforme o uso, basicamente, em dois tipos: licenciado e não-licenciado.

As bandas de frequências do tipo *licenciado* somente podem ser acessadas por dispositivos de determinadas tecnologias e/ou que detêm uma licença de operação para utilizar esse recurso em uma dada região geográfica. Entretanto, estudos realizados por agências reguladoras e universidades constataram que a alocação estática do espectro praticada atualmente favorece, na prática, uma subutilização das bandas de frequência, mesmo em áreas urbanas [3]. Uma das causas deste problema são as diversas faixas licenciadas alocadas para dispositivos que utilizam tecnologias legadas, que são pouco eficientes na utilização desse recurso ou até mesmo que já caíram em desuso na maioria das regiões.

Já as bandas de frequências do tipo *não-licenciado*, que representam pequenas fatias do espectro de RF, podem ser utilizadas por qualquer dispositivo, desde que certos limites de potência de transmissão sejam respeitados. Este modelo simplificado atraiu o desenvolvimento de diversas tecnologias que atualmente possuem ampla utilização em diferentes aplicações, como por exemplo, as redes do padrão IEEE 802.11 [4]. Justamente por isso, essas poucas bandas do tipo não-licenciado apresentam altos índices de utilização, limitantes para o desempenho dos dispositivos que delas se utilizam.

[1] Entidade integrante da Administração Pública Federal indireta, submetida a regime autárquico especial e vinculada ao Ministério das Comunicações, com a função de órgão regulador das telecomunicações no Brasil, com sede no Distrito Federal.

Portanto, é possível notar que na prática existe uma má utilização do espectro de RF: as bandas do tipo licenciado permanecem pouco utilizadas por tecnologias legadas, que são ineficientes ou têm baixa utilização, enquanto as bandas do tipo não-licenciado tornam-se muito utilizadas perante a crescente demanda por este tipo de espectro [5–8].

A FCC estuda maneiras de minimizar este problema através da regulamentação do acesso dinâmico e oportunista ao espectro de RF [9]. Esse acesso dinâmico apoia-se no advento de um dispositivo de rede reconfigurável, denominado *rádio cognitivo* [5, 10], capaz de adaptar dinamicamente seus parâmetros e modos de operação às condições do ambiente onde ele se encontra, e cujo desenvolvimento foi possibilitado a partir dos avanços na área do rádio definido por *software* (*Software Defined Radio* - SDR). Esta classe de dispositivos, por sua vez, permite que suas características de operação, tais como, frequência da portadora, potência de transmissão, largura de banda, tipo de modulação, e outras baseadas em *hardware* programável sejam controladas por *software* [11].

Basicamente, o rádio cognitivo tem o potencial de solucionar o problema do uso ineficiente do espectro de RF sem comprometer a Qualidade de Serviço (QoS) dos usuários licenciados, pois ele somente acessa uma determinada faixa de frequências quando os usuários licenciados estiverem inativos [12].

Caso contrário, ele deve suspender sua operação e migrar para outra faixa de frequências disponível, evitando causar interferência prejudicial ao funcionamento dos rádios licenciados presentes na região [5, 7, 8]. Assim, faixas que estejam temporariamente disponíveis podem ser utilizadas de modo oportunista [13], permitindo, desta forma, um uso mais eficiente desse recurso. Em consequência, o sensoreamento do espectro é parte importante no funcionamento desses dispositivos.

Devido a sua característica oportunista e não-prioritária de acesso as bandas do tipo licenciado, o rádio cognitivo é usualmente denominado usuário *secundário*. Já o rádio licenciado, que possui prioridade no acesso, é denominado usuário *primário*. Ao longo do texto, os termos "secundário" e "primário" serão utilizados, respectivamente, como sinônimos para "rádio cognitivo" e "rádio licenciado".

Na realidade, para a realização eficiente do acesso dinâmico ao espectro (DCS) nas redes cognitivas, os processos de investigação (*spectrum exploration*) e exploração (*spectrum exploitation*) do espectro de RF são mandatórios [14]. O *processo de investigação* pressupõe que o secundário determine que o canal escolhido para sua comunicação esteja livre de primários o mais rapidamente possível, enquanto que o *processo de exploração* está relacionado com a eficiência do mecanismo empregado pelo secundário para descobrir e utilizar um canal livre.

Um problema interessante a ser estudado neste contexto é o impacto da falta de prioridade no acesso as bandas do tipo licenciado no desempenho das redes formadas

por rádios cognitivos, chamadas redes cognitivas [15]. Por serem dispositivos oportunistas, que realizam acesso não-prioritário a essas bandas de frequência, os rádios cognitivos são diretamente dependentes da atividade dos primários. Dependendo do conjunto de oportunidades de acesso ao espectro disponível para os rádios de uma rede cognitiva, podem existir períodos de tempo onde os enlaces entre eles tornam-se temporariamente indisponíveis. Este problema é especialmente prejudicial em cenários onde a disponibilidade de oportunidades de acesso é heterogênea e a atividade dos primários é intensa e dinâmica [16]. Nestes casos, a topologia formada pelos enlaces entre rádios cognitivos também será dinâmica, dificultando a descoberta e a manutenção de faixas de frequências para a comunicação.

Em um cenário de múltiplas faixas de frequências (ou canais) e secundários dotados de um único transceptor, apenas um canal pode ser sensoreado por vez na busca de possíveis "oportunidades" de uso (também chamadas *white spaces*) e, o mais rapidamente possível, utilizar-se o canal. Nesse cenário, a sequência de sensoreamento dos canais a ser seguida pelos secundários com maior chance de se obter um canal "bom" para utilização, no nosso caso, livre de primários e com alta capacidade, chamada *ordem de sensoreamento*, pode ter um grande impacto no desempenho da rede secundária, pois busca minimizar o tempo de busca e acesso a uma faixa livre do espectro de RF, visando tanto o aperfeiçoamento da utilização desse recurso finito quanto a manutenção da Qualidade de Serviço (QoS) para os primários.

Embora a teoria da parada ótima [17] aliada a um mecanismo de força bruta (ou busca exaustiva) possam resolver o problema, há uma forte dependência do conhecimento prévio e preciso das estatísticas dos canais e do custo computacional dessa solução, que é elevado, crescendo ainda mais com o aumento da quantidade de usuários e canais, tornando o problema **NP**-difícil [18]. Ambos os problemas impactam diretamente na utilização dessa solução embarcada no rádio cognitivo, que assumimos dispor de poucos recursos computacionais.

Quando há conhecimento prévio e preciso das estatísticas dos canais, por exemplo, as suas capacidades médias e probabilidades de disponibilidade, a ordem de sensoreamento *intuitiva*, aquela que segue a ordem decrescente das probabilidades de disponibilidade dos canais, é reconhecidamente a ordem ótima para o cenário com somente um secundário e sem a utilização de modulação adaptativa [19]. Entretanto, em muitos cenários práticos, a probabilidade de disponibilidade dos canais não é conhecida previamente, e por isso, a ordem de sensoreamento ótima exige um enorme esforço computacional para ser obtida, que piora com o aumento da quantidade de secundários.

Uma primeira abordagem para resolver esse problema seria a observação histórica dessa estatística, porém, isso tornaria o tempo de espera para o efetivo acesso ao

canal demorado devido a variação dessa probabilidade até a estacionariedade e da análise necessária para se obter uma estatística precisa.

Outro aspecto interessante a ser analisado recai no *compromisso* entre maximizar o ganho imediato através da exploração dos canais com maiores "disponibilidades" conhecidas até o momento *ou* a investigação dos canais aparentemente sub-ótimos, que podem ter maiores ganhos futuros. Entretanto, a tarefa envolvida não é simples, uma vez que o número possível de sequências ordenadas de canais cresce exponencialmente com a quantidade de canais (e também, de secundários). Além do processo de escolha de qual canal que será sensoreado e, efetivamente, utilizado a cada vez, ser aleatório, em razão da presença ou não de um primário, mesmo seguindo uma determinada ordem de sensoreamento.

Em complemento, para um cenário genérico de múltiplos canais, em razão das características distintas de propagação multicaminho com forte dependência da frequência dos canais, um secundário pode experimentar diferentes ganhos de canal, ou seja, variações nas capacidades médias dos canais que determinam, por sua vez, variações nas taxas de transmissão obtidas, conforme a faixa de frequência selecionada [20]. E nesse caso, tanto "quantidade" de investigação de canais quanto o ganho a ser obtido são difíceis de serem quantificados deterministicamente.

Existem soluções para o estabelecimento da ordem de sensoreamento baseadas em diversas técnicas, sejam estruturadas e determinísticas ou adaptativas, que veremos com mais detalhes na pesquisa bibliográfica realizada na Seção 2.4, e dentre todas, a que entendemos ser a mais promissora em razão da sua adaptabilidade às variações dinâmicas das características dos canais (desconhecendo antecipadamente qualquer das suas estatísticas e o modelo de comportamento do primário) e, principalmente, da sua baixa complexidade computacional aliada a um desempenho elevado foi a técnica de aprendizagem por reforço (*Reinforcement Learning*), que adotamos para o desenvolvimento da nossa solução.

As técnicas de aprendizagem por reforço podem ser utilizadas para desenvolver políticas de seleção de ações para otimizar o retorno do ambiente através da formação de um mapeamento entre as recompensas e as ações. Uma vantagem particular do aprendizado por reforço é a sua utilização em ambientes onde os agentes possuem pouca ou nenhuma experiência e informação sobre as capacidades e objetivos dos demais agentes.

A posição defendida aqui é que a abordagem do problema a partir da aprendizagem por reforço pode ser utilizada como uma nova técnica de coordenação para os ambientes onde os mecanismos de coordenação atualmente disponíveis são ineficazes, pois quase todos eles dependem fortemente de conhecimento do ambiente e das informações compartilhadas entre os agentes.

Embora a comunicação de controle e coordenação entre os agentes seja frequen-

temente útil e indispensável como um auxílio para as atividade em grupo, ela não garante por si um comportamento coordenado [21], pode ser demorada, e pode prejudicar outra atividade de resolução de problemas se não for cuidadosamente controlada [22]. Além disso, agentes excessivamente dependente dessa comunicação serão severamente afetados se a sua qualidade estiver comprometida (canais de comunicação com altas taxas de erro, informações incorretas ou deliberadamente enganosas, etc.). Em outras ocasiões, essa comunicação pode ser arriscada ou até mesmo fatal (como em algumas situações de combate, onde o adversário pode interceptar as mensagens transmitidas). Mesmo quando essa comunicação é viável e segura, é prudente utilizá-la somente quando absolutamente necessária.

Na forma independente de aprendizagem discutida aqui, cada agente aprende a otimizar o seu reforço a partir do ambiente onde os demais agentes não são explicitamente modelados, contudo a resposta do ambiente para as ações tomadas é imediata e a solução para o problema de seleção de canal possui um conjunto grande de soluções ótimas, em razão da dinâmica do ambiente envolvida no problema. Portanto, nem o conhecimento prévio sobre as características do ambiente nem um modelo explícito sobre as capacidades dos demais agentes é necessária. A limitação dessa abordagem reside na sua incapacidade para desenvolver uma coordenação eficaz quando as ações dos agentes são fortemente acopladas, a resposta do ambiente para as ações tomadas é atrasada e existe apenas uma única ou algumas poucas combinações para as ações que levam ao comportamento ótimo.

Assim, nesta tese, nos concentramos no processo de exploração do espectro de RF (*spectrum exploitation*) e investigamos a utilização da ordem de sensoreamento dos canais para o aproveitamento eficiente (e eficaz) do espectro temporariamente desocupado pelo primário, aproveitando a diferença que pode ocorrer entre as características dos canais, dentre aqueles selecionados pelo secundário. Contudo, assumimos que o comportamento do primário é desconhecido para os secundários e consideramos a utilização de canais de comunicação, preferencialmente, heterogêneos.

Finalmente, as contribuições que destacamos do nosso trabalho podem ser divididas em duas partes:

- Em *primeiro* lugar, propomos um mecanismo baseado em aprendizagem por reforço (Reinforcement Learning) de baixa complexidade ($\mathcal{O}(N)$, onde N é o número de canais disponíveis) que fornece uma ordem dinâmica de sensoreamento de canais para usuários não-licenciados (secundários), capaz de decidir se deve parar em um canal e utilizá-lo, visando maximizar os ganhos de uma métrica de interesse, ou continuar a busca, mesmo se o canal estiver livre de primários. Além disso, nossa proposta não exige um conhecimento a priori das

capacidades médias e/ou das probabilidades de disponibilidade de cada canal, sendo esta considerada um indicador da atividade do primário;

- Em *segundo* lugar, nós comparamos nossa solução do problema de exploração do espectro de RF (*spectrum exploitation*) com um conjunto de outras abordagens para o mesmo problema, obtidas da literatura. Os resultados da nossa avaliação baseada em simulação mostraram que a nossa ordem dinâmica de sensoreamento é promissora, mantendo um desempenho superior, mesmo quando há variação no nível de atividade dos primários, o que acarreta variação das oportunidades disponíveis para os secundários.

1.1 Organização da Tese

O Capítulo 1 faz uma introdução ao tema, apresentando os seus aspectos gerais.

Diversas soluções existentes, empregando a ordem de sensoreamento de canais aplicadas para rádios cognitivos, são apresentadas no Capítulo 2, onde também são apresentados os conceitos básicos que os envolvem e as considerações assumidas neste trabalho a respeito dos usuários primários, dos rádios cognitivos e de suas funcionalidades, encerrando com um detalhamento da técnica de aprendizado por reforço, utilizada nos mecanismos propostos.

O Capítulo 3 descreve a modelagem do sistema e apresenta as propostas do mecanismo de escolha da ordem dinâmica de sensoreamento de canais em um ambiente multicanal, para apenas um usuário e, em seguida, com escopo ampliado para múltiplos usuários.

O Capítulo 4 apresenta novas estratégias para o balanceamento do dilema investigação-exploração e sua implementação no simulador desenvolvido neste trabalho, bem como, realiza a avaliação comparativa da nossa proposta com outras da literatura. Em seguida, são discutidas as questões referentes a análise da convergência do nosso algoritmo multiusuário e apresentadas as recomendações para a escolha dos seus parâmetros de modo a melhorar a taxa de convergência do mecanismo a partir dos resultados numéricos, obtidos por simulação, assim como toda a metodologia empregada nas avaliações.

Por fim, o Capítulo 5 apresenta as conclusões gerais e as oportunidades de continuidade ao trabalho de pesquisa realizado nesta tese.

Capítulo 2

Rádios Cognitivos, Conceitos e Considerações Assumidas

Neste capítulo apresentamos os conceitos básicos relacionados aos rádios cognitivos e os problemas causados pela disponibilidade dinâmica de oportunidades de acesso ao espectro de frequências. Estes conceitos são importantes para a compreensão dos assuntos abordados nesta tese e também para motivar a necessidade de novas soluções para o problema de exploração do espectro de RF (*spectrum exploitation*) em redes de rádios cognitivos com oportunidades dinâmicas.

Em seguida, realizamos um detalhamento da técnica de aprendizado por reforço, utilizada nos mecanismos propostos, e apresentamos os principais trabalhos publicados nos últimos anos envolvendo a ordem de sensoreamento dos canais e rádios cognitivos.

Além disso, no final do capítulo, serão apresentadas as considerações assumidas neste trabalho a respeito dos rádios primários e secundários e, também, sobre o ambiente em que estes dispositivos operam.

2.1 Rádios Cognitivos

Os avanços tecnológicos das últimas décadas permitiram o surgimento de uma nova classe de dispositivos configuráveis, denominada rádio definido por *software* (*Software Defined Radio* - SDR). Neste tipo de dispositivo, a maior parte do tratamento dos sinais transmitidos e/ou recebidos é realizada por *software* [11, 23]. De maneira geral, os sinais recebidos e transmitidos são convertidos, respectivamente, por conversores analógico/digital (*Analog to Digital Converters* - ADC) e digital/analógico (*Digital to Analog Converters* - DAC), o que permite que sejam processados digitalmente em microprocessadores genéricos. Assim, o uso de rádios definidos por *software* permite que diversas características de operação sejam modificadas sem

a necessidade de modificação do *hardware* do rádio. Esta flexibilidade, aliada ao uso de antenas, interfaces de rádio, ADCs e DACs que operam em largas faixas de frequências, permitem que estes rádios transmitam e recebam sinais em diferentes faixas do espectro e utilizem variadas técnicas de tratamento dos sinais, além de diferentes modulações.

O rádio cognitivo por sua vez é uma tecnologia que alia inteligência computacional à flexibilidade fornecida pelo rádio definido por *software*. Na definição original apresentada por Mitola III [11, 23], o rádio cognitivo é um arcabouço ou conjunto de funcionalidades que permitem que o dispositivo consiga de forma autônoma: observar o ambiente, inferir seu contexto, descobrir as ações possíveis, gerar planos, supervisionar os serviços fornecidos ao usuário, e aprender com a experiência. Estas funcionalidades foram organizadas na forma de um ciclo cognitivo (Figura 2.1), onde, basicamente, o dispositivo observa os estímulos do mundo exterior, planeja estratégias, toma decisões, e aprende com os resultados de suas ações. Na ideia visionária de Mitola III, o rádio cognitivo é um dispositivo inteligente, que é capaz de perceber as características do ambiente em que opera e adaptar seu modo de operação de forma autônoma visando atender as demandas do usuário.

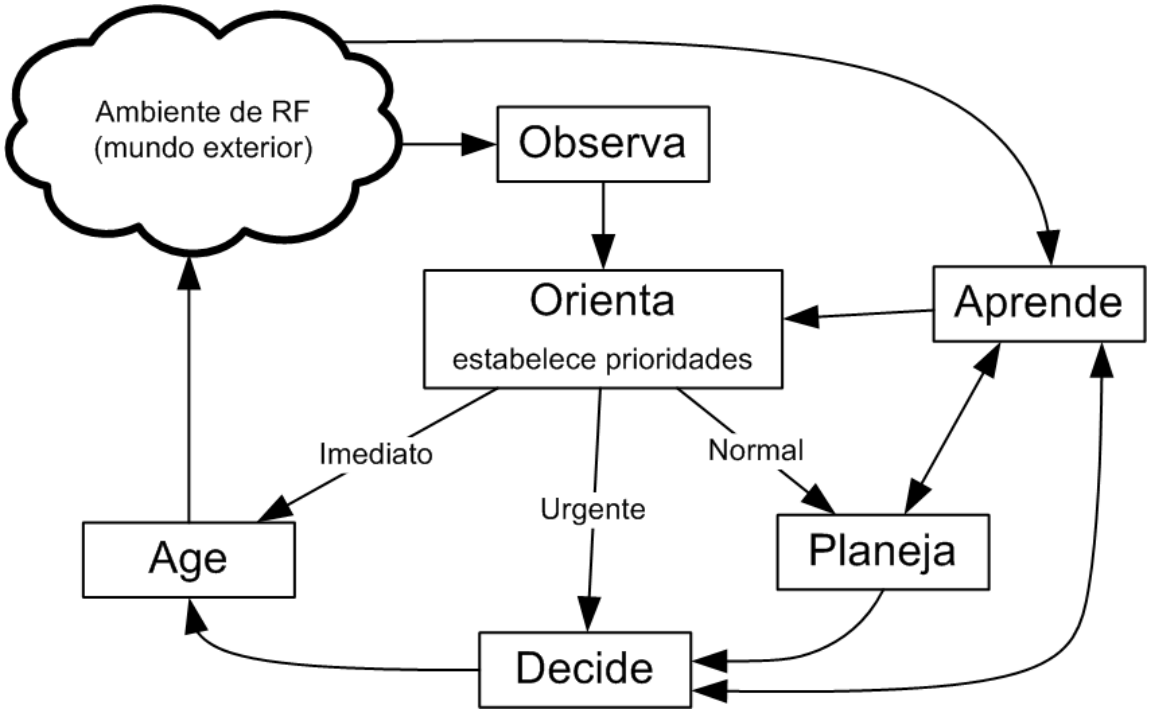


Figura 2.1: O ciclo cognitivo

Uma das vantagens intrínsecas a esta nova tecnologia é o aprimoramento da utilização do espectro de frequências. Isto porque esses dispositivos podem, por exemplo, utilizar suas habilidades cognitivas e de reconfiguração para descobrir e ocupar faixas de frequências não-utilizadas, ou parcialmente utilizadas, por outros

dispositivos de tecnologias legadas. Ou seja, os rádios cognitivos podem ser utilizados em aplicações de acesso dinâmico e oportunista ao espectro de frequências originalmente alocado para tecnologias legadas [7, 8]. De fato, como na atualidade as faixas disponíveis do espectro de frequências se tornaram um recurso escasso e a demanda por aplicações utilizando transmissões sem fio vem aumentando, o uso de rádios cognitivos é apontado como uma solução para o problema da escassez de espectro [5, 7, 8, 24].

Ao mesmo tempo em que se apresenta como uma tecnologia promissora, o rádio cognitivo também impõe uma série de novos desafios para a sua realização prática. Com isso, o assunto se tornou objeto de estudo de diversos trabalhos científicos recentes. Dentre eles, [5, 7, 8, 24] destacam-se por serem os primeiros a abordar o assunto de maneira mais ampla, apresentando os desafios existentes na área e classificando os diversos trabalhos já publicados.

Especialmente em [7] e [24], apresenta-se uma classificação interessante para os desafios em aberto na pesquisa sobre rádios cognitivos, baseada nas funcionalidades básicas de um rádio cognitivo. Segue abaixo uma breve descrição de tais funcionalidades:

- **Sensoreamento de Espectro (*Spectrum Sensing*)** - habilidade que permite ao rádio cognitivo detectar as faixas do espectro de frequências que estão livres, ou seja, as oportunidades de acesso ao espectro que podem ser utilizadas de maneira oportunista.

- **Escolha e Gerenciamento de Espectro (*Spectrum Decision and Management*)** - escolha da oportunidade de acesso ao espectro que melhor atende às necessidades dos seus usuários.

- **Mobilidade de Espectro (*Spectrum Mobility*)** - capacidade de trocar a oportunidade de acesso ao espectro sempre que um usuário primário for detectado, com o objetivo de evitar causar interferência prejudicial às suas comunicações.

- **Compartilhamento de Espectro (*Spectrum Sharing*)** - permitir o compartilhamento justo da capacidade disponível nas oportunidades de acesso ao espectro entre os rádios cognitivos.

A escolha e gerenciamento, a mobilidade e o compartilhamento espectral são desafios da área de rádios cognitivos que foram abordados neste trabalho de pesquisa.

2.2 Oportunidades de Acesso ao Meio

Um dos problemas que pode prejudicar a formação de redes cognitivas é o acesso ao meio oportunista. Os rádios cognitivos possuem prioridade secundária no acesso ao meio e, por isso, podem utilizar apenas as faixas do espectro deixadas livres pelos primários, tendo a obrigação de modificar as suas características de operação sempre que um desses dispositivos entra em funcionamento na região.

Apesar de viabilizar sua coexistência com os usuários primários, o acesso não-licenciado ao espectro dificulta a comunicação entre os rádios cognitivos. Para ser bem sucedida, a comunicação entre dois rádios cognitivos na faixa licenciada exige que ambos os dispositivos possuam ao menos uma oportunidade em comum de acesso a esta faixa. Essa oportunidade pode ser representada de diferentes maneiras. O modelo mais comum, que foi adotado neste trabalho, considera que cada canal temporariamente não utilizado pelos primários da região é uma oportunidade de acesso.

O exemplo da Figura 2.2 mostra a disponibilidade das oportunidades de acesso a faixa licenciada em função do tempo. Neste exemplo é possível perceber que os canais disponíveis para o usuário secundário mudam de acordo com o tempo. Estas mudanças fazem com que o rádio cognitivo tenha que reconfigurar frequentemente suas características de operação, as quais, ainda assim, podem não ser suficientes para evitar períodos sem oportunidades de comunicação (períodos P1, P2 e P3). Além disso, os usuários da rede secundária possuem diferentes visões dos canais disponíveis, devido ao posicionamento geográfico e às características de propagação dos sinais. Desta forma, as interrupções nas comunicações podem ser frequentes devido à falta de oportunidades de acesso em comum.

De acordo com o exemplo anterior, fica evidente que a comunicação entre os secundários está fortemente relacionada ao comportamento dos primários [16]. Assim, a comunicação entre os nós de uma rede secundária pode sofrer mudanças repentinas de qualidade e passar por frequentes períodos de indisponibilidade, os quais podem ser especialmente prejudiciais na descoberta e manutenção de canais para comunicação.

Os problemas causados pelo acesso não-licenciado ao espectro podem ser agravados devido à natureza potencialmente dinâmica da atividade dos primários e, consequentemente, da influência dinâmica destes sobre os secundários [16, 25, 26]. Em [16], os autores classificam os cenários de aplicação de redes cognitivas de acordo com o padrão de atividade dos primários em três tipos: **estáticos**, **dinâmicos** e **oportunistas**.

Em cenários do tipo **estático**, os primários permanecem por longos períodos de tempo ligados ou desligados. Desta forma, a disponibilidade das oportunidades

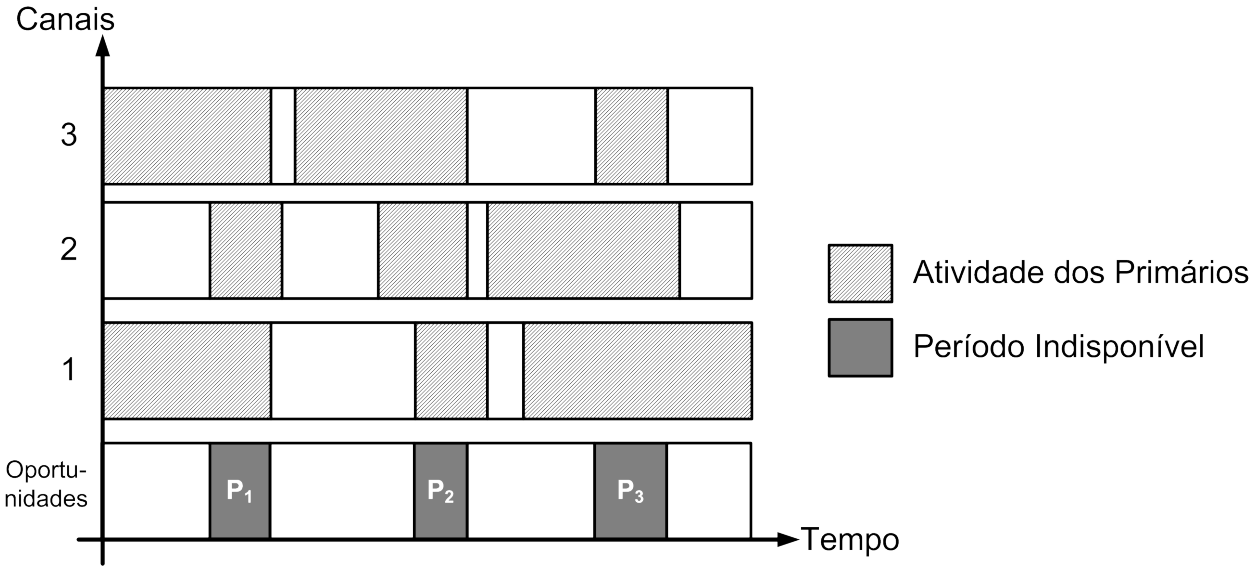


Figura 2.2: Exemplo de oportunidades de acesso a faixa licenciada

permanece inalterada por períodos de tempo maiores do que a duração média de uma comunicação na rede cognitiva secundária [16]. Para os secundários é como se os canais licenciados estivessem disponíveis ou indisponíveis por tempo indeterminado.

Nos cenários do tipo **dinâmico**, os intervalos médios entre mudanças de estado dos primários são menores que a duração média de uma comunicação na rede cognitiva secundária. Os secundários, por sua vez, têm liberdade para acessar a faixa licenciada apenas nos períodos de "silêncio" dos primários, fazendo com que a disponibilidade de oportunidades para os secundários apresente um comportamento dinâmico. Para poder utilizar de maneira eficiente as oportunidades de acesso disponíveis, os rádios secundários devem ser capazes de adaptar suas características de operação dinamicamente.

Por fim, estão os cenários do tipo **oportunista**, que são um tipo de cenário mais desafiador onde os primários possuem um padrão de atividade intenso e com alta dinamicidade. Isto faz com que a disponibilidade das oportunidades de acesso a faixa licenciada seja escassa e de curta duração. Os longos períodos de falta de conectividade entre os dispositivos da rede cognitiva tornam este tipo de cenário semelhante a uma DTN (*Disruption/Delay-Tolerant Network*), onde soluções do tipo *store-and-forward* seriam as mais indicadas [16].

O objetivo deste trabalho é buscar soluções para o problema da seleção de canais em redes cognitivas com oportunidades dinâmicas, em razão dos múltiplos canais, em cenários do tipo dinâmico. Foram desenvolvidas propostas para a seleção de canais em uma rede de rádios cognitivos em um ambiente multicanal, com um único secundário e com múltiplos secundários, detalhadas no Capítulo 3, onde não é necessário o conhecimento prévio dos momentos das variáveis de disponibilidade e capacidade

dos canais, podendo se adaptar dinamicamente às variações desses momentos.

Os resultados obtidos, discutidos nos Capítulos 3 e 4, mostraram um desempenho muito próximo do ótimo, para alguns casos. Esta característica, aliada a baixa complexidade computacional do mecanismo proposto, baseado em uma máquina de aprendizagem por reforço, detalhada na próxima seção, pode fornecer intuições para trabalhos futuros cuja plataforma destino possua restrições em termos de *hardware* e de consumo de energia.

2.3 Máquina de Aprendizagem por Reforço

O *aprendizado por reforço* (*reinforcement learning*) é uma área do aprendizado de máquina inspirada pela psicologia comportamental, onde um agente aprende através da interação com um ambiente (ou mundo), onde realiza o mapeamento de sequências de observações, chamadas *estados*, para *ações*, de modo a maximizar um valor escalar correspondente a resposta do ambiente para a ação tomada, chamado *recompensa* (ou sinal de reforço) [27].

Essa técnica difere do aprendizado supervisionado comum porque os pares (entrada, saída) corretos não são conhecidos, nem as ações sub-ótimas são explicitamente corrigidas. Além disso, existe um foco no desempenho durante a sua execução, o que envolve o ajuste do balanço entre a investigação das ações ainda não mapeadas e a exploração do conhecimento adquirido.

A tarefa do agente é aprender uma política (ou estratégia de controle) que lhe possibilite escolher o melhor conjunto de ações (regulado por uma métrica) que atinja o seu objetivo, no longo prazo. Para este efeito, o agente armazena, cumulativamente, uma recompensa para cada estado visitado ou para cada par (estado, ação). O objetivo final de um agente é o de maximizar a recompensa acumulada no longo prazo, a partir do estado atual e todos os subsequentes próximos estados até o último.

Os sistemas baseados no aprendizado por reforço possuem quatro elementos principais [27]: uma política, uma função de recompensa, uma função de valor e, opcionalmente, um modelo do ambiente de aplicação.

A *política* define o comportamento do agente, consistindo no mapeamento dos estados para as ações. A *função de recompensa* especifica como as ações escolhidas são consideradas "boas", mapeando cada par (estado,ação) para uma única recompensa numérica. A *função de valor* descreve o valor de um determinado estado através do total esperado de recompensas coletadas no longo prazo, considerando as ações tomadas a partir desse estado. Com isso, essa função pode ser utilizada para a tomada de decisões.

O *modelo do ambiente de aplicação* normalmente é desconhecido inicialmente e

o seu aprendizado ocorre durante a execução da técnica. Porém, caso exista, ele simula o comportamento do ambiente, sendo capaz de realizar a previsão da sua evolução a partir do par (estado, ação) atual. Geralmente é representado como um Processo de Decisão Markoviano (MDP) [27], onde a escolha de um novo estado depende somente do estado atual do agente e da ação que ele decidir tomar.

Um MDP é definido como uma quádrupla (S, A, T, R) caracterizada da seguinte forma: S é um conjunto de estados do ambiente; A é o conjunto de ações disponíveis no ambiente; T é uma função de transição do estado s para o novo estado s', mediante a tomada da ação a; e, R é a função de recompensa.

A solução ótima para um MDP é tomar a melhor ação possível em um estado, ou seja, aquela que coletou o máximo de recompensas possível ao longo do tempo. Para isso, existe um processo de tentativa-e-erro para maximizar as recompensas obtidas a partir do ambiente, chamado de investigação. O esquema empregado na investigação é chamado de estratégia (ou política).

Para obter uma recompensa maior, o agente prefere ações que foram testadas no passado e mostraram-se efetivas na coleta de recompensas. Porém, para descobrir tais ações, o agente necessitou testar ações que ele não havia testado antes. Assim, o agente precisa explorar o conhecimento obtido de forma a coletar melhores recompensas, mas ele também precisa investigar novas ações com a finalidade de realizar uma seleção melhor das ações futuramente. Este dilema entre investigação (*exploration*) e exploração (*exploitation*) [27] demonstra que nem a investigação nem a exploração devem ser realizadas exclusivamente. Assim, o agente deve testar uma variedade de ações e progressivamente favorecer aquelas que aparentam ser as melhores, segundo alguma métrica de interesse.

Sob o ponto de vista teórico, os métodos baseados na técnica de aprendizado por reforço realizam o aprendizado a partir de uma função de valor (por exemplo, a função Q) que pode ser utilizada para obter a ação ótima, a partir do estado corrente. Dentro desse princípio e pelo fato de existirem, frequentemente, mais de uma estatística sendo estimada através do mesmo conjunto de dados, essas estimações podem apresentar uma convergência lenta, sendo esperado que o mecanismo se comporte da mesma forma. Mesmo assim, algumas aplicações de métodos de aprendizado por reforço demonstraram possuir uma velocidade de convergência acima do esperado [27].

No *aprendizado off-policy*, o agente aprende o valor da política ótima, independentemente das suas ações, contanto que haja investigação suficiente no conjunto de ações. Entretanto, existem problemas onde o desconhecimento das ações em execução é perigoso, pois pode haver grandes penalizações (ou recompensas negativas), por exemplo, no controle do movimento de um robô sobre uma mesa, onde uma tomada de ação pode levar o robô a cair da mesa. Uma alternativa é avaliar a função

de valor no estado atual do agente de modo que o processo de seleção da ação a ser tomada possa ser melhorado de forma iterativa. Como consequência, o agente deve levar em consideração os custos associados ao processo de investigação das ações na função de valor. Assim, no *aprendizado on-policy*, o agente considera o valor da política em execução por ele próprio, incluindo as etapas de investigação.

Dentre as diversas técnicas na área de aprendizagem por reforço, o *Q-learning* [28] possui algumas vantagens [27]. Em primeiro lugar, o *Q-learning* é um algoritmo *on-line* que não necessita de qualquer conhecimento prévio do seu ambiente de aplicação. Em segundo lugar, o algoritmo do *Q-learning* adota uma abordagem simples, cuja complexidade envolvida na modelagem do problema pode ser minimizada [29], levando também a uma baixa complexidade computacional. Portanto, é bastante indicada para aplicação em sistemas com múltiplos agentes, onde cada agente possui pouco conhecimento dos demais e onde o ambiente se modifica durante o período do aprendizado. Outra vantagem importante da técnica reside na prova formal sobre a sua convergência para a ação ótima depois de um número suficiente de interações com o ambiente, conforme será visto ao final desta seção.

Desta forma, a complexidade envolvida tanto na modelagem do ambiente quanto do canal de comunicações pode ser minimizada e, pelo conjunto de vantagens, adotamos neste trabalho a técnica *Q-learning*, que apresentaremos em mais detalhes.

Q-learning

O *Q-learning* é uma técnica de aprendizado *off-policy* que busca determinar, sem conhecimento prévio do ambiente, qual ação ou decisão melhor se aplica ao agente, a cada instante. O modelo básico deste algoritmo consiste em:

- Instante de decisão, que representa o momento de execução do algoritmo, indicado por $t \in T$, $T = \{1, 2, ...\}$;
- Um conjunto de estados, próprios da modelagem de cada problema, indicados por $s \in S$;
- Um conjunto de ações, que representam as decisões que podem ser tomadas, indicadas por $a \in A$, e que levam a um novo estado;
- Uma temporalidade definida por episódio, que representa o período equivalente a progressão do agente, entre o estado inicial e o final, no espaço de estados S;
- Regras que determinam a recompensa de uma ação em um dado estado, indicada por $r_t(s, a)$;
- Regras de transição entre estados.

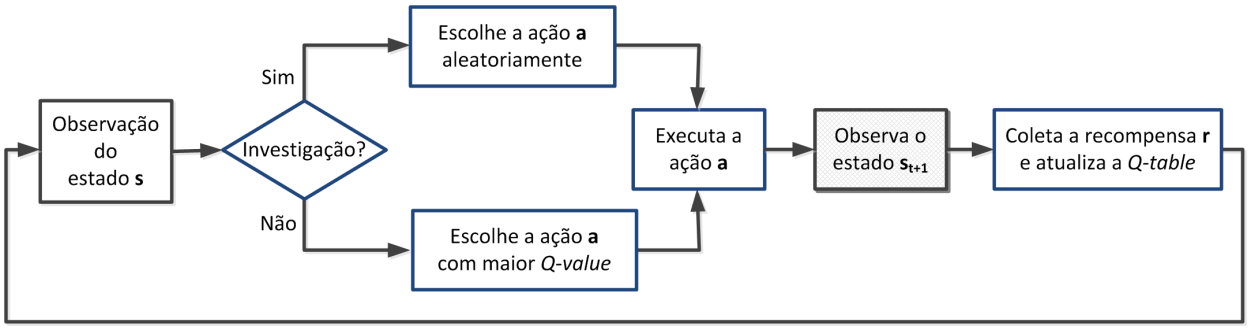


Figura 2.3: Fluxograma simplificado do *Q-learning*.

Cada agente mantém uma *Q-table*, que é uma matriz com $|S| \times |A|$ entradas, onde as linhas representam os estados e as colunas indicam as ações.

O algoritmo *Q-learning* funciona através da estimação dos valores de pares (estado, ação). Os elementos da *Q-table*, $Q(s, a)$, são chamados *Q-value's*, sendo definidos como o valor esperado da soma das recompensas futuras obtidas pelo agente ao tomar a ação a a partir do estado s, segundo uma política ótima [30].

No *Q-learning*, a partir do estado corrente s, seleciona-se uma ação a e, posteriormente, recebe-se uma recompensa r, prosseguindo para o próximo estado s_{t+1}.

A partir da recompensa obtida, o agente atualiza a respectiva entrada na *Q-table*, $Q(s, a)$, no tempo $t + 1$ conforme:

$$Q_{t+1}(s, a) = \\ (1 - \alpha) \times Q(s, a) + \alpha \times [r(s, a) + \gamma \; max_{a \in \mathbf{A}} Q(s_{t+1}, a)] \tag{2.1}$$

Na Equação 2.1, α é chamado de taxa de aprendizagem e, $0 \leq \gamma \leq 1$ é chamado de fator de desconto. Maiores valores de α indicam maior importância para a experiência recente em relação ao histórico. Valores maiores de γ indicam que o agente baseia-se mais na recompensa futura que na imediata [31].

O fluxograma do *Q-learning* é apresentado na Figura 2.3. O algoritmo inicia com a identificação do estado de entrada s. Em seguida, uma ação a é escolhida na lista das possíveis ações a partir de s, através de uma política predeterminada. Com os valores inicializados nas etapas anteriores, o *Q-value* para a ação a tomada no estado s é calculado usando a Equação 2.1 e, na sequência, armazenado na *Q-table* (em outras palavras, a experiência do agente é guardada na *Q-table*). As recompensas (e penalizações) são avaliadas por um conjunto de regras simples traduzidas na função de recompensa.

Em continuação, o próximo estado s_{t+1} é determinado após a ação selecionada a ser executada. Depois, o critério de parada para o algoritmo é verificado simul-

taneamente a determinação do próximo estado s_{t+1}. Caso s_{t+1} seja o objetivo final do algoritmo, o processo será encerrado, senão s_{t+1} se tornará o atual estado s para uma nova iteração. O processo, então, continuará até que um determinado estado ou critério de parada seja atingido.

O agente inicializa a sua *Q-table* com um valor arbitrário e, no instante de escolha da ação a ser tomada, uma estratégia é geralmente escolhida de forma que garanta uma investigação suficiente enquanto favorece as ações com "melhores" *Q-values* [27].

Uma estratégia comumente usada é a *ε-greedy* [27], que tenta fazer com que todas as ações e seus efeitos sejam experimentados igualmente [32]. No *ε-greedy* o agente utiliza a probabilidade dada por ε para decidir entre a exploração (*exploitation*) da *Q-table* ou a investigação (*exploration*) aleatória de novos estados. Na maioria dos casos, com probabilidade dada por $1 - \varepsilon$, o agente explora de forma gananciosa a ação que satisfaça $max_{a \in \mathbf{A}} Q(s, a)$, ou seja, aquela que historicamente oferece a maior recompensa. E, ocasionalmente, o agente seleciona uma ação aleatória com uma probabilidade ε pequena, que pode ser fixa ou variável.

Apesar da estratégia *ε-greedy* ser uma forma eficaz e popular para o balanceamento entre exploração e investigação, uma desvantagem é que durante a investigação a escolha pode recair igualmente entre todas as ações.

Outra estratégia, também comum, é chamada *softmax* [27], onde a probabilidade de escolha de uma determinada ação varia conforme o valor correspondente do *Q-value*. Nessa estratégia, utiliza-se uma distribuição de probabilidades para a escolha de uma ação a a partir de um estado s. Uma distribuição normalmente utilizada para essa situação é a de *Boltzmann*, ou *Gibbs*, (Equação 2.2). Essa distribuição possui o parâmetro t, chamado temperatura, que realiza o controle da quantidade de exploração que será praticada. Valores altos de temperatura tornam a escolha das ações quase equiprovável. Valores baixos, pelo contrário, causam uma grande diferença na probabilidade de escolha das ações. No limite, com $t \to 0$, a estratégia *softmax* funciona como se fosse a *ε-greedy*.

$$\Pr(a_i) = \frac{e^{\frac{Q(s,a_i)}{t}}}{\sum_a e^{\frac{Q(s,a)}{t}}} \tag{2.2}$$

Em comum, ambas as estratégias possuem apenas um parâmetro que deve ser configurado. Entretanto, se entre elas, *ε-greedy* ou *softmax*, existe uma que é melhor na maior parte dos cenários, isso não é claro [27], pois existe dependência das tarefas para a solução do problema e de fatores subjetivos da avaliação. Na prática, ambas as estratégias possuem vantagens e desvantagens, conforme descrito em [27]. Na literatura, ambas as estratégias também tem sido utilizadas para a modelagem do processo de seleção/ação existente no cérebro humano, onde o *softmax* demonstrou ser o mais representativo e possuidor de uma maior acurácia [33, 34].

Sob o ponto de vista do *Q-learning*, a aprendizagem do modelo do problema não é necessária para o propósito principal de aprender quais ações precisam ser realizadas visando um objetivo definido. Assim, o *Q-learning* é independente de modelo, sendo considerado um verdadeiro aprendizado primitivo [28]. Portanto, nessa técnica interessa apenas ganhar conhecimento sobre a recompensa que pode ser coletada em um determinado par (estado, ação), deixando de lado a quantidade de possíveis ações a partir de um determinado estado s, no espaço de estados, e evitando também o problema do armazenamento desse mapeamento, que pode ser grande e complexo.

Desta forma, o agente pode evoluir de forma adaptativa em um problema sem conhecer o seu modelo, realizando a seleção das ações que renderam as maiores recompensas em detrimento daquelas que tiveram uma recompensa menor.

Convergência do Q-learning, usuário único [30]

Definindo $n^i(s,a)$ como o índice da $i-$ésima vez que a ação a é experimentada no estado s e r_t como a recompensa a ser coletada com a tomada dessa mesma ação a nesse estado s.

E uma vez que $|r_t| \leq \Re$, $0 \leq \alpha < 1$:

$$\sum_{i=1}^{\infty} \alpha_{n^i(s,a)} = \infty, \sum_{i=1}^{\infty} {\alpha_{n^i(s,a)}}^2 < \infty, \forall s, a$$

Pode-se afirmar que com probabilidade 1, $n \to \infty, \forall s, a$, a função de valor Q converge para o seu ótimo:

$$Q_{t+1}(s,a) \to Q^*(s,a)$$

O teorema vale para γ igual a 1 e também para o caso onde são atualizados mais de um *Q-value* por vez.

Assim, o algoritmo do *Q-learning converge* para os *Q-values* ótimos com probabilidade 1, se e somente se: o ambiente é estacionário e Markoviano; uma tabela de armazenamento é utilizada para armazenar os *Q-values*; nenhum par estado-ação é negligenciado (com o tempo tendendo ao infinito) e a taxa de aprendizado é reduzida apropriadamente com o tempo.

2.4 Estratégias de Seleção de Canal

Em um cenário multicanal e de múltiplos secundários, onde cada um deles é dotado de um único transceptor, devido à restrições de *hardware*, apenas um canal pode ser sensoreado por vez para detectar possíveis oportunidades de uso. Nesse cenário, a

chamada *ordem de sensoreamento* dos canais, que é a sequência de sensoreamento dos canais a ser seguida pelos secundários com maior chance de se obter um canal "bom" para utilização (no nosso caso, livre de primários e com alta capacidade), pode ter um grande impacto no desempenho da rede secundária.

Assim, a busca pela ordem de sensoreamento ótima é um problema de grande importância e interesse que vem se intensificando desde 2006, quando surgiram os primeiros trabalhos com foco neste assunto específico, inicialmente, para apenas um (ou um par de) secundário (s) funcionando isoladamente. Nos trabalhos em [19, 35–38] a regra de parada ótima (*optimal stopping*) [17] é utilizada para encontrar a melhor sequência de sensoreamento.

Nesses trabalhos, a modelagem do sistema utilizada divide o tempo em *slots*, e em cada *slot* de tempo, os canais são sensoreados seguindo uma determinada sequência até que um canal livre seja encontrado, não sendo permitido o retorno a um canal sensoreado previamente (*recall*). O restante do tempo do *slot* é, então, utilizado para a transmissão. Pelo conhecimento antecipado das taxas alcançáveis e das probabilidades de disponibilidade de cada canal (indicativa da atividade dos primários), é possível se determinar a recompensa esperada no uso de uma sequência de canais em termos da taxa de transmissão efetiva, ou seja, do produto da taxa alcançável e da efetividade de uso do *slot*. Assim, a sequência ótima pode ser encontrada calculando-se a recompensa esperada de cada uma das sequências possíveis e escolhendo-se a de maior recompensa. No entanto, a complexidade temporal desse algoritmo por força bruta é de $\mathcal{O}(N.N!)$, onde N é a quantidade de canais, considerando-se que o cálculo da recompensa esperada para cada sequência é de $\mathcal{O}(1)$.

Com a finalidade de diminuir a complexidade dessa busca pela ordem de sensoreamento ótima, os autores em [19, 37] fornecem soluções sub-ótimas. Como desvantagem, esses trabalhos têm a deficiência de precisarem do conhecimento antecipado das taxas alcançáveis e/ou das probabilidades de disponibilidade de cada canal. E, por apresentarem uma alta complexidade computacional com o aumento do número de canais, não é possível embarcá-los nos rádios cognitivos com facilidade.

Em [19] é proposto o uso da programação dinâmica para reduzir a complexidade da busca pela ordem de sensoreamento ótima, encontrando uma solução sub-ótima com complexidade de tempo igual a $\mathcal{O}(N.2^{N-1})$, para canais com probabilidade de disponibilidade distintos. Seguindo a mesma linha, os autores em [37] fornecem também uma solução sub-ótima, baseada em árvore de decisão com uma complexidade de $\mathcal{O}(N^3)$. Essas duas soluções são comparadas nesse último trabalho, além das sequências aleatória e em ordem decrescente de disponibilidade, denominada como "sequência intuitiva" em [19], variando-se o grau de atividade dos primários.

O trabalho em [39] faz uso da regra de parada tradicional (*traditional stopping*

rule) para encontrar a melhor sequência de sensoreamento em um par de secundários, assumindo que as estatísticas dos canais são conhecidas. Quando expandida para o caso de múltiplos secundários, a regra de parada tradicional não leva a sequência ótima [39].

Em [38], as consequências dos erros de sensoreamento estão incluídas na avaliação da proposta de um mecanismo, baseado na técnica *Q-learning*, que fornece a ordem de sensoreamento dos canais para apenas um secundário, comparativamente com outras sequências de ordenação simples.

No trabalho descrito em [40], uma proposta para o problema de seleção de canal para único secundário, baseada em aprendizagem por reforço, fornece a sequência de sensoreamento a ser seguida pelo secundário visando reduzir o tempo de sensoreamento dos canais e, consequentemente, aumentando o tempo destinado à comunicação (e, indiretamente, a vazão). A proposta é avaliada em comparação com a sequência aleatória de canais, obtendo bons resultados.

Embora existam propostas empregando variadas técnicas para a solução do problema de seleção dinâmica de canal envolvendo múltiplos secundários, podemos reuni-las em dois grupos principais: aqueles com solução *não-coordenada* ou *coordenada*, conforme a existência ou não de troca de informações entre os secundários.

2.4.1 Não-coordenada

Aprendizado por Reforço

As técnicas de aprendizagem por reforço podem ser utilizadas para desenvolver políticas de seleção de ações para otimizar o retorno do ambiente através da formação de um mapeamento entre as recompensas e as ações. Uma vantagem particular do aprendizado por reforço é a sua utilização em ambientes onde os agentes possuem pouca ou nenhuma experiência e informação sobre as capacidades e objetivos dos demais agentes.

A posição defendida aqui é que a abordagem do problema a partir da aprendizagem por reforço pode ser utilizada como uma nova técnica de coordenação para os ambientes onde os mecanismos de coordenação atualmente disponíveis são ineficazes, pois quase todos eles, atualmente, dependem fortemente de conhecimento do ambiente e das informações compartilhadas entre os agentes.

Embora a comunicação de controle e coordenação entre os agentes seja frequentemente útil e indispensável como um auxílio para as atividade em grupo, ela não garante por si um comportamento coordenado [21], pode ser demorada, e pode prejudicar outra atividade de resolução de problemas se não for cuidadosamente controlada [22]. Além disso, agentes excessivamente dependente dessa comunicação serão severamente afetados se a sua qualidade estiver comprometida (canais de

comunicação com altas taxas de erro, informações incorretas ou deliberadamente enganosas, etc.). Em outras ocasiões, essa comunicação pode ser arriscada ou até mesmo fatal (como em algumas situações de combate, onde o adversário pode interceptar as mensagens transmitidas). Mesmo quando essa comunicação é viável e segura, é prudente utilizá-la somente quando absolutamente necessária.

Na forma independente de aprendizagem discutida aqui, cada agente aprende a otimizar o seu reforço a partir do ambiente onde os demais agentes não são explicitamente modelados, contudo a resposta do ambiente para as ações tomadas é imediata e a solução para o problema de seleção de canal possui um conjunto grande de soluções ótimas, em razão da dinâmica do ambiente envolvida no problema. Portanto, nem o conhecimento prévio sobre as características do ambiente nem um modelo explícito sobre as capacidades dos demais agentes é necessária. A limitação dessa abordagem reside na sua incapacidade para desenvolver uma coordenação eficaz quando as ações dos agentes são fortemente acopladas, a resposta do ambiente para as ações tomadas é atrasada e existe apenas uma única ou algumas poucas combinações para as ações que levam ao comportamento ótimo.

Os algoritmos da técnica de aprendizado por reforço (vista na Seção 2.3) estão no estado-da-arte das soluções para o problema de seleção dinâmica de canal envolvendo múltiplos secundários a partir da abordagem não-coordenada [13, 31, 41–50].

Na proposta descrita em [31], a técnica *Q-learning* é utilizada em um mecanismo que realiza a seleção dinâmica de canais, otimizando o desempenho do mecanismo para as métricas: utilização do canal pelos primários e taxa de erro de pacote (*Packet Error Rate* - PER) e comparado com o mecanismo que estabelece a escolha aleatória de canais. Mais tarde, a mesma proposta é embarcada em um *GNU radio* [13], obtendo resultados satisfatórios.

Os trabalhos descritos em [41, 42] obtêm algumas evidências empíricas que descrevem as propriedades de convergência de métodos de aprendizado por reforço em sistemas multiagente considerando que em muitas aplicações práticas não é razoável assumir que as ações dos demais agentes possam ser observadas. A maioria dos agentes interagem com ambiente ao redor apoiando-se em informações de "sensores", pois sem nenhum conhecimento das ações (nem recompensas) dos demais agentes, o problema torna-se ainda mais complexo.

Os autores em [43] estudaram a classe multiagente do tipo *independent learners*, incluindo a convergência do mecanismo, em ambiente determinístico para alguns cenários. Em trabalhos futuros [44, 45], a mesma classe é estudada em um ambiente probabilístico.

A técnica de aprendizagem por reforço é utilizada em um mecanismo para controle do tráfego gerado por uma fonte de vídeo em uma rede cognitiva com múltiplos saltos, no trabalho descrito em [48]. O mecanismo seleciona os canais para utiliza-

ção nos enlaces formados na rede entre 3 secundários e, simultaneamente, evita a interferência com 1 primário, dentro do contexto próprio de rede de vigilância por vídeo. São realizados experimentos utilizando o USRP (*Universal Software Radio Peripheral*) e o *software GNU radio*, mantendo o fluxo do vídeo com qualidade.

O trabalho descrito em [49] apresenta uma proposta conjunta de seleção de canal e roteamento em redes de rádios cognitivos baseada na técnica *Q-learning*. A proposta se baseia na formação de *clusters* e na otimização das métricas: interferência nos primários e taxa de entrega de pacotes, comparando-a ao protocolo de roteamento AODV (*Ad hoc On-Demand Distance Vector*).

Em [50], os autores apresentam uma proposta para seleção de canal em redes *ad hoc* cognitivas utilizando aprendizagem por reforço, que "aprende" o comportamento do primário a partir das características do tráfego gerado por ele, e, em seguida, realiza a seleção apropriada do melhor canal para transmitir, visando reduzir a interferência com o primário.

Teoria de Jogos

As propostas geralmente baseiam-se na teoria de jogos, assim como muitos estudos se basearam nela para a modelagem das interações entre secundários [51], quando múltiplos secundários competem de forma egoísta pelas oportunidades no espectro de RF.

Nos trabalhos descritos em [52, 53], a convergência é um requisito para estabilidade; e racionalidade, definida nesse mesmo trabalho como o requisito onde um agente converge para a melhor resposta enquanto os demais agentes permanecem estacionários, é considerada um critério de adaptação. Para um algoritmo ser considerado convergente, os autores estabelecem que o agente deve convergir para uma estratégia estacionária, dado que os demais agentes utilizem um algoritmo de uma classe-alvo pré-definida de algoritmos. Embora a convergência para um equilíbrio ótimo (*Nash*) não aconteça explicitamente, é esperado que ocorra naturalmente se todos os agentes no sistema possuam racionalidade e sejam convergentes.

Um conceito importante é o de arrependimento (*regret theory*), que é definido como o requisito onde um agente obtém uma recompensa que é, no mínimo, tão boa quanto aquela obtida através de qualquer estratégia estacionária, independente da estratégia adotada pelos demais agentes [54]. Para certos tipos de problemas, os algoritmos de aprendizado que seguem esse conceito convergem para o equilíbrio ótimo (Nash) [55, 56].

No trabalho descrito em [57], a convergência da classe *independent learners* é estabelecida, enquanto aplicada ao procedimento do jogador fictício [58] em jogos competitivos. Com uma abordagem diferenciada, o trabalho em [46] propôs um algoritmo da classe *independent learners* para jogos repetitivos (onde se permite

que o jogador tenha memória de jogadas passadas) que converge para uma política garantidamente justa, mas que periodicamente se alterna entre alguns pontos de equilíbrio.

Oa autores em [59], propõem uma solução ótima para a busca da ordem de sensoreamento para um cenário multicanal com múltiplos secundários, baseada em um jogo não-cooperativo, que realiza também o controle da interferência entre secundários.

Em [60], é apresentada uma proposta para a configuração e otimização autônomas de femto-células em redes LTE heterogêneas, baseada em aprendizagem por reforço, que realiza a redução da interferência tanto com os primários como com as demais femto-células, simultaneamente satisfazendo requisitos de utilização do espectro e vazão.

Bandido de N-Braços

Com uma abordagem diferenciada, alguns trabalhos buscam uma solução ótima para o problema de seleção de canal em redes cognitivas se valendo da semelhança existente com o problema do bandido de n-braços [61], porém a complexidade espacial envolvida nesse problema clássico é $\mathcal{O}(N^N)$. No trabalho descrito em [62], uma estratégia ótima guiada através do índice Gittins [63] é proposta para atingir o balanceamento entre os processos de investigação e exploração do espectro de RF (*spectrum exploration-exploitation dilemma*).

No trabalho descrito em [64], os autores aplicaram formulações derivadas de outro trabalho [65] para encontrar soluções para o problema de seleção de canal em redes cognitivas.

Os autores em [66] empregaram uma abordagem descentralizada para a solução do mesmo problema, além de apresentarem uma heurística que tende para a solução ótima, quando o tempo de observação dos canais cresce muito. Entretanto, nesses trabalhos assumiu-se que dentro de um tempo fixo de observação, não seria possível sensorear mais de um canal.

E no trabalho descrito em [67] é proposta uma solução sub-ótima de complexidade de tempo igual a $\mathcal{O}(N^2 \log SLOTS)$, onde N é igual ao número de canais possíveis de utilização e $SLOTS$ é a quantidade de *slots*, para o problema de seleção de canal através do estabelecimento de uma ordem de sensoreamento.

Processo Markoviano

O trabalho em [68] estuda o problema do acesso oportunístico dos secundários em um cenário multicanal através de um processo de decisão Markoviano parcialmente observável (POMDP). Através da metodologia empregada, a probabilidade de ocupação do canal pelo primário é obtida. Com esse conhecimento, os secundários

poderiam sensorear e selecionar os canais, de forma oportunista.

Os autores em [69] demonstram que com o conhecimento preciso da atividade do primário, os secundários podem explorar melhor os canais, evitando interferir com a atividade do primário, e ainda melhorar a utilização do espectro de RF.

Diferente da nossa abordagem, nenhum desses trabalhos [68, 69] tratam do mecanismo de contenção para acesso ao meio, no caso de múltiplos secundários.

Análise Estatística

Mais recentemente, no trabalho descrito em [70], os autores propuseram uma solução ótima para o escalonamento de pacotes de dados em um sistema multicanal através de uma análise dos tempos de espera. Contudo, diferente da nossa abordagem, esse trabalho desconsidera o mecanismo de contenção para acesso ao meio, no caso de múltiplos secundários.

Outras Técnicas

A utilização da regra de parada ótima (*optimal stopping*) [17] é indicada no trabalho em [71] para atingir um ganho de desempenho na vazão de uma rede sem fio baseada no padrão IEEE 802.11, contudo, sua aplicação para redes de rádios cognitivos é ineficiente.

Ainda na abordagem não-coordenada, na solução apresentada em [36], um protocolo de camada MAC é proposto considerando que todos os canais possuem a mesma probabilidade de disponibilidade. Neste caso, o problema de ordem de sensoreamento se reduz ao uso da sequência de canais em ordem decrescente de suas taxas alcançáveis, como em [72].

A nossa abordagem, comparada com [36], considera os canais heterogêneos e se aproveita da diferença que pode ocorrer entre as características dos canais, dentre aqueles selecionados pelo secundário, para o aproveitamento eficiente (e eficaz) do espectro temporariamente desocupado pelo primário; e com [72], desconsidera a existência de variação de um mesmo canal para diferentes secundários, devido a *assumirmos* que a rede secundária ocupa uma área pequena, onde ocorrem poucos fenômenos que influenciam a propagação entre pontos geograficamente próximos.

O trabalho descrito em [73], aborda o processo de exploração do espectro (*spectrum exploitation*) e propõe uma solução ótima para o caso de um par de secundários, a partir do conhecimento a priori da probabilidade de ocupação do canal pelo primário. Além disso, é permitido o retorno a um canal previamente utilizado (*recall*) ou sem sensoreamento prévio (*guessing*). Quando comparada com [73], a nossa abordagem desconsidera a utilização do *recall* e do *guessing*.

O trabalho descrito em [74] considerou apenas o sensoreamento (e acesso) conforme a ordem crescente dos canais, assumindo limitações de *hardware* e de energia

nos rádios cognitivos, e utilizou como métrica a disponibilidade e a qualidade dos canais na sua proposta de uma estratégia de vazão eficiente, embora tenha assumido somente a existência de canais homogêneos.

Na literatura, as abordagens referentes ao processo de investigação do espectro de RF (*spectrum exploration*) para redes de rádios cognitivos vem recebendo atenção [24]. Sem o conhecimento antecipado das transmissões primárias, a detecção por energia é mostrada como ótima [75].

Variadas propostas para estratégias de seleção de canal, envolvendo o processo de exploração do espectro de RF (*spectrum exploitation*), vêm sendo apresentadas [36, 62, 64, 65, 68, 69, 71, 73, 76] visando melhorar a utilização do espectro finito e aumentar o desempenho de uma métrica de interesse.

O trabalho descrito em [77] propõe um mecanismo híbrido, que fornece a ordem de sensoreamento dos canais através de uma escolha aleatória ou pela busca baseada em programação dinâmica, conforme o valor instantâneo de uma variável aleatória comparada com um limiar. Ao final, é realizada uma avaliação comparativa com outras sequências, como: a sequência de canais em ordem decrescente das suas taxas de transmissão [72] e a sequência obtida por programação dinâmica proposta em [19]. Esses trabalhos assumem o conhecimento antecipado das probabilidades de disponibilidade dos canais (ou seja, possuem conhecimento do comportamento do primário).

Em [78], os autores apresentam uma proposta de estratégia adaptativa (e persistente) para o problema de seleção de canais em redes cognitivas, "livre de colisão" entre os secundários para o caso onde a sua quantidade é menor ou igual a de canais, e comparam essa proposta com outros mecanismos, obtendo resultados satisfatórios.

Como solução para o exploração do espectro (*spectrum exploitation*), o trabalho descrito em [79] propõe uma partição do espectro em conjuntos de canais em número equivalente ao de secundários e obtém a ordem de sensoreamento a partir da "sequência intuitiva" (ordem decrescente das probabilidades de disponibilidade dos canais) [19], a qual mitiga também as colisões entre secundários.

No trabalho descrito em [80], a sequência de sensoreamento dos canais em redes 5G heterogêneas é obtida considerando o tráfego redirecionado para as chamadas "células pequenas" (*Small Cells* - SC, ou sejam, femto-células, pico-células e micro-células.) dos usuários das macro-células, que estiverem congestionadas. A avaliação do mecanismo proposto demonstra seu desempenho superior, comparativamente aos que fornecem as sequências em ordem decrescente das: probabilidades de disponibilidade dos canais, capacidades dos canais e da métrica conjunta das probabilidades de disponibilidade e capacidades dos canais.

Em [81], é apresentada uma proposta de complexidade de tempo igual a $\mathcal{O}(\ln T)$, onde T é a quantidade de *slots*, para o problema de seleção de canal para único

secundário, que é expandida para múltiplos secundários, seguindo a teoria da estratégia míope (*myopic strategy*), onde cada secundário toma suas decisões e ações considerando apenas a sua visão da rede. Na proposta, o mecanismo otimiza a utilização dos canais e a vazão na rede, após "aprender" o comportamento do primário e descobrir os canais com maior probabilidade de disponibilidade. A avaliação do mecanismo no cenário onde apenas as probabilidades de disponibilidade variam a cada *slot*, comparativamente a uma solução ótima, baseada no problema do bandido de n-braços [61], demonstra seu bom desempenho.

Na proposta de complexidade de tempo igual a $\mathcal{O}(N \log N)$ descrita em [82], a sequência de canais obtida é em ordem decrescente da métrica conjunta da taxa de transmissão estimada e da probabilidade estimada da disponibilidade dos canais, considerando o comportamento dos primários equivalente ao modelo *ON-OFF* e o conceito de fator de utilização dos canais (também utilizado na nossa proposta (Subseção 3.3.3)). Contudo, a duração da ocupação dos canais pelos secundários segue uma distribuição de Poisson e os secundários mantém um sensoreamento periódico visando evitar interferir com os primários. Na avaliação comparativa do mecanismo contra a sequência aleatória dos canais, referente as métricas vazão e tempo de transmissão de pacotes, o desempenho é superior, porém baixo.

2.4.2 Coordenada

A principal desvantagem das propostas coordenadas recai sobre a necessidade de um ponto central para a coordenação dos secundários (canal de controle comum (CCC)) ou sobre o *overhead* envolvido na solução distribuída. Além disso, essas abordagens podem ser alvos de ataques, em razão da utilização da informação de coordenação obtida de um outro usuário [83].

No trabalho descrito em [84], os secundários se organizam em grupos baseados na similaridade entre os canais disponíveis para cada secundário. E no trabalho em [85], a coordenação entre rádios cognitivos é formulada como um problema envolvendo simultaneamente o controle de potência e a taxa de transmissão e um problema de otimização para designação dos canais.

No trabalho descrito em [86], cada secundário recomenda para os demais os canais que foram utilizados com sucesso. Alguns trabalhos comentados em [87] utilizam a abordagem baseada em métodos da negociação de espectro (*spectrum trading*).

A avaliação do consumo de energia em uma proposta para a seleção de canal baseada no *Q-learning* é realizada no trabalho descrito em [88], onde os secundários trocam informações entre eles. O mecanismo inclui as observações de utilização de canal pelos primários na análise e, na comparação com o mecanismo de seleção aleatória de canais, obtém resultados superiores.

2.5 Considerações Assumidas na Tese

Como ainda não estão definidos padrões para a utilização do rádio cognitivo, se torna importante a definição das considerações a respeito das funcionalidades e características de operação dos rádios cognitivos que foram assumidas neste trabalho.

Assim, nesta seção, são apresentadas as considerações importantes a respeito dos rádios primários e secundários e, também, sobre o ambiente em que estes dispositivos operam.

2.5.1 Dispositivos Primários e Canais Licenciados

A primeira consideração importante diz respeito ao comportamento do primário. Na prática, o secundário desconhece as características do sinal do primário, não conseguindo depreender as informações do canal a partir do conhecimento do primário ou vice-versa [75].

Assim, o secundário precisa adotar alguma forma de detecção não-coerente (como a detecção por energia, que é ótima em termos de percepção do nível de um sinal desconhecido do receptor [75]) durante o sensoreamento dos canais. Além disso, mesmo nessa condição, o secundário não dispõe de informações da sua localização em relação ao primário e, portanto, desconhece os efeitos de sombreamento e das perdas no caminho dos sinais do transmissor primário até o receptor secundário. Por isso, não é razoável supor que o secundário conheça a exata distribuição das observações sob a hipótese de ocupação do canal pelo primário, embora possa ser assumido que a distribuição das observações sob a hipótese de canal desocupado (ou livre) do primário seja conhecida [89].

Baseado nisso, diversos trabalhos na literatura modelam esse comportamento através de uma cadeia de Markov de tempo contínuo de dois estados, como na Figura 2.4. Neste modelo, os estados 1 e 0 representam, respectivamente, os períodos em que o rádio primário está "ON", ativo (canal ocupado) e "OFF", inativo (canal livre). Por ser uma cadeia de Markov, o tempo de permanência nos estados é dado por variáveis aleatórias com distribuições exponenciais.

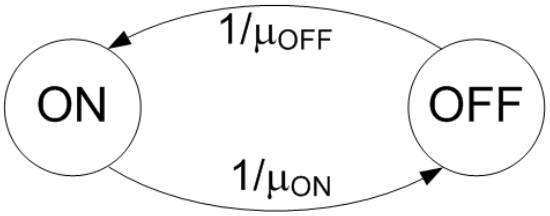


Figura 2.4: Cadeia de Markov representando a utilização dos canais licenciados (modelo ON-OFF exponencial).

Apesar de ser um modelo amplamente adotado, alguns trabalhos específicos argumentam que este tipo de modelagem do comportamento dos primários é inadequado em algumas situações [90–95]. Estes trabalhos utilizam dados coletados através de medições do espectro e buscam determinar modelos matemáticos que representem os padrões de utilização e, consequentemente, o padrão de atividade dos primários. Eles indicam que o tempo médio de permanência nos estados ocupado e livre seria mais bem representado por variáveis aleatórias com outros tipos de distribuições. Com isso, o modelo de atividade dos primários passa a ser um modelo semi-Markoviano de dois estados. Entretanto, todos estes trabalhos apresentam limitações no que diz respeito aos dados utilizados na análise, não se aplicando às tecnologias licenciadas.. Alguns deles limitam-se a dados coletados na banda ISM, que é a faixa de frequências utilizada por redes do padrão IEEE 802.11 e Bluetooth [90–92].

Em [93] e [94], os autores utilizam dados coletados em largas faixas do espectro, cobrindo diversas tecnologias licenciadas. Entretanto, para realizar medições de tal magnitude, o espectro é dividido em canais de banda estreita com tamanho fixo e é realizada uma medição em varredura. Para varrer uma faixa tão grande do espectro e obter resultados acurados, o intervalo entre medições torna-se consideravelmente grande devido ao tempo gasto na medição do nível de sinal em cada canal.

Em [93], o intervalo entre medições é da ordem de dezenas de segundos, logo os dados utilizados na modelagem podem não conseguir capturar o comportamento dinâmico das tecnologias avaliadas. Já em [94], o intervalo entre medições é menor, de aproximadamente 1 segundo. Neste caso, na maioria dos canais, o tempo de permanência nos estados ocupado e livre é bem modelado por uma distribuição exponencial. Apenas nos casos onde a utilização dos canais é muito baixa ou muito alta é que a distribuição exponencial torna-se ineficiente. Este resultado motiva o uso do modelo markoviano mais simples.

Em outro trabalho [95], voltado exclusivamente para redes celulares, utilizam-se *traces* com os instantes de início e término das chamadas de voz dos usuários que foram coletados diretamente com os provedores do serviço móvel celular. Os autores argumentam que estes dados são mais representativos, pois indicam os momentos em que, de fato, os usuários primários estão se comunicando. Entretanto, não se apresenta um modelo específico para representar os padrões de utilização dos canais ao longo do tempo. Os modelos apresentados visam representar o comportamento do sistema como um todo, modelando a dinâmica da chegada e saída de usuários. Logo, a adaptação destes modelos para cenários onde se deseja representar ocupação dos canais não é trivial.

Devido a estas limitações e especificidades dos modelos mais realistas presentes na literatura, este trabalho *adota* o modelo Markoviano mais simples, onde os tempos médios de permanência do primário nos estados ativo e inativo são dados,

respectivamente, por variáveis aleatórias com distribuição exponencial de médias iguais a μ_{ON} e μ_{OFF}.

Portanto, não é feita nenhuma consideração a respeito da tecnologia utilizada pelos rádios primários. Consideramos apenas que o rádio primário é um rádio genérico, que utiliza um dos canais de uma faixa licenciada de N canais seguindo um padrão dinâmico, onde pode haver transmissão em qualquer momento. Os canais da faixa licenciada são multiplexados em frequência, sendo representados por faixas de espectro de mesma largura de banda B, como na Figura 2.5. Este modelo foi capaz de cumprir o objetivo de gerar cenários com disponibilidade dinâmica dos canais e, também, apresentou a vantagem de ser fácil de parametrizar. Além disso, como mencionado anteriormente, de acordo com os resultados em [94] ele é representativo em vários cenários.

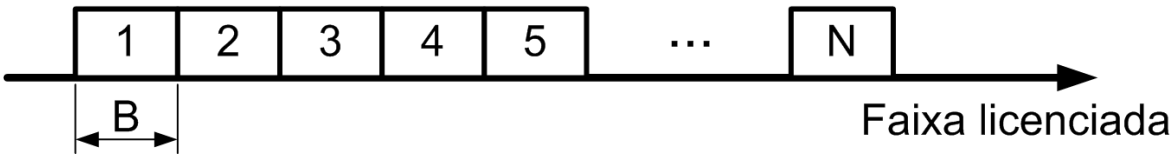


Figura 2.5: Modelo de faixa licenciada utilizado no trabalho

2.5.2 Detecção das Oportunidades de Acesso

Outra funcionalidade importante de um rádio cognitivo é a descoberta das oportunidades de acesso ao meio. Os rádios cognitivos podem descobrir as oportunidades através de consultas a bases de dados de uso do espectro, baseadas nas suas localizações geográficas, ou através de técnicas de sensoreamento do espectro executadas localmente [96].

Em cenários onde a disponibilidade das oportunidades de acesso ao espectro varia dinamicamente é improvável que uma solução do tipo base de dados seja eficiente. Isso porque, quando a disponibilidade dos canais varia em curtas escalas de tempo, seria muito custoso construir e manter a base de dados atualizada. Nestes cenários dinâmicos, o uso de técnicas de sensoreamento seria o mais indicado para poder determinar as oportunidades disponíveis a cada instante. Entretanto, é notório que as técnicas de sensoreamento existentes atualmente apresentam limitações no que diz respeito a eficiência na detecção da presença de rádios primários, e, além disso, também impõem uma sobrecarga aos dispositivos secundários que precisam reservar parte do tempo para realizar medições dos canais licenciados.

Apesar de ser um problema interessante e muito estudado na literatura [96], a detecção eficiente dos primários e das oportunidades de acesso ao espectro é um

problema que está fora do escopo desta tese por ser ortogonal aos problemas de seleção de canal aqui abordados. Assim, *assumimos* que os rádios secundários conseguem obter informação sobre a disponibilidade momentânea dos canais em escalas de tempo uma ordem de grandeza menor que aquelas das durações médias dos períodos ocupados e livres dos canais. Desta forma, é possível ter uma visão mais realista da ocupação instantânea dos canais onde a tarefa de sensoreamento representa uma sobrecarga para o funcionamento do rádio cognitivo.

Nesta tese, *assumimos* que nossa detecção de primário é precisa e isenta de erros, $\Pr_{MISDETECTION} = 0.0$ e $\Pr_{FALSEALARM} = 0.0$, pois entendemos que o foco do trabalho não é direcionado para o problema de detecção do primário, preferindo deixar essa análise para os trabalhos futuros.

Outro aspecto importante relaciona-se às características dos canais propriamente ditos. Alguns trabalhos afetos a seleção dinâmica de canais (*Dynamic Channel Selection - DCS*) em redes de rádios cognitivos assumiram os canais de comunicação como homogêneos [97–99]. Isto significa que todos os canais disponíveis partilham propriedades físicas semelhantes, tais como alcance de transmissão, qualidade de canal, atenuação e influência da interferência de radiofrequência.

Na prática, os canais suportados por um rádio cognitivo podem ser localizados em faixas de frequência separadas e diferentes canais podem experimentar significativa diferença nas suas características de propagação [20]. Além disso, por causa da atenuação e efeitos de multicaminho, um canal com maior alcance de transmissão pode não cobrir toda a área coberta por outro com um alcance de transmissão mais curto. Como tais propriedades do canal variam com a frequência da portadora utilizada e com as condições do canal variando no tempo, o pressuposto de homogeneidade do canal pode não ser apropriado.

Por essa razão, *avaliamos* nossa proposta considerando dois conjuntos distintos de canais, sendo um deles homogêneo e o outro heterogêneo.

2.6 Conclusões

No início deste capítulo são mostrados os conceitos básicos relacionados aos rádios cognitivos e os problemas causados pela disponibilidade dinâmica de oportunidades de acesso ao espectro de frequências. Estes conceitos são importantes para a compreensão dos assuntos abordados nesta tese e também para motivar a necessidade de novas soluções para o problema de exploração do espectro de RF (*spectrum exploitation*) em redes de rádios cognitivos com oportunidades dinâmicas.

Em seguida, é realizado um detalhamento da técnica de aprendizado por reforço, utilizada nos mecanismos propostos, e são discutidos os principais trabalhos publicados nos últimos anos envolvendo a ordem de sensoreamento dos canais e rádios

cognitivos.

E finalmente, no final do capítulo, são mostradas as considerações assumidas neste trabalho a respeito dos rádios primários e secundários e, também, sobre o ambiente em que estes dispositivos operam.

A medida que evoluem as plataformas de *hardware* capazes de suportar as funcionalidades (*software*) presentes no rádio cognitivo, é esperado que sejam lançados dispositivos de comunicação seguindo essa tecnologia, que a consolidarão, certamente, como a evolução do antigo conceito de "rádio".

Capítulo 3

Propostas

Este capítulo apresenta os detalhes das duas propostas que compõem esta tese. Inicialmente, na Seção 3.2, são discutidos os conceitos da proposta do mecanismo para a busca dinâmica da ordem de sensoreamento ótima de canais de comunicação para o caso de um único usuário secundário. Em seguida, na Seção 3.3, essa proposta é ampliada para o caso de múltiplos usuários secundários, incluindo uma nova estratégia para o balanceamento do dilema investigação-exploração existente.

Antes, realizaremos a descrição da modelagem do sistema utilizada pelos mecanismos propostos.

3.1 Modelagem do Sistema

Nesta seção, descrevemos o modelagem do sistema utilizada para o desenvolvimento e implementação das nossas duas propostas baseadas em uma máquina de aprendizagem por reforço (*reinforcement learning*) [27] (Seção 2.3). Esta modelagem permite determinar a sequência de sensoreamento ótima através da aplicação da teoria da parada ótima (*optimal stopping*) [17]. Com isso, o objetivo é fornecer uma decisão sobre o momento para finalizar o sensoreamento de novos canais de tal forma que a recompensa obtida na escolha de um canal seja maximizada.

Inicialmente faremos uma breve introdução, recapitulando o que foi comentado na Seção 2.2 e que forma o contexto necessário para descrevermos o modelo do sistema que adotamos.

O exemplo da Figura 2.2 mostra a disponibilidade das oportunidades de acesso a faixa licenciada em função do tempo. Neste exemplo é possível perceber que os canais disponíveis para o usuário secundário mudam de acordo com o tempo. Estas mudanças fazem com que o rádio cognitivo tenha que reconfigurar frequentemente suas características de operação, as quais ainda assim podem não ser suficientes para evitar períodos sem oportunidades de comunicação (períodos P1, P2 e P3). Além disso, os usuários da rede secundária possuem diferentes visões dos canais

disponíveis, devido ao posicionamento geográfico e as características de propagação dos sinais. Desta forma, as interrupções nas comunicações podem ser frequentes devido a falta de oportunidades de acesso em comum.

De acordo com o exemplo anterior, fica evidente que a comunicação entre os secundários está fortemente relacionada ao comportamento dos primários [16]. Assim, a comunicação entre os nós de uma rede secundária pode sofrer mudanças repentinas de qualidade e passar por frequentes períodos de indisponibilidade, os quais podem ser especialmente prejudiciais na descoberta e manutenção de canais para comunicação. Os problemas causados pelo acesso não-licenciado ao espectro podem ser agravados devido a natureza potencialmente dinâmica da atividade dos primários e, consequentemente, da influência dinâmica destes sobre os secundários [16, 25, 26].

Nos cenários do tipo **dinâmico** considerados no nosso problema, os intervalos médios entre mudanças de estado dos primários são menores que a duração média de uma comunicação na rede cognitiva secundária. Os secundários, por sua vez, têm liberdade para acessar a faixa licenciada apenas nos períodos de "silêncio" dos primários, fazendo com que a disponibilidade de oportunidades para os secundários apresente um comportamento dinâmico. Para poder utilizar de maneira eficiente as oportunidades de acesso disponíveis, os rádios secundários devem ser capazes de adaptar suas características de operação dinamicamente.

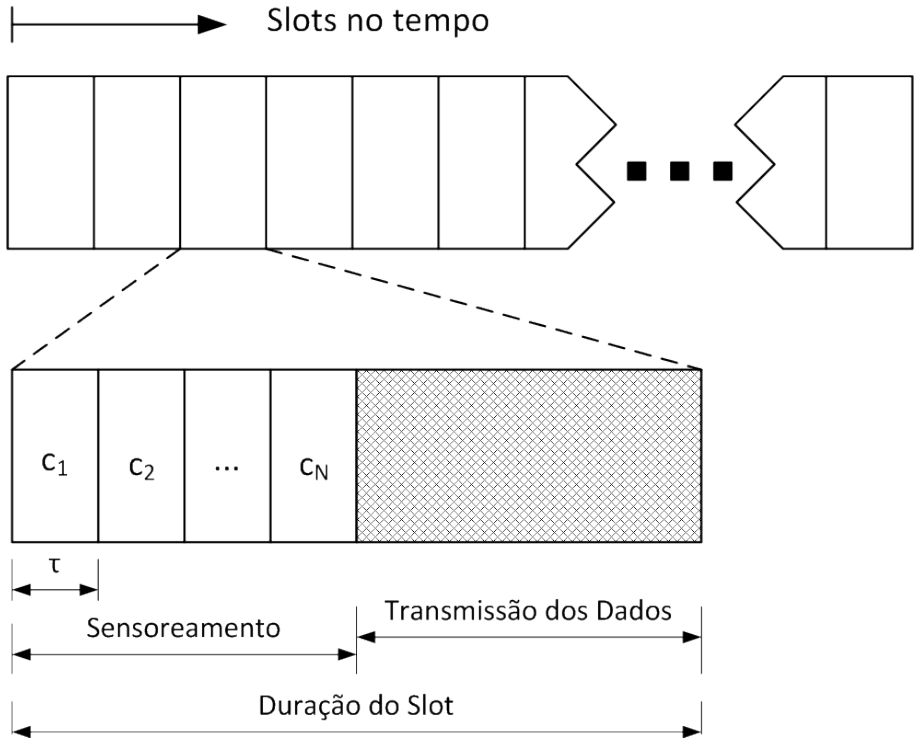
Como estamos interessados em aproveitar as oportunidades disponíveis ao longo do tempo, conforme mostrado na Figura 2.2, é interessante utilizar um modelo de sistema que possa conter essa dinâmica dos primários e, também, preocupado com a necessidade de sensorear e descobrir um canal "livre" de primários e estimar a sua capacidade, tudo isso em um intervalo de tempo muito curto, da ordem de milhares de microssegundos, se tomarmos como referência a duração do sensoreamento praticada na norma IEEE 802.22 [100].

A modelagem que adotamos é similar a apresentada em [19]. Assim, considere a existência de um número finito de canais de comunicação, N, e um usuário secundário ou, para o caso de múltiplos secundários, considere ainda que eles formam uma rede e competem para o acesso ao espectro, sem nenhuma coordenação entre eles nem consciência individual da presença um do outro, e que, embora muitos secundários possam escolher o mesmo canal, apenas um, no máximo, pode ser bem sucedido na sua utilização efetiva.

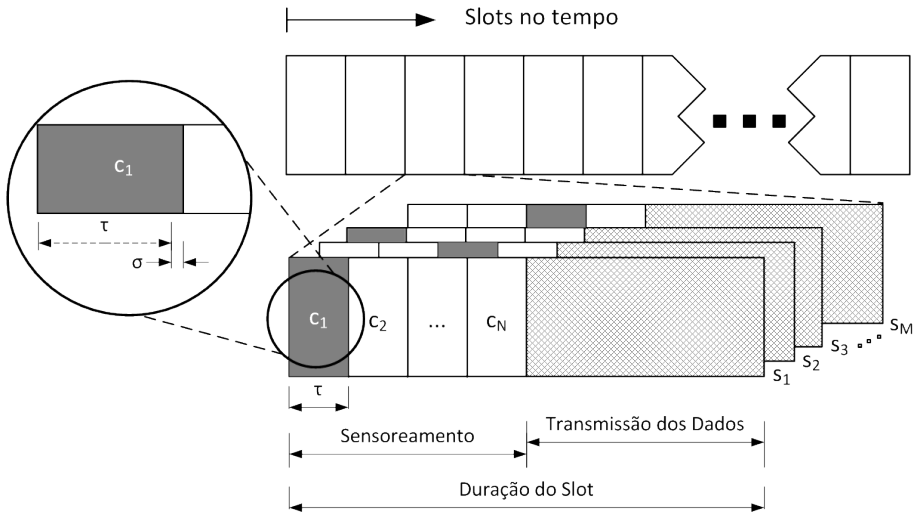
O secundário é equipado com um transceptor que é sintonizado em um dos N canais da rede, de mesma largura de banda, cujo tempo de acesso destinado à observação do canal e à transmissão de dados, que chamaremos *slot*, é constante, com duração T. Se tomarmos a referência da norma IEEE 802.22 [100], a duração do *slot* é semelhante a duração do quadro (*frame*), ou seja, poucos milissegundos.

Além disso, em cada *slot*, cada canal i possui a probabilidade $\Pr_{CH-AV_i}$ de estar

"livre" da atividade de usuários primários, i. e., de apresentar-se no estado "livre" ou "ocupado" durante todo o *slot*. Assume-se que $\Pr_{CH-AV_i}$ é independente do seu estado prévio e do estado de outros canais, dentro de cada *slot*, e i.i.d. entre *slots* [19, 64].



(a) Único secundário.



(b) Múltiplos secundários.

Figura 3.1: Modelo de um *slot* para os secundários.

A Figura 3.1 exemplifica a atividade de um (ou múltiplos) secundário (s) em um *slot*, a qual possui duas fases: *sensoreamento*, e *transmissão de dados*. Antes de decidir utilizar um canal para transmissão de dados em um determinado *slot*, o secundário deve realizar o sensoreamento desse canal (e apenas desse por restri-

ção do modelo adotado), no tempo igual a τ, com a finalidade de evitar colisões, especialmente com primários em atividade.

Como não existe conhecimento prévio a respeito dos estados dos canais, durante a fase de sensoreamento, cada secundário realiza o sensoreamento sequencial dos N canais. Chamaremos de k o número máximo de sensoreamentos por *slot*, sendo esse valor igual ao $\min(N, \lfloor(\frac{T}{\tau})\rfloor)$. Com isso, a quantidade de sequências possíveis é dada por $M = \begin{pmatrix} N \\ k \end{pmatrix} \times\ k\ !$. Assim, cada usuário secundário segue uma ordem preestabelecida dada pela sequência $s_n = \{o_1, o_2, ..., o_N\}, n < M$, correspondente a uma permutação dos N canais disponíveis, até decidir em qual canal parar.

Devido ao efeitos de desvanecimento nos canais, a relação sinal-ruído (SNR) obtida em um canal varia aleatoriamente entre *slots*. Assumimos que essa SNR aleatória é i.i.d entre os *slots* e os diferentes canais, e que é regida por uma distribuição arbitrária.

Assim, se o usuário secundário decidir transmitir em um canal considerado "livre", c_i, reserva-se o tempo igual a σ para a *estimação de canal*, que possui duração muito menor que a da fase de transmissão, é dependente da largura de banda do canal, da modulação e etc., e cujo resultado fornece a taxa de transmissão que será obtida, função da SNR momentânea desse secundário, nesse canal. Esta função, $F(SNR_i)$, mapeia de forma monotônica e crescente a SNR do canal c_i na taxa de transmissão que será obtida neste canal.

No modelo, assume-se que a atividade dos primários é desconhecida pelos secundários e cada tentativa de sensoreamento de canal utilizando um detector por energia é bem sucedida de acordo com a sua probabilidade de falha na detecção ($\mathrm{Pr}_{MISDETECTION}$) e que a *estimação de canal* é precisa e isenta de erros.

A eficiência no uso de uma sequência de sensoreamento está relacionada com o tempo gasto até que seja encontrado um canal adequado para ser utilizado pelo secundário e com a taxa de transmissão momentânea que pode ser obtida nesse canal. Desta forma, durante a fase de sensoreamento, caso o i-ésimo canal a ser sensoreado, c_i, seja considerado como "ocupado", o secundário realiza o sensoreamento no canal c_{i+1}, que é o próximo canal na ordem de sensoreamento que está sendo utilizada. Entretanto, caso o canal c_i seja considerado como "livre", a recompensa coletada corresponde à taxa de transmissão efetiva obtida pelo secundário no uso deste canal durante o tempo remanescente do *slot*.

Para o caso de múltiplos usuários, cada secundário pode ser considerado um nó da rede secundária onde, após estabelecer o canal que será utilizado, cada um realiza o controle de acesso ao meio com a adoção de uma camada de enlace específica, existindo portanto contenção ou colisão entre eles.

Inicialmente, o secundário que deseja acessar o meio efetua o sensoreamento do

canal para verificar se este está sendo utilizado por um primário. Caso o canal esteja "livre" de primários, pode haver disputa com outros secundários pelo canal, sendo o resultado da comparação do valor instantâneo de uma variável aleatória com um limiar, a indicação se haverá ou não colisão entre os secundários. No caso, utilizamos para esse limiar a *probabilidade de colisão*, segundo a fórmula $\Pr_{COLLISION} = 1 - (1 - \frac{1}{W_m})^{(US-1)}$, derivada a partir de observações de saturação da rede no padrão IEEE 802.11 [101, 102], onde o número de usuários secundários é dado por US e W_m é o tamanho médio da janela de contenção no 802.11. Nesse estudo [101], observa-se que o comportamento do parâmetro W_m é inversamente proporcional ao da $\Pr_{COLLISION}$ e o contrário ocorre com o a quantidade de nós na rede, que é diretamente proporcional a $\Pr_{COLLISION}$.

Se houver disputa, esses secundários aguardam durante um período de tempo, cuja duração é sorteada dentro de um intervalo determinado, e ao final desse período apenas um deles retorna ao sensoreamento do canal, para garantir que esteja "livre", e depois o utiliza efetivamente, e os demais perdem o *slot*. Essa estratégia de contenção é chamada $FAIL\_THEN\_QUIT$[39] foi adotada como caso extremo. Nessa situação, a recompensa é coletada apenas pelo vencedor da disputa, enquanto se houver colisão, nenhuma recompensa é coletada por nenhum dos secundários.

3.2 Seleção da Ordem de Sensoreamento de Canais em uma Rede Cognitiva Oportunista

Nesta seção são apresentados a ideia básica e os conceitos envolvidos na proposta inicialmente descrita nos nossos trabalhos em [103, 104].

O mecanismo, focado no problema da escolha da ordem de sensoreamento dos canais para o caso de um usuário secundário, em um sistema multicanal, utiliza um método de baixa complexidade baseado em uma máquina de aprendizagem por reforço (Seção 2.3), para determinar de maneira dinâmica uma ordem de sensoreamento a ser utilizada em cada *slot*.

Uma das vantagens desse mecanismo é a ausência da necessidade de conhecimento prévio a respeito da probabilidade de cada canal estar disponível, e da qualidade estimada de cada canal, por meio de suas SNRs médias. Outra vantagem importante é quanto a sua adaptabilidade as mudanças de características dos canais garantida pelo aprendizado com as tomadas de ação.

Logo, o mecanismo torna-se imune as possíveis mudanças nas probabilidades de disponibilidade dos canais, que podem ocorrer devido a uma alteração no padrão de atividade dos usuários primários, e as possíveis mudanças na qualidade dos canais, que podem ocorrer devido a mobilidade e aos efeitos de desvanecimento de larga

escala.

3.2.1 Mecanismo Proposto

A modelagem dos estados e ações é um dos maiores desafios encontrados no emprego da ferramenta de aprendizado por reforço no problema da escolha da ordem de sensoreamento. Uma modelagem descuidada pode gerar um modelo com muitos estados e/ou muitas ações, o que tornaria lenta a convergência do processo de investigação (*exploration*).

No nosso modelo, definimos o estado como o par ordenado formado pela posição na ordem de sensoreamento, o_k, e o canal que é sensoreado naquela posição, c_i. As ações possíveis de serem tomadas por um usuário secundário a partir de um estado (o_k, c_i) correspondem a escolher o canal que será sensoreado na próxima posição da ordem de sensoreamento, o_{k+1}.

Com isso, a *Q-table* será uma matriz de dimensões $N^2 \times N$ ($estados \times$ ações). Repare que essa modelagem faz com que não exista um estado ótimo a ser alcançado e, sim, uma sequência de ações que maximizam a recompensa imediata a cada estado, criando uma ordem de sensoreamento dinâmica.

Algumas *restrições* devem ser levadas em consideração no momento das tomadas de ação e na atualização da *Q-table*:

- Uma ação tomada no estado $(o_k, *)$, com $1 \leq k \leq (N-1)$, sempre leva a um estado onde a posição na ordem de sensoreamento é o_{k+1}.

- No estado onde a posição é $(o_N, *)$, que representa o último canal da ordem de sensoreamento, as ações indicam o primeiro canal a ser sensoreado no próximo *slot*, ou seja, leva a um estado onde a posição na ordem de sensoreamento é a $(o_1, *)$.

- Quando o usuário secundário decide usar um canal c_i na posição o_k, o sensoreamento nesse *slot* é finalizado. Nesse caso, o primeiro canal a ser sensoreado no próximo *slot* será determinado pela melhor ação no estado (o_N, c_i).

- O retorno a um canal sensoreado previamente, chamado *recall*, é proibido. Para isso, é necessário armazenar os canais já sensoreados no *slot* corrente. Desta forma, antes de tomar uma ação, o usuário secundário deve eliminar, das ações possíveis, os canais já sensoreados.

Uma parte importante do modelo diz respeito ao processo de atualização da *Q-table*, seguindo o modelo descrito na Seção 3.1:

- Quando o canal c_i é sensoreado como "livre" na posição o_k da ordem de sensoreamento, a recompensa obtida r_t será dada pela taxa de transmissão efetiva obtida pelo uso daquele canal durante o tempo remanescente do *slot*.

- Caso o canal seja considerado como "ocupado", é necessário que exista alguma penalização para reduzir o *Q-value* referente aquela ação. Desta forma, introduzimos o parâmetro δ, a quem chamamos *fator de perda*, que assume valores no intervalo $]0, 1]$ e multiplica o *Q-value* atual referente aquela ação. Assim, garante-se que quando uma ação leva a um canal "ocupado", o *Q-value* referente aquela ação será reduzido. Essa estratégia faz com que o *Q-value* represente também a disponibilidade dos canais, e não apenas a taxa de transmissão efetiva.

  Assim, os valores da *Q-table* são atualizados conforme a Equação 3.1, onde α é a taxa de aprendizagem, $\gamma \in [0, 1]$ é o fator de desconto, e onde o valor de cada variável corresponde ao instante t, exceto quando explicitado em contrário:

$$Q_{t+1}(s,a) =$$

$$\begin{cases} (1-\alpha)Q(s,a) + \alpha\left[r(s,a) + \gamma\ max_{a\in\mathbf{A}}Q(s_{t+1},a)\right], & se \text{ canal "livre"} \\ \delta\ Q(s,a), & se \text{ canal "ocupado"} \end{cases} \tag{3.1}$$

Com isso, a *recompensa* no uso de cada canal na sequência, obtida após cada *ação* selecionada (escolha de canal), é obtida através da solução fornecida pelo nosso mecanismo (Equação 3.2), onde adotamos como critério de parada a comparação da *recompensa* atual com o maior valor do *Q-value* referente ao *estado* atual e as possíveis *ações* na *Q-table*, em razão desse valor ser um bom indicador da *recompensa* futura, conforme tratado na Seção 2.3. Assim, torna-se possível estimar se a *recompensa* do canal sensoreado como "*livre*" ainda poderia ser maior, comparando o valor da *recompensa* atual com o histórico armazenado na informação do $Q-value$. Repare que mesmo no caso onde o canal "*livre*" não é utilizado, o *Q-value* referente aquela *ação* também é atualizado pela Equação 3.1.

$$r_i =$$

$$\begin{cases} e_i F(SNR_i), & se\ r_i > max_{a\in\mathbf{A}}Q(s,a) \\ r_{i+1}, & \text{nos demais casos} \end{cases} \tag{3.2}$$

onde e_i é a efetividade da transmissão, calculada pela fórmula $e_i = 1 - \frac{i\tau}{T}$, e r_{i+1}

para $i \leq N-1$ é a *recompensa* caso o usuário decida prosseguir no sensoreamento, a qual é esperada ser maior que a *recompensa* atual, conforme comentado acima.

Nessa proposta com *apenas um secundário*, calculamos como referência comparativa superior, o valor ótimo de recompensa, calculada pelo método de força bruta (ou busca exaustiva) de complexidade $O(N!)$, onde devido ao número de canais a serem sensoreados ser finito e igual a N, o método de indução reversa [1] (*backward induction*) pode ser aplicado para encontrar a recompensa esperada de uma sequência de N canais. Assim:

$$r_i = \begin{cases} e_i F(SNR_i), & se\ e_i F(SNR_i) > R_{i+1} \\ R_{i+1}, & \text{nos demais casos} \end{cases} \tag{3.3}$$

onde i corresponde ao canal atual, e_i é a efetividade da transmissão, calculada pela fórmula $e_i = 1 - \frac{i\tau}{T}$, e R_{i+1}, para $i \leq N-1$, é a recompensa esperada caso o usuário decida prosseguir no sensoreamento. O cálculo da recompensa esperada é dado por:

$$R_{i+1} = \begin{cases} p_{i+1}E[r_{i+1}] + (1-p_{i+1})R_{i+2}, & se\ i < N-1 \\ p_{i+1}E[r_{i+1}], & se\ \ i = N-1 \end{cases} \tag{3.4}$$

onde p_{i+1} e $E[r_{i+1}]$, para $i \leq N-1$, são, respectivamente, a *probabilidade de disponibilidade* do canal e o valor esperado da recompensa.

Repare que o conjunto de recompensas esperadas $\{R_1, R_2, ..., R_n\}$ pode ser obtida recursivamente a partir de R_N, através das Equações 3.3 e 3.4, segundo o método de indução reversa. Logo, R_1 representa a recompensa esperada pelo usuário secundário no uso de uma sequência de N canais. De forma genérica, R_i representa o valor esperado da recompensa no uso de uma sequência parcial de canais $(o_i, o_{i+1}, ..., o_N)$. Desta forma, o uso de um canal sensoreado como "livre" é vantajoso caso a recompensa no uso do canal, r_i, seja superior a recompensa esperada do restante da ordem de sensoreamento, R_{i+1}. Caso contrário, o usuário deve prosseguir no sensoreamento do próximo canal da sequência, não podendo retornar ao canal anterior (*no recall*), de modo a se manter o emprego da teoria da parada ótima (*optimal stopping*).

Uma estratégia puramente gananciosa (*greedy*) escolheria sempre a melhor ação, aquela que maximizasse $Q(s, a)$. Nosso mecanismo utiliza a estratégia ε-*greedy*, que regula esse comportamento rígido através de um limiar dado pela probabilidade ε, permitindo que em algumas ocasiões, sejam escolhidas outras ações além da gananciosa.

[1] É o processo de análise reversa no tempo, a partir do último estado de um problema, para determinar a sequência de ações ótimas. Em primeiro lugar, considera-se o último estado do problema e escolhe-se quais ações são possíveis naquele estado. Usando essas informações, pode-se então determinar as possíveis ações no penúltimo estado do problema e, assim sucessivamente, para todos os estados possíveis (ou seja, para cada conjunto possível de informações) em cada ponto no tempo.

E de forma intuitiva, utilizamos o procedimento de decrescer o valor dessa probabilidade após a progressão do mecanismo ao longo dos episódios, realizando no início mais investigação das ações em busca de melhores estimativas de recompensa e assumindo as ações mais gananciosas no final [105].

Algoritmo e Complexidade

O funcionamento do mecanismo é descrito em detalhes no Algoritmo 1.

Os passos 3 e 4 deste algoritmo correspondem à *fase de inicialização*, onde todos os pares estado-ação da *Q-table* são completados com zeros. Realizada a inicialização, começa a *fase de aprendizado*, que é repetida durante todo o período de funcionamento do mecanismo nos episódios, cada um com duração equivalente a um *slot*.

Inicialmente, nos passos 8 a 13, sob a estratégia *ε-greedy*, toma-se a decisão entre *investigação*, onde uma ação é escolhida aleatoriamente, e *exploração*, onde a melhor ação é escolhida, baseando-se na *Q-table*. Após a execução da ação, o mecanismo torna-se capaz de calcular a recompensa obtida e atualizar o correspondente *Q-value*.

Em seguida, no passo 14, apresenta-se uma característica importante da nossa proposta, comentada acima, que diz respeito ao uso dos canais sensoreados e considerados como "livres". De acordo com o modelo apresentado na Seção 3.1, a regra de parada ótima consiste em verificar se a recompensa instantânea é maior do que a recompensa esperada para o restante da sequência. Isso indica que nem sempre será vantajoso utilizar o primeiro canal "livre" encontrado. De forma similar, a nossa proposta também utiliza um critério de parada que consiste em comparar a recompensa atual, r_i, com o melhor *Q-value* das ações possíveis a partir daquele estado (passo 18) (Equação 3.2). Assim, é possível estimar se a recompensa do canal "livre" atual é superiora a recompensa esperada da melhor ação existente. Repare que mesmo no caso onde o canal "livre" não é utilizado, o *Q-value* referente aquela ação também é atualizado.

Caso o canal seja considerado "ocupado", o fator de perda δ é empregado para reduzir o *Q-value* referente aquela ação (passos 24 a 27).

Complexidade

Determinamos a eficiência do nosso algoritmo a partir da análise teórica da complexidade de pior caso para o tempo de execução ($T(n)$) e para a utilização de recursos ($S(n)$).

- *Eficiência Temporal*: observando o Algoritmo 1, é possível notar a repetição

dos passos de 5 a 28 durante todo o funcionamento do mecanismo, correspondente a *fase de aprendizado*.

Esse laço guiado pelo valor da variável *fim_do_sensoreamento* pode se repetir, no pior caso, até $N-1$ vezes, onde N é a quantidade máxima de canais.

Dentro desse laço, existem operações simples, que consomem 1 unidade de tempo de execução cada, tornando a complexidade de tempo constante, $\mathcal{O}(1)$.

Desta forma, nosso algoritmo possui complexidade $T(n) = \mathcal{O}(N)$.

- *Eficiência Espacial*: o maior consumo de recursos está relacionado ao armazenamento da *Q-table*, que é uma matriz de dimensões $N^2 \times N$ ($estados \times$ ações).

  Contudo, dentro do laço existente entre os passos 5 e 28, precisamos ter armazenado em memória apenas as ações possíveis a partir de um estado, ou seja, no pior caso, $N-1$ ações. Assim, $S(n) = \mathcal{O}(N)$.

3.2.2 Implementação

Esta subseção enumera as premissas usadas na implementação do simulador próprio e descreve o modelo de simulação adotado, para a avaliação do mecanismo proposto aplicado ao caso de apenas um secundário.

Simulador Próprio

Baseado no modelo do sistema descrito na Seção 3.1 e para avaliar o comportamento do nosso mecanismo de aprendizado por reforço (Subseção 3.2.1) na solução do problema da ordem de sensoreamento, desenvolvemos um simulador, utilizando a linguagem Tcl [106] e *software* "livre", sob o sistema operacional GNU/Linux, que emula o funcionamento de um secundário utilizando sequências de sensoreamento arbitrárias. Nesse simulador, as seguintes ordens de sensoreamento foram avaliadas:

- a sequência ótima obtida por força bruta (`Ótima`) [39];
- a sequência dinâmica dos canais fornecida pela nossa proposta (`RL`);
- a sequência de canais na ordem decrescente de suas probabilidades de disponibilidade (`Prob`) [19];
- a sequência dada pela ordem decrescente das capacidades médias de cada canal (`Cap`);

Algoritmo 1: Mecanismo proposto baseado em aprendizado por reforço.

```
1  fim_do_sensoreamento = 0;
2  /* inicialização da Q-table */
3  foreach s ∈ S, a ∈ A do
4      Q(s,a) = 0;
5  while ! fim_do_sensoreamento do
6      /* aprendizado */
7      x = Uniforme(0, 1);
8      if (x < ε) then
9          /* investigação → seleciona uma ação a aleatoriamente */
10         a = Uniforme(c_1, c_N);
11     else
12         /* exploração → escolhe ação a que possua o maior Q-value para o
               estado atual s */
13         a = argmax_s(Q(s, ∗));
14     if (canal "livre") then
15         /* canal c_a correspondente à ação a */
16         calcula recompensa r_t(s, a);
17         Q_{t+1}(s, a) ← (1 − α)Q_t(s, a) + αr_t(s, a);
18         if r_t(s, a) > max_{a∈A} Q(s′, a′) then
19             /* usa canal c_a */
20             fim_do_sensoreamento = 1;
21         else
22             /* não usa canal c_a */
23             continua sensoreamento;
24     else
25         /* canal c_a "ocupado" */
26         Q_{t+1}(s, a) ← δQ_t(s, a);
27         continua sensoreamento;
28     s_t = s_{t+1};
```

- a sequência dada pela ordem decrescente do produto das capacidades médias de cada canal pela sua respectiva probabilidade de disponibilidade (`Prob` × `Cap`); e,

- a sequência de canais na ordem definida através de sorteio, utilizando uma distribuição uniforme (`Aleatória`).

Em destaque, servindo como referência comparativa, a sequência ótima obtida pelo método de força bruta (`Ótima`) dentre todas as sequências possíveis de N canais, $N \in \mathbb{Z}$, em razão do método de indução reversa (*backward induction*) ser capaz de encontrar a recompensa esperada de qualquer dessas sequências. A sequência `Prob`, que é considerada ótima para o cenário com somente um secundário e sem a utilização de modulação adaptativa [19], e a sequência `Cap`, que ao seguir a ordem

decrescente das capacidades médias dos canais, intuitivamente conduz a uma boa expectativa do seu desempenho. A sequência `Aleatória` está presente também apenas como referência, devendo ser o limite inferior de desempenho para qualquer mecanismo.

Vale ressaltar que todas as sequências acima, com exceção da sequência `RL`, são *estáticas*, ou seja, não mudam durante toda a simulação. No caso da sequência `RL`, devido ao próprio aprendizado por reforço, a sequência pode variar durante a simulação. Além disso, todas as sequências, exceto a `RL` e a `Aleatória`, assumem o conhecimento a priori das capacidades médias de cada ($\overline{C}$) e/ou de suas probabilidades de disponibilidade (Pr_{CH-AV}).

Modelo de Simulação

No início de cada experimento, a capacidade média de cada canal ($\overline{C}$) é sorteada seguindo uma distribuição uniforme dentro do intervalo $[FHC*C_{MAX}, C_{MAX}]$, onde FHC é o *fator de homogeneidade* dos canais e $\overline{C}_{MAX}$ é a *capacidade média máxima* dos canais.

O parâmetro FHC deve assumir valores no intervalo [0,1]. Quanto maior for esse valor, maior é a homogeneidade das capacidades médias entre os canais e mais próximas de $\overline{C}_{MAX}$. Em outras palavras, esse parâmetro modifica a diferença entre as capacidades médias dos canais, tornando-os mais ou menos heterogêneos.

A probabilidade de disponibilidade (Pr_{CH-AV}) de cada canal i, $i \in \{1, ..., N\}$, é sorteada uniformemente dentro do intervalo [0,1].

Além disso, em cada slot T, a capacidade instantânea de cada canal (C_{INST}) é sorteada utilizando-se uma distribuição uniforme dentro do intervalo $[\overline{C} \times (1 - \frac{FVA}{2}), \overline{C} \times (1 + \frac{FVA}{2})]$, onde FVA é o *fator de variabilidade* do ambiente e $\overline{C}$ é a *capacidade média* de cada canal.

Quanto maior for o valor de FVA, a capacidade instantânea do canal (C_{INST}) apresenta uma alta variação. O parâmetro FVA deve assumir valores no intervalo [1,2]. Assim, FVA parametriza a intensidade da atenuação (*fading*) que pode ocorrer em cada canal.

Durante um experimento, em cada *slot* T, adotamos o modelo de comportamento para o primário representado pela ocupação dos canais de acordo com a sua probabilidade de disponibilidade, na forma "livre" ou "ocupado", usando uma distribuição uniforme ou, um modelo *ON-OFF* exponencialmente distribuído, conforme discutido na Subseção 2.5.1.

Desta forma, se o canal permaneceu no estado "ocupado" segundo uma distribuição exponencial de média "*OFF*" igual a μ_{OFF}. A média "*ON*" (estado "livre"), μ_{ON} pode ser obtida por:

$$\mu_{ON} = \frac{(1 - FU) \times \mu_{OFF}}{FU} \tag{3.5}$$

onde FU é a *fator de utilização do canal* pelo primário, ou seja, equivalente a probabilidade do canal estar indisponível.

Assim, o nosso cenário de simulação pode ser resumido como:

- Características físicas dos canais:

$$\overline{C} = \overline{C}_{MAX} \ (FHC + ((1 - FHC \ rand()))) \ , \ FHC \ \in \ [0, 1]$$

$$C_{INST} = \overline{C} \ (1 + FVA \ (0.5 - rand())) \ , \ FVA \ \in \ [1, 2]$$

  onde $\overline{C}$ é a *capacidade média* de cada canal, $\overline{C}_{MAX}$ é a *capacidade média máxima* dos canais, FHC é o *fator de homogeneidade* dos canais, FVA é o *fator de variabilidade* do ambiente e C_{INST} é a capacidade instantânea de cada canal, que é a única característica que varia a cada *slot*.

- Ocupação dos canais (modelo de comportamento para o usuário primário), "livre" ou "ocupado", variando a cada *slot* segundo:
  - uma distribuição uniforme; ou,
  - um modelo *ON-OFF* exponencial.

Finalmente, a cada *slot* T, o simulador calcula a recompensa obtida por cada uma das sequências implementadas, correspondente a taxa de transmissão efetiva (Seção 3.1), utilizando-se dos mesmos estados e das mesmas capacidades instantâneas dos canais (critério de justiça).

Uma rodada de simulação (ou experimento) consiste na execução de X *slots*. Ao final de cada rodada, o simulador fornece a recompensa média obtida por cada uma das sequências em todos os X *slots*.

3.2.3 Avaliação

Foram realizadas simulações com 50.000 *slots*, cada um correspondente a um episódio do *Q-learning*, e uma quantidade de canais variando de 3 a 8, incluindo um tempo de transiente e aprendizado para os mecanismos equivalente a 20% do número total de *slots*. A capacidade média máxima, $\overline{C}_{MAX}$, assumiu o valor fixo de 10.

O parâmetro W_m, descrito na Seção 3.1, correspondente ao tamanho médio da janela de contenção no 802.11, e parte da fórmula para o cálculo da probabilidade

de colisão derivada a partir de observações de saturação da rede no padrão IEEE 802.11, assumiu o valor de 8. Nesse estudo [101], observa-se que o comportamento do parâmetro W_m é inversamente proporcional ao da $\mathrm{Pr}_{COLLISION}$ e o contrário ocorre com o a quantidade de nós na rede, que é diretamente proporcional a $\mathrm{Pr}_{COLLISION}$.

O parâmetro α do `RL` assumiu o valor de 0.1 e o parâmetro γ, o valor 0, beneficiando a experiência recente face ao histórico. A *Q-table* foi inicializada com o valor 0.

Adotamos o valor inicial de 0.7 para o fator de investigação, ε, da estratégia implementada pelo nosso mecanismo, baseado no compromisso entre número de estados e os resultados obtidos, favorecendo a investigação de um número maior de estados dentro do período de aprendizado, passando para 0.1 ao final desse período.

O tamanho do *slot* T é um múltiplo inteiro do tempo necessário para sensorear um canal (τ), tendo sido configurado com o valor 20.

Os valores dos demais parâmetros foram escolhidos como os mais representativos dentro de seus intervalos de validade, estabelecidos após a realização de alguns testes.

Foram feitas 200 rodadas de simulação para cada conjunto de parâmetros, para o cenário anteriormente descrito, onde comparamos o desempenho das sequências apresentadas na Subseção 3.2.2. Em todos os resultados, apresentamos a média das recompensas coletadas a cada rodada, com barras de erro correspondentes ao intervalo de confiança de 95%. Os parâmetros usados, bem como os valores que eles assumem estão resumidos na Tabela 3.1.

No primeiro conjunto de simulações, o modelo de ocupação dos canais é escolhido de forma aleatória usando uma distribuição uniforme, isto é, eles são definidos como "ocupados" ou "livres" em cada *slot* de acordo com a correspondente probabilidade de disponibilidade do canal (Pr_{CH-AV}). No outro grupo de simulações, o modelo de ocupação dos canais é definido considerando que o primário possui um comportamento conhecido que pode ser modelado por um processo *ON-OFF* exponencial, com tempo médio "ocupado" (μ_{OFF}) fixo e o tempo médio "livre" obtido conforme a Equação 3.5.

Os resultados obtidos para esses dois modelos de ocupação de canal são apresentados nas duas subsecções seguintes.

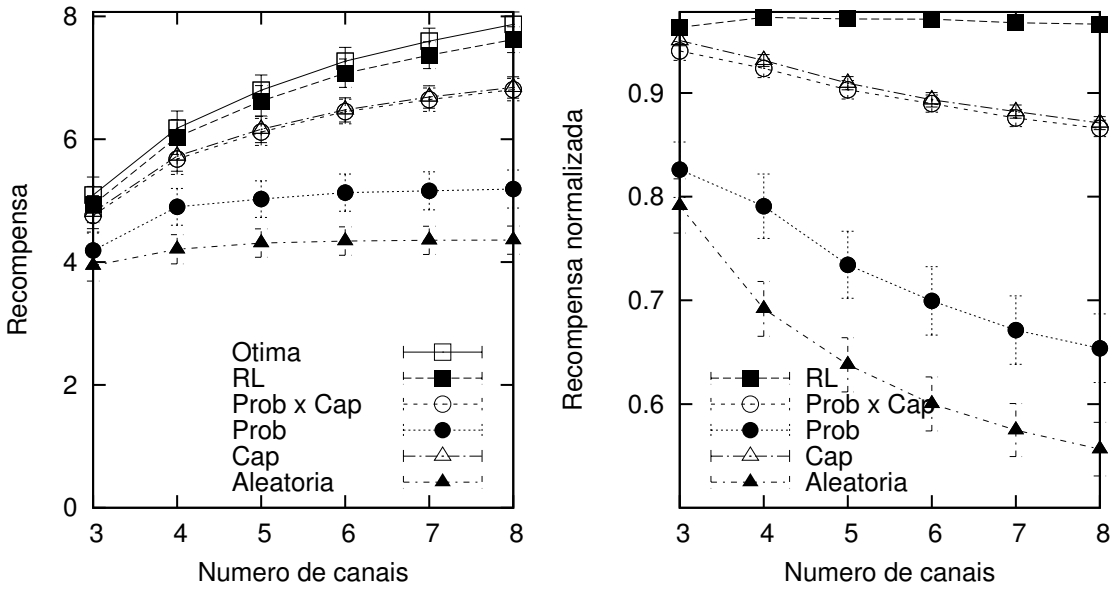
Nessa avaliação, consideramos que nossa detecção de primário é precisa e isenta de erros, $\mathrm{Pr}_{MISDETECTION} = 0.0$ e $\mathrm{Pr}_{FALSEALARM} = 0.0$.

Ocupação do Canal Uniformemente Distribuída

Na primeira etapa de simulações, a Figura 3.2 apresenta resultados da variação do número de canais de 3 até 8. Nestes resultados, $\overline{C}_{MAX}$ e o tamanho do *slot* foram configurados em 10, e os parâmetros FHC e FVA em 0.1 e 2.0, respectivamente.

Nome	Valor	Conteúdo
ε	0.1, $\varepsilon \in [0,1]$ (0.7 no transiente)	fator de investigação do RL
α	0.1, $\alpha \in [0,1]$	taxa de aprendizado do RL
δ	0.95, $\delta \in [0,1]$	fator de perda do RL
γ	0.0, $\gamma \in [0,1]$	fator de desconto do RL
FHC	$FHC \in [0,1]$	fator de homogeneidade dos canais
FVA	$FVA \in [1,2]$	fator de variabilidade do ambiente
μ_{OFF}	20, 100 e 200	duração média da ocupação do canal
$\overline{C}_{MAX}$	10	capacidade média máxima do canal
$\overline{C}$	$U(FHC * \overline{C}_{MAX}, \overline{C}_{MAX})$	capacidade média do canal
C_{INST}	$U(\overline{C} * (1 - FVA/2), \overline{C} * (1 + FVA/2))$	capacidade instantânea do canal
T	$20 \times \tau$	tamanho do *slot*
$Modelo\ de\ Canal$	Uniforme ou *ON-OFF* Exponencial	comportamento do primário
$\mathrm{Pr}_{MISDETECTION}$	0.0	prob. de falha na detecção do primário
$\mathrm{Pr}_{FALSEALARM}$	0.0	prob. de alarme falso de primário
$\#_{SLOTS}$	50.000	número de *slots*
$\#_{CH}$	3 a 8	número de canais
$\#_{RUNS}$	200	número de rodadas
$\#_{SEC}$	1	número de secundários

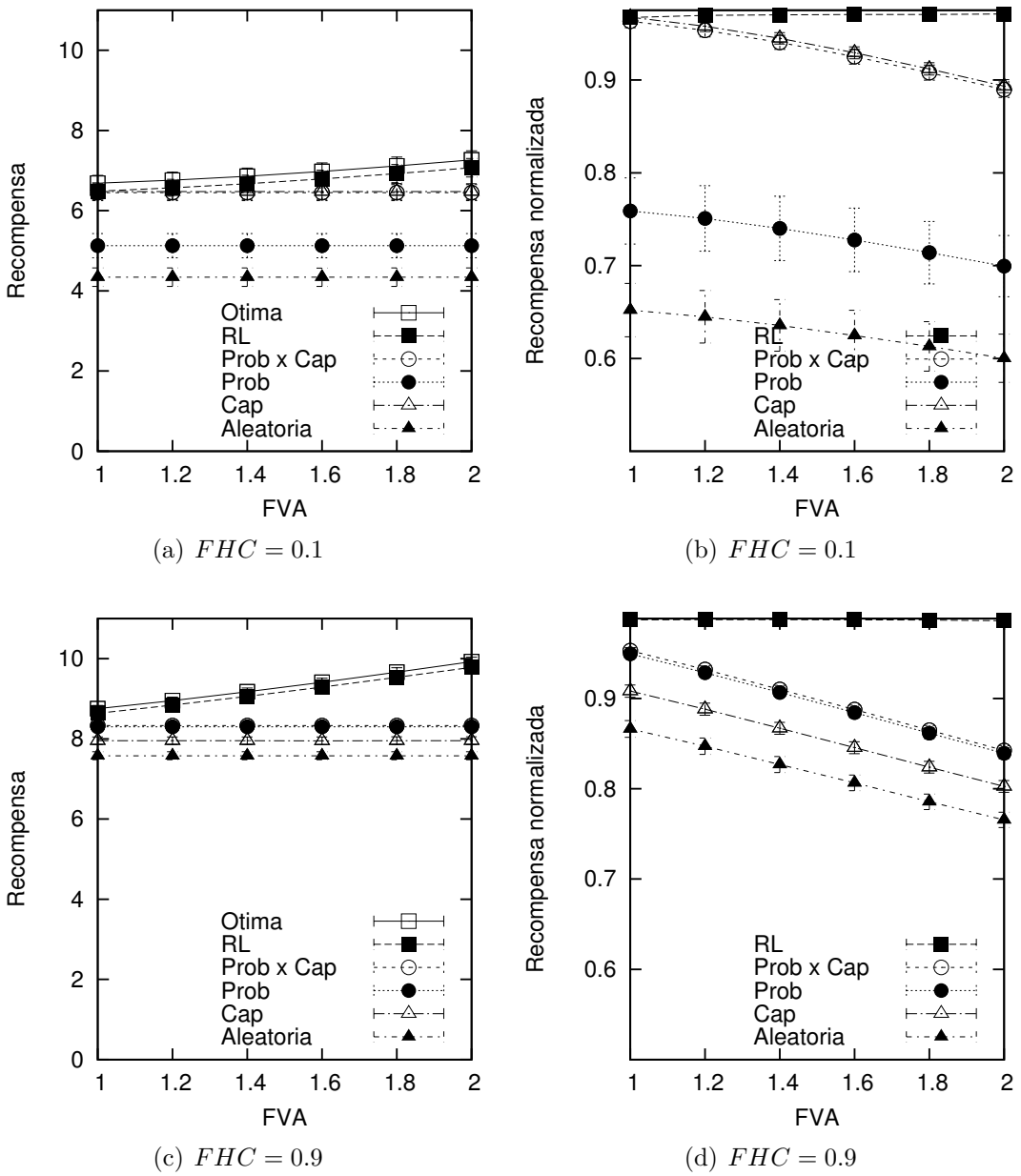
Tabela 3.1: Parâmetros da simulação e avaliação da proposta para um secundário.



(a) recompensas absolutas

(b) recompensas normalizadas

Figura 3.2: Resultados $FHC = 0.1$, $FVA = 2$ e $\delta = 0.95$.



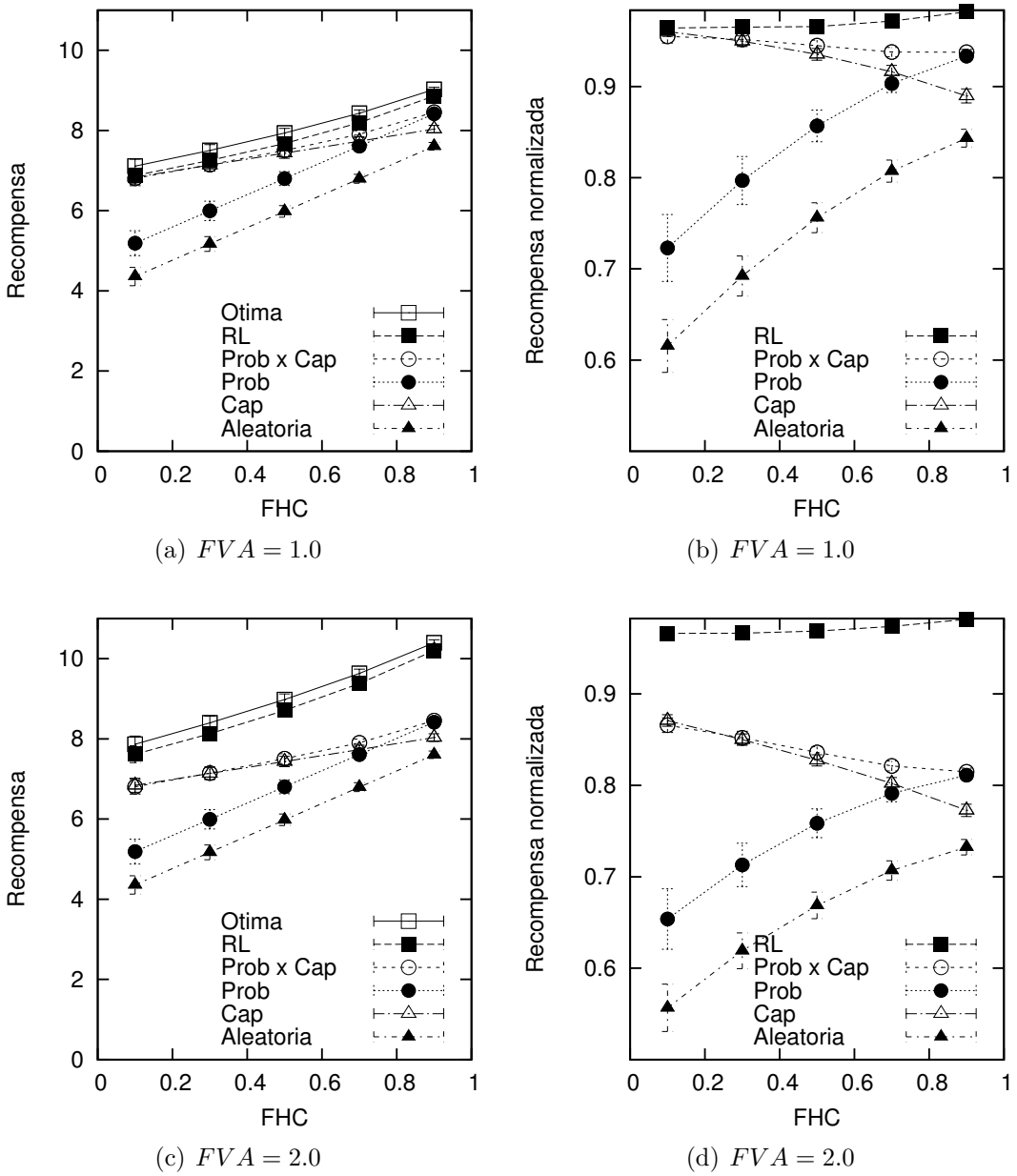
(a) $FHC = 0.1$ (b) $FHC = 0.1$

(c) $FHC = 0.9$ (d) $FHC = 0.9$

Figura 3.3: Resultados para 6 canais, com tamanho do *slot* = 10, capacidade = 10.

Analisando os resultados absolutos apresentados na Figura 3.2(a), percebe-se que o aumento do número de canais causa um aumento da recompensa média em todas as sequências simuladas. Este comportamento ocorre, pois o aumento do número de canais aumenta a possibilidade de existir um canal com alta capacidade média e grande probabilidade de disponibilidade. Também por este motivo, as sequências dadas por `Prob`, `Cap`, `Prob` × `Cap`, `RL` e `Ótima`, que são conscientes das capacidades de canal e das probabilidades de disponibilidade, obtêm um desempenho melhor do que o da sequência `Aleatória`.

A Figura 3.2(b) apresenta os resultados normalizados pelo desempenho obtido pela sequência `Ótima`. A comparação do desempenho das sequências neste gráfico



(a) $FVA = 1.0$ (b) $FVA = 1.0$

(c) $FVA = 2.0$ (d) $FVA = 2.0$

Figura 3.4: Resultados para 6 canais, com tamanho do *slot* = 10, capacidade = 10.

mostra que a nossa proposta, `RL`, é a única que alcança resultados próximos ao valor ótimo. O desempenho das outras sequências é desfavorável, pois nenhuma delas utiliza regras de parada baseadas na predição do desempenho estimado de se continuar sensoreando os próximos canais da sequência, ou seja, nestas outras soluções o primeiro canal sensoreado como "livre" sempre é utilizado. Desta forma, o `RL`, que utiliza as experiências passadas armazenadas na *Q-table*, consegue determinar de maneira eficiente se é vantajoso utilizar um determinado canal sensoreado como "livre". Outra observação interessante a respeito das curvas da Figura 3.2(b) é que o desempenho da sequência `Prob` é inferior ao desempenho das sequências `Cap` e `Prob` × `Cap`. Isso indica que neste cenário a diferenciação entre as capacidades médias

dos canais ($\overline{C}$) é mais importante do que a diferenciação entre suas probabilidades de disponibilidade. Com isso, é melhor ordenar os canais pela ordem decrescente de suas capacidades médias, pois aumenta-se a probabilidade de o primeiro canal sensoreado como "livre" ser um canal de maior capacidade.

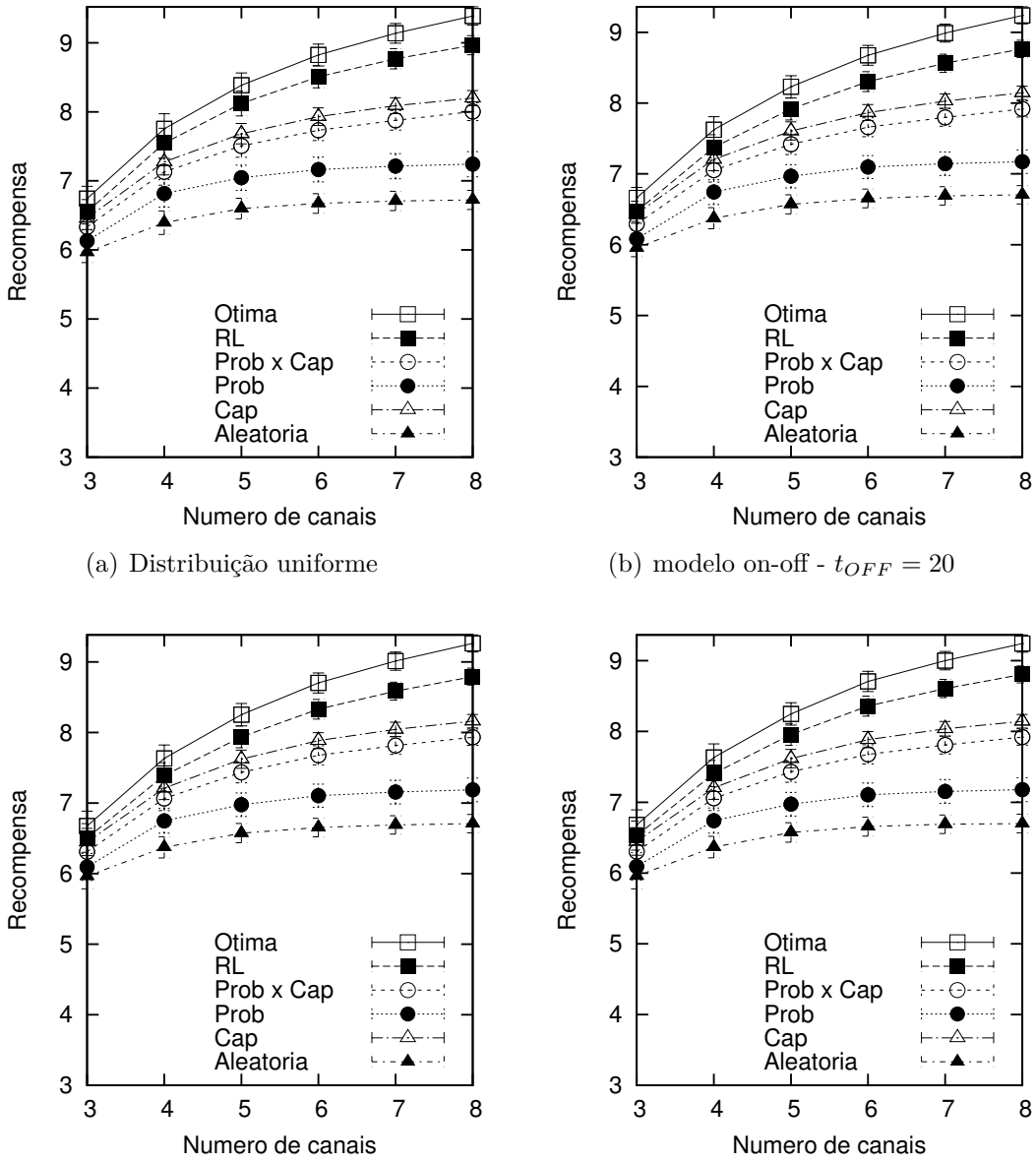
Na segunda etapa de simulações, variamos os parâmetros FHC e FVA, conforme descrito na Seção 3.2.2, tamanho do *slot* e capacidade máxima iguais a 10, e comparamos as recompensas absolutas e normalizadas pelo valor da recompensa obtida pela sequência ótima.

O FVA modifica a variabilidade da capacidade instantânea dos canais em torno da capacidade média, que varia a cada *slot*, e é representativa da variação dinâmica da SNR. As estratégias `Prob`, `Cap`, `Prob` $\times$ `Cap` e `Aleatória` são invariantes com relação a esse parâmetro, pois elas utilizam as médias das variáveis aleatórias para a geração das suas respectivas sequências. Assim, observando-se as Figuras 3.3(a) e 3.3(c), que consideram a recompensa absoluta, e as Figuras 3.3(b) e 3.3(b), que consideram a recompensa normalizada, para todos os valores de FVA, as sequências utilizadas serão sempre as mesmas, e as suas recompensas tendem para um valor médio constante. Para as estratégias `FB` e `RL`, o aumento da recompensa absoluta se deve ao fato dessas estratégias apenas utilizarem um canal "livre" caso a recompensa instantânea do uso desse canal for maior do que a recompensa esperada do resto da sequência. Assim, elas tendem a utilizar canais "livre"s com grandes capacidades instantâneas, devido a sua maior variabilidade. As demais estratégias sempre utilizam um canal quando ele é encontrado "livre", independente da sua capacidade instantânea.

O FHC modifica a homogeneidade dos canais com relação as suas capacidades médias. Com o aumento do valor desse parâmetro, as capacidades médias ficam mais próximas da $\overline{C}_{MAX}$, aumentando a recompensa com o aumento do FHC para todos os mecanismos, conforme a Figura 3.4. No entanto, os mecanismos baseados em ordenação por capacidade, i.e. `Cap` e `Prob` $\times$ `Cap`, não crescem na mesma proporção que os demais. A explicação para isso está no fato de que ao se homogeneizar os canais, o peso da capacidade na escolha da sequência se torna menos importante. Por isso, a estratégia `Prob` melhora a medida que os canais se tornam mais homogêneos.

Ocupação do Canal Segundo o Modelo *ON-OFF* Exponencial

Nas simulações correspondentes a esse modelo de ocupação, os valores do *fator de utilização* do canal para cada *slot* é variado de acordo com uma variável uniformemente distribuída dentro do intervalo [0.1, 0.9] e os μ_{OFF} são definidos em número de *slots*.



(a) Distribuição uniforme

(b) modelo on-off - $t_{OFF} = 20$

(c) modelo on-off - $t_{OFF} = 100$

(d) modelo on-off - $t_{OFF} = 200$

Figura 3.5: Distribuição Uniforme versus modelo ON-OFF.

Na primeira rodada de simulações, avaliamos o impacto da utilização de um modelo *ON-OFF* para caracterizar a atividade dos usuários primários. Para estas simulações, μ_{OFF} assume valores iguais a 20, 100 e 200. Além disso, os parâmetros δ, FHC e FVA assumem valores de 0.95, 0.1 e 1.5, respectivamente. Os resultados obtidos para recompensa absoluta estão na Figura 3.5.

Ao comparar os resultados nas Figuras 3.5(b), 3.5(c) e 3.5(d) com aqueles na Figura 3.5(a), pode-se observar que o desempenho do `RL` é mais impactado com o aumento de t_{OFF}. No entanto, a degradação de desempenho é pequena, o que demonstra a robustez do `RL` contra diferentes disponibilidades do canal, sem qualquer conhecimento prévio.

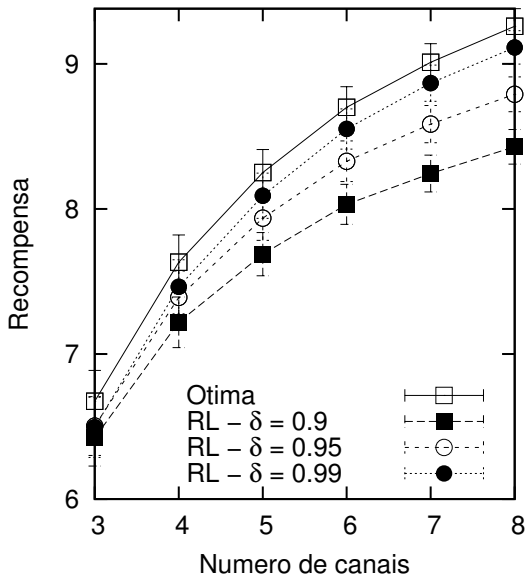
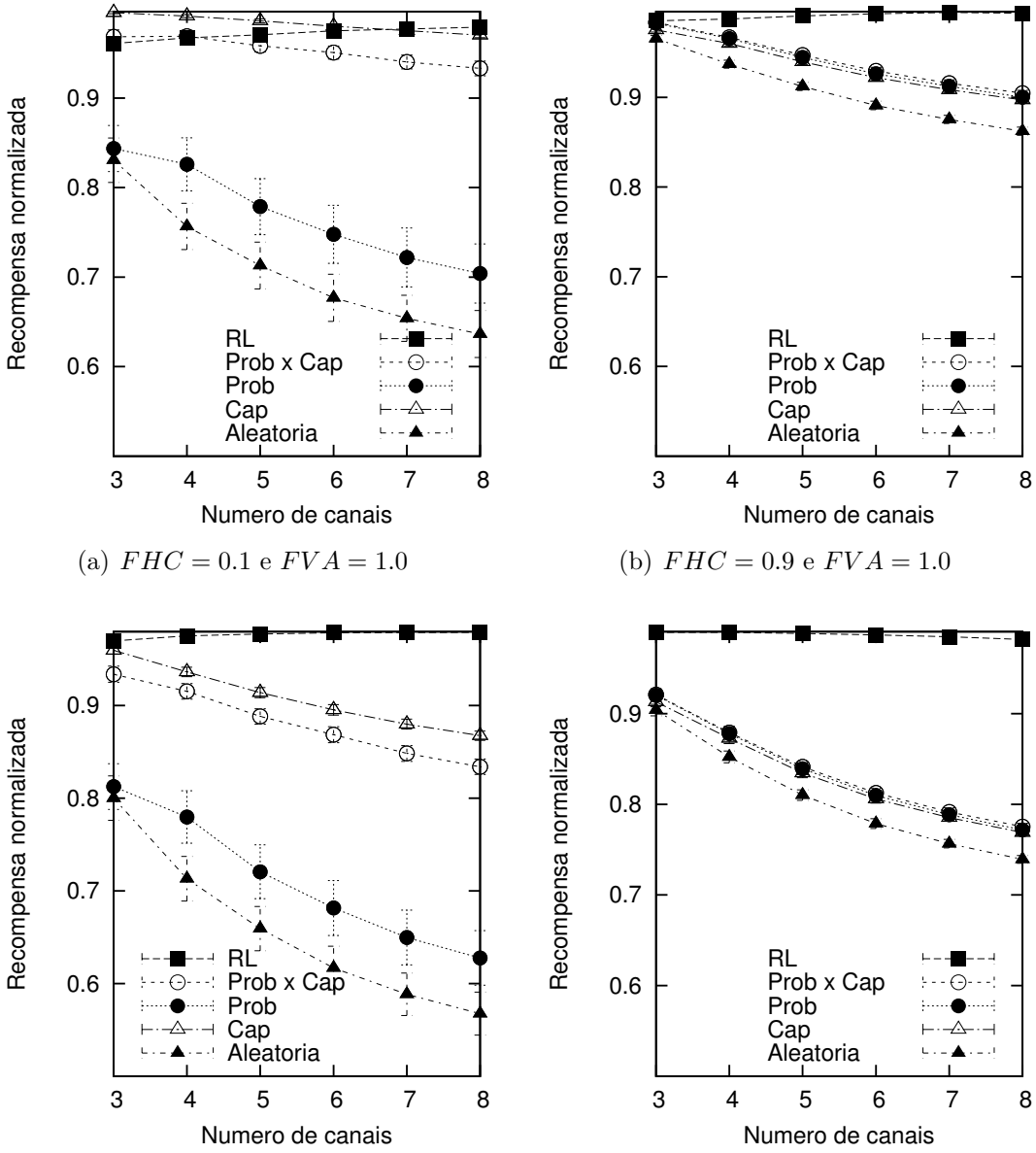


Figura 3.6: Influência do parâmetro δ.

Nessa rodada de simulações, variamos o parâmetro δ (Subseção 3.2.2). Este parâmetro determina a penalização incorrida pela escolha de um canal "ocupado" na sequência, levando a uma redução do *Q-value* para essa ação. Mantivemos o valor de μ_{OFF} igual a 100 e os parâmetros α, FHC e FVA, assumiram valores iguais a 0.9, 0.1 e 1.5, respectivamente. A Figura 3.6 mostra a recompensa absoluta para diferentes valores de δ. Os resultados mostram que quanto maior é o valor de δ, melhor é o desempenho do `RL`. Isso ocorre porque no modelo on-off, a indisponibilidade de um canal ocorre em rajadas de *slot*s. Desta forma, a recompensa no uso de um canal "ocupado" torna-se reduzida, levando a sequência `RL` a mudar mais frequentemente.

O objetivo dessa simulação foi avaliar o impacto da penalização sofrida pelo mecanismo proposto `RL`, dada pelo parâmetro δ, quando da tomada de ações desfavoráveis, e a sua relação com a velocidade de reação do mecanismo na direção desejada, ajustando-se às variações dinâmicas do ambiente (de RF) no menor tempo e corrigindo as suas ações para melhorar a recompensa coletada.

Finalmente, na última etapa de simulações, avaliamos os parâmetros FHC e FVA quando é utilizado o modelo *ON-OFF* de ocupação do canal. Mantivemos o valor de μ_{OFF} igual a 100 e δ assume o valor de 0.99. Os resultados com as recompensas normalizadas são mostrados na Figura 3.7. No cenário com a heterogeneidade maior e menor variabilidade dos canais ($FHC = 0.1$ e $FVA = 1.0$), o `RL` apresenta o pior desempenho em relação a `Ótima` para um pequeno número de canais. Isso ocorre devido a natureza dinâmica do `RL`. Nesses cenários, qualquer mudança temporária da ordem ideal provoca uma degradação significativa no desempenho. Em contraste, o `RL` alcança um desempenho muito próximo da `Ótima`



(a) $FHC = 0.1$ e $FVA = 1.0$

(b) $FHC = 0.9$ e $FVA = 1.0$

(c) $FHC = 0.1$ e $FVA = 2.0$

(d) $FHC = 0.9$ e $FVA = 2.0$

Figura 3.7: Influência da heterogeneidade e variabilidade dos canais.

nos outros cenários.

3.3 Ordem de Sensoreamento de Canais em uma Rede de Rádios Cognitivos Multiusuário

Em continuidade ao trabalho, a nossa proposta anterior é ampliada para a inclusão de múltiplos usuários, conforme descrita no nosso trabalho em [107], evoluindo o nosso mecanismo para uma solução baseada em aprendizado por reforço multiagente utilizando a técnica *Q-learning* aplicada ao problema da busca da ordem de

sensoreamento ótima de canais em uma rede de rádios cognitivos, guardando sua característica original de baixa complexidade computacional.

As vantagens originais do mecanismo, apontadas na Secão 3.2, que o tornam imune as possíveis mudanças nas probabilidades de disponibilidade e na qualidade dos canais (SNRs médias), foram mantidas:

- Desnecessário o conhecimento prévio a respeito da probabilidade de cada canal estar disponível;
- Desnecessário o conhecimento prévio a respeito da qualidade estimada de cada canal por meio de suas SNRs médias; e,
- Adaptabilidade as mudanças de características dos canais garantida pelo aprendizado com as tomadas de ação.

3.3.1 Mecanismo Proposto

Cada usuário na rede secundária foi modelado como um agente de aprendizagem seguindo uma estratégia independente, conforme as características da classe multiagente do tipo *independent learners* [41]. Esse tipo de agente não possui conhecimento dos demais agentes, interagindo com o ambiente de RF como se estivesse sozinho.

A escolha dessa classe multiagente, em particular, foi motivada pela característica autônoma de cada agente, que impusemos para a solução do problema, pelo custo da criação de um canal de comunicação, necessário para coordenação entre os agentes, pela escalabilidade do mecanismo através do crescimento do número de agentes, além da enorme dificuldade de satisfazer, na prática, o requisito de observação das ações conjuntas dos agentes, necessário para a aplicação de mecanismos da classe *joint learners* [42].

Assim, em um determinado instante de decisão, o agente observa somente o seu próprio ambiente de RF e, no próximo instante, ele coleta sua recompensa local seguindo sua decisão de melhor ação. Com isso, o mecanismo de aprendizagem obtém consciência do ambiente de RF observando as consequências das ações tomadas anteriormente [27]. E com o tempo, o agente aprende a melhor ação, que leva à maximização da recompensa.

Nesse caso, ele é incapaz de observar as recompensas e ações dos demais agentes e, por consequência, pode aplicar a técnica *Q-learning* na sua forma tradicional (Figura 3.8).

A modelagem dos *estados*, *ações* e o cálculo da *recompensa* coletada mantiveram-se conforme descrito na Subseção 3.2.1.

Na avaliação realizada para um secundário (Subseção 3.2.3), utilizamos um valor baixo para α, que aproveitava pouco o valor da recompensa coletada a cada

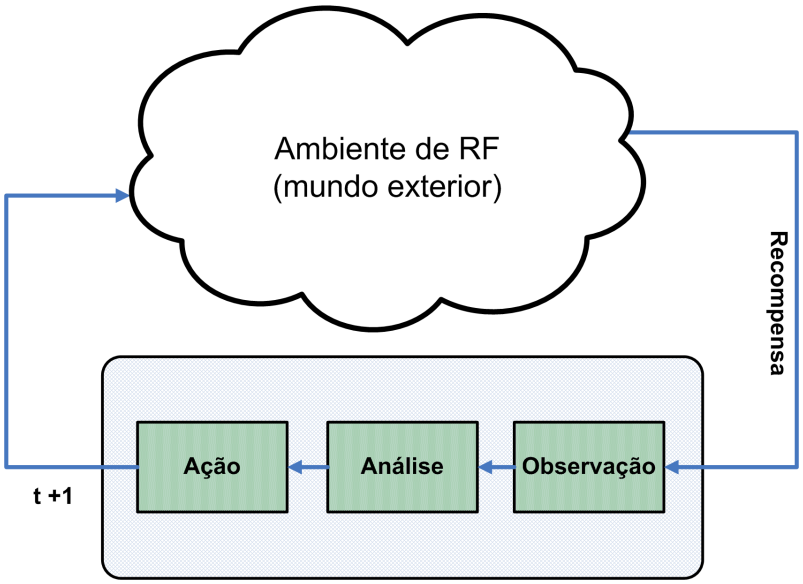


Figura 3.8: Agente e ambiente de aprendizagem.

ação e privilegiava muito o valor armazenado na *Q-table*, conforme demostrado pela Equação 2.1. Isso é uma boa estratégia a partir do momento que a *Q-table* tiver armazenado informação suficiente sobre o ambiente de RF - o que pode ser verificado quando houver pouca ou nenhuma variação dos *Q-values* -, fruto da experiência obtida na evolução do mecanismo ao longo dos episódios. No início, entretanto, parece ser mais promissor aproveitar o valor da recompensa coletada no presente, até para melhorar (e acelerar) o processo de aprendizado que é armazenado na *Q-table*.

Por isso, introduzimos no nosso mecanismo um meta-parâmetro, β, onde $\beta \in [0, 1]$, com o objetivo de realizar o ajuste dinâmico da sua taxa de aprendizado α, segundo a Equação 3.6, onde I é um contador da quantidade de visitas aos estados do modelo de sistema que adotamos, descrito na Seção 3.1. Maiores valores de α indicam maior importância para a experiência recente em relação ao histórico (ou "aprendizado" realizado) [31].

$$\alpha = \frac{1}{1 + \beta \times I} \tag{3.6}$$

Ao evoluirmos o nosso mecanismo, incluímos a outra estratégia clássica que faltava, *softmax*, a qual possui o parâmetro *temperatura* t, responsável pelo controle da quantidade de investigação que será praticada a cada estado. Um valor alto para t torna a escolha das ações quase equiprovável. Valores baixos, pelo contrário, causam uma grande diferença na probabilidade de escolha das ações.

Assim, para permitir uma quantidade maior de investigação de ações no início do funcionamento do nosso mecanismo e, reduzir a investigação depois dele já ter adquirido consciência suficiente do ambiente (de RF), introduzimos o controle do

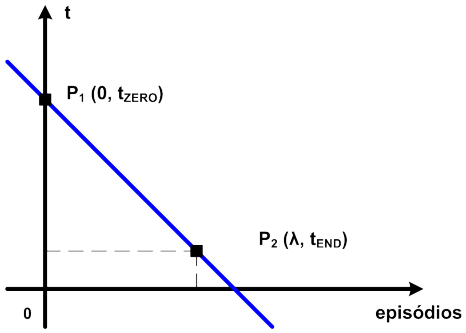


Figura 3.9: Ajuste dinâmico do parâmetro *temperatura*.

parâmetro t de forma dinâmica, segundo a Equação 3.7, onde t_{ZERO} e t_{END} correspondem, respectivamente, ao valor do parâmetro t no início do funcionamento do mecanismo e ao valor final que ele assume, e que não será mais modificado, após terem decorridos λ episódios, correspondentes a duração do período transiente e de aprendizado do mecanismo. A variável ι corresponde ao episódio corrente. Assim, o parâmetro temperatura é ajustado de modo dinâmico, segundo uma reta (Figura 3.9).

$$t = \frac{t_{END} - t_{ZERO}}{\lambda} \times \iota + t_{ZERO} \tag{3.7}$$

De resto, o mecanismo proposto manteve-se conforme descrito na Seção 3.2.1.

Algoritmo e Complexidade

O algoritmo adotado (Algoritmo 2) está ligeiramente modificado em relação ao original (Algoritmo 1), apresentado na Seção 3.2.1. A diferença está no passo 7, que realiza a seleção da Estratégia, entre a ε-*greedy*, presente na versão original do algoritmo, e a $softmax$ (passo 13), implementada agora, ambas descritas na Seção 2.3.

Para o $softmax$, inicialmente é realizado o cálculo das probabilidades das ações a usando a distribuição de *Boltzmann* e, em seguida, elas são ordenadas segundo a informação presente no parâmetro *temperatura* t.
Complexidade

Determinamos a eficiência do nosso algoritmo a partir da análise teórica da complexidade de pior caso para o tempo de execução ($T(n)$) e para a utilização de recursos ($S(n)$).

- *Eficiência Temporal*: observando o Algoritmo 2, é possível notar a repetição

dos passos de 5 a 32 durante todo o funcionamento do mecanismo, correspondente a *fase de aprendizado*.

Esse laço guiado pelo valor da variável *fim_do_sensoreamento* pode se repetir, no pior caso, até $N-1$ vezes, onde N é a quantidade máxima de canais.

Dentro desse laço, existem operações simples, que consomem 1 unidade de tempo de execução cada, tornando a complexidade de tempo constante, $\mathcal{O}(1)$.

Desta forma, nosso algoritmo manteve a complexidade de $T(n) = \mathcal{O}(N)$.

- *Eficiência Espacial*: o maior consumo de recursos continua relacionado ao armazenamento da *Q-table*, que é uma matriz de dimensões $N^2 \times N$ ($estados \times$ ações).

  Contudo, dentro do laço existente entre os passos 5 e 32, precisamos ter armazenado em memória apenas as ações possíveis a partir de um estado, ou seja, no pior caso, $N-1$ ações. Assim, $S(n) = \mathcal{O}(N)$.

3.3.2 Implementação

Nesta subseção, são ressaltadas as premissas usadas na implementação do simulador próprio e é descrito o modelo de simulação adotado, para a avaliação do nosso mecanismo aplicado ao caso de múltiplos secundários.

Simulador Próprio

Embora existam alguns simuladores, como o NetSim [108], o ns-2 [109] e o ns-3 [110], conhecidos da comunidade e que permitem a avaliação de uma rede de rádios cognitivos, resolvemos continuar o desenvolvimento do nosso próprio simulador, principalmente em razão do relacionamento intrínseco existente entre o simulador que já havíamos desenvolvido e o nosso mecanismo, que objetiva a análise de desempenho de uma sequência de sensoreamento de canais, e como validação da prova de conceito das propostas.

Desta forma, evoluímos o nosso simulador, desenvolvido em linguagem Tcl [106] sob GNU/Linux, e descrito na Subseção 3.2.2, para que emulasse o funcionamento de uma rede secundária, com a adoção de uma camada de enlace específica, onde existe contenção entre os usuários.

Nesse simulador, onde cada um dos usuários da rede secundária utiliza sequências de sensoreamento individuais, as seguintes ordens de sensoreamento foram avaliadas:

Algoritmo 2: Ampliação do mecanismo baseado em aprendizado por reforço para multiusuário.

```
1  fim_do_sensoreamento = 0;
2  /* inicialização da Q-table */
3  foreach s ∈ S, a ∈ A do
4      Q(s,a) = 0;
5  while ! fim_do_sensoreamento do
6      /* aprendizado */
7      if (ε−greedy) then
8          sorteia número aleatório x entre 0 e 1;
9          if (x < ε) then
10             /* investigação → seleciona uma ação a aleatoriamente */
11         else
12             /* exploração → escolhe ação a que possua o maior Q-value para o
                   estado atual s */
13     else if (softmax) then
14         /* calcula a probabilidade da ação a usando a distribuição de
               Boltzmann */
15         Pr(a_i) = e^{Q(s,a_i)/t} / Σ_a e^{Q(s,a)/t};
16         /* ordena as probabilidades */
17         /* escolhe a ação de maior probabilidade a partir do estado atual s */
18     if (canal "livre") then
19         /* canal c_a correspondente à ação a */
20         calcula recompensa r_t(s, a);
21         Q_{t+1}(s, a) ← (1 − α)Q(s, a) + α [r(s, a) + γ max_{a∈A} Q(s_{t+1}, a)[;
22         if r_t(s, a) > max_{a∈A} Q(s′, a′) then
23             /* usa canal c_a */
24             fim_do_sensoreamento = 1;
25         else
26             /* não usa canal c_a */
27             continua sensoreamento;
28     else
29         /* canal c_a "ocupado" */
30         Q_{t+1}(s, a) ← δQ_t(s, a);
31         continua sensoreamento;
32     s_t = s_{t+1};
```

- a sequência dinâmica dos canais fornecida pela nossa proposta (`RL`);
- a sequência de canais na ordem decrescente de suas probabilidades de disponibilidade (`Prob`) [19];
- a sequência dada pela ordem decrescente das capacidades médias de cada canal

(`Cap`);

- a sequência de canais na ordem definida através de sorteio, utilizando uma distribuição uniforme (`Aleatória`).

Em destaque, a sequência *Prob*, que é considerada ótima para o cenário com somente um secundário [19] e sem a utilização de modulação adaptativa, e a sequência *Cap*, que ao seguir a ordem decrescente das capacidades médias dos canais, intuitivamente conduz a uma boa expectativa do seu desempenho. A sequência `Aleatória` está presente apenas como referência, devendo ser o limite inferior de desempenho para qualquer mecanismo.

Vale ressaltar que todas as sequências acima, com exceção da sequência `RL`, são *estáticas*, ou seja, não mudam durante toda a simulação. No caso da sequência `RL`, devido ao próprio aprendizado por reforço, a sequência pode variar durante a simulação.

Além disso, todas as sequências, exceto a `RL` e a `Aleatória`, assumem o conhecimento a priori das capacidades médias de cada ($\overline{C}$) e/ou de suas probabilidades de disponibilidade (Pr_{CH-AV}).

Modelo de Simulação

O modelo de simulação manteve-se na maior parte conforme descrito na Subseção 3.2.2.

Recordando:

- Fator de Homogeneidade dos Canais, $FHC \in [0, 1]$. Quanto maior for esse valor, maior é a homogeneidade das capacidades médias dos canais ($\overline{C}$) no entorno da capacidade média máxima ($\overline{C}_{MAX}$); e,

- Fator de Variabilidade do Ambiente, $FVA \in [1, 2]$. Quanto maior for o valor de FVA, a capacidade instantânea do canal (C_{INST}) apresenta uma alta variação.

Assim, o nosso cenário de simulação pode ser resumido como:

- Características físicas dos canais:

$$\overline{C} = \overline{C}_{MAX}\ (FHC + ((1 - FHC\ rand()))\ ,\ FHC\ \in\ [0, 1]$$

$$C_{INST} = \overline{C}\ (1 + FVA\ (0.5 - rand()))\ ,\ FVA\ \in\ [1, 2]$$

onde $\overline{C}$ é a *capacidade média* de cada canal, $\overline{C}_{MAX}$ é a *capacidade média máxima* dos canais, FHC é o *fator de homogeneidade* dos canais, FVA é o

fator de variabilidade do ambiente e C_{INST} é a capacidade instantânea de cada canal, que é a única característica que varia a cada *slot*.

- Ocupação dos canais (modelo de comportamento para o usuário primário), "livre" ou "ocupado", variando a cada *slot* segundo:
  - uma distribuição uniforme; ou,
  - um modelo *ON-OFF* exponencial.

A cada *slot* T, o simulador calcula a recompensa obtida por cada uma das sequências implementadas, correspondente a taxa de transmissão efetiva (Seção 3.1), utilizando-se dos mesmos estados e das mesmas capacidades instantâneas dos canais (critério de justiça).

Uma rodada de simulação consiste na execução de X *slots*. Ao final de cada rodada, o simulador fornece a recompensa média obtida por cada uma das sequências em todos os X *slots*.

3.3.3 Avaliação

Os detalhes da parametrização para a avaliação do nosso mecanismo, ampliado para o caso de múltiplos secundários, mantiveram-se em grande parte conforme descrito no início da Subseção 3.2.3. Passaremos a comentar sobre as modificações realizadas.

A escolha dos parâmetros FHC e FVA é importante para o desempenho do mecanismo `RL`, que é sensível a esses parâmetros, conforme visto na avaliação discutida na Subseção 3.2.3. Desta forma, atribuímos os valores de 0.1 e 1.0, respectivamente, para FHC e FVA, que torna o cenário de simulação para cada canal, menos homogêneo em relação $\overline{C}_{MAX}$ e menos variável em relação C_{INST}, e onde o nosso mecanismo apresentou maior sensibilidade, conforme o resultado da Figura 3.7.

O tamanho do *slot* T é um múltiplo inteiro do tempo necessário para sensorear um canal (τ), tendo sido configurado como variável, com valor igual ao dobro do número de canais utilizados na rodada correspondente multiplicado por τ. Essa modificação tem o objetivo de manter a proporcionalidade entre a quantidade de canais e o tempo do *slot*.

Em razão da existência de múltiplos secundários, u_j, onde $j \in \{1..J\} | J \in \mathbb{Z}$, em todos os resultados apresentamos o valor médio das recompensas agregadas calculadas conforme uma combinação linear das recompensas individuais de cada secundário (Equação 3.8), e coletadas a cada rodada, com barras de erro correspondentes ao intervalo de confiança de 95%.

$$r_{AGREGGATE} = \sum_{j=1}^{J} \kappa \times r_{u_j} \tag{3.8}$$

onde κ é uma constante e r_{u_j} o valor da recompensa individual do secundário u_j.

Foram feitas 200 rodadas de simulação para cada conjunto de parâmetros, para o cenário anteriormente descrito, onde comparamos o desempenho das sequências listadas na Subseção 3.3.2. Os valores dos parâmetros foram escolhidos como os mais representativos dentro de seus intervalos de validade, estabelecidos após a realização de alguns testes, e os valores que eles assumem estão resumidos na Tabela 3.2, onde destacamos, na parte superior, os parâmetros variados em relação a avaliação realizada para o caso de um secundário (Subseção 3.2.3).

Nome	Valor	Conteúdo
$\#_{SEC}$	20	número máximo de secundários
T	$2 \times \#_{CH} \times \tau$	tamanho do *slot*
ε	$\varepsilon \in [0, 1]$	fator de investigação da ε-*greedy*
t	$t \in]0, \infty[$	temperatura do *softmax*
$Modelo\ de\ Canal$	*ON-OFF* Exponencial	comportamento do primário
β	$\beta \in [0, 1]$	meta-parâmetro para α
W_m	8	parâmetro para prob. de colisão [101]
κ	1	constante no cálculo da $r_{AGREGGATE}$
FHC	0.1, $FHC \in [0, 1]$	fator de homogeneidade dos canais
FVA	1.0, $FVA \in [1, 2]$	fator de variabilidade do ambiente
$\#_{CH}$	9	número máximo de canais
FU	U(0.1,0.9)	fator de utilização do canal
δ	0.95, $\delta \in [0, 1]$	fator de perda do `RL`
γ	0.0, $\gamma \in [0, 1]$	fator de desconto do `RL`
C_{MAX}	10	capacidade média máxima do canal
$\overline{C}$	$U(FHC * C_{MAX}, C_{MAX})$	capacidade média do canal
C_{INST}	$U(\overline{C} * (1 - FVA/2), \overline{C} * (1 + FVA/2))$	capacidade instantânea do canal
$\mathrm{Pr}_{MISDETECTION}$	0.0	prob. de falha na detecção do primário
$\mathrm{Pr}_{FALSEALARM}$	0.0	prob. de alarme falso de primário
$\#_{SLOTS}$	50.000	número de *slots*
$\#_{RUNS}$	200	número de rodadas

Tabela 3.2: Parâmetros da simulação e avaliação da proposta multiusuário (destaque, na parte superior, dos parâmetros variados em relação a avaliação realizada para o caso de um secundário (Tabela 3.1)).

Diferente do que fizemos na avaliação do mecanismo para o caso de um secundário (Subseção 3.2.3), nesta avaliação, consideramos apenas um modelo de ocupação dos canais pelos primários: o modelo *ON-OFF* exponencialmente distribuído. A nossa decisão foi baseada na inexistência de consenso sobre qual modelagem é mais representativa, conforme discutido na Subseção 2.5.1, e pela adoção do modelo *ON-OFF*

em muitos trabalhos [94]. Outro detalhe diferente nesse aspecto é que avaliamos a ocupação do canal não pela sua duração absoluta, mas pela relativa, representada pelo fator de utilização do canal pelo primário (FU), equivalente a probabilidade do canal estar indisponível.

Além disso, separamos a análise por contextos, sendo cada um detalhado nas próximas subseções. Os contextos utilizados para as análises foram:

- Parametrização do *Q-learning*;
- Quantidade de secundários;
- Quantidade de canais;
- Medida de justiça entre os secundários; e,
- Resultados comparativos.

Continuamos assumindo que nossa detecção de primário é precisa e isenta de erros, $\Pr_{MISDETECTION} = 0.0$ e $\Pr_{FALSEALARM} = 0.0$, pois entendemos que o foco do trabalho não é direcionado para o problema de detecção do primário, preferindo deixar essa análise para os trabalhos futuros.

Parametrização do *Q-learning*

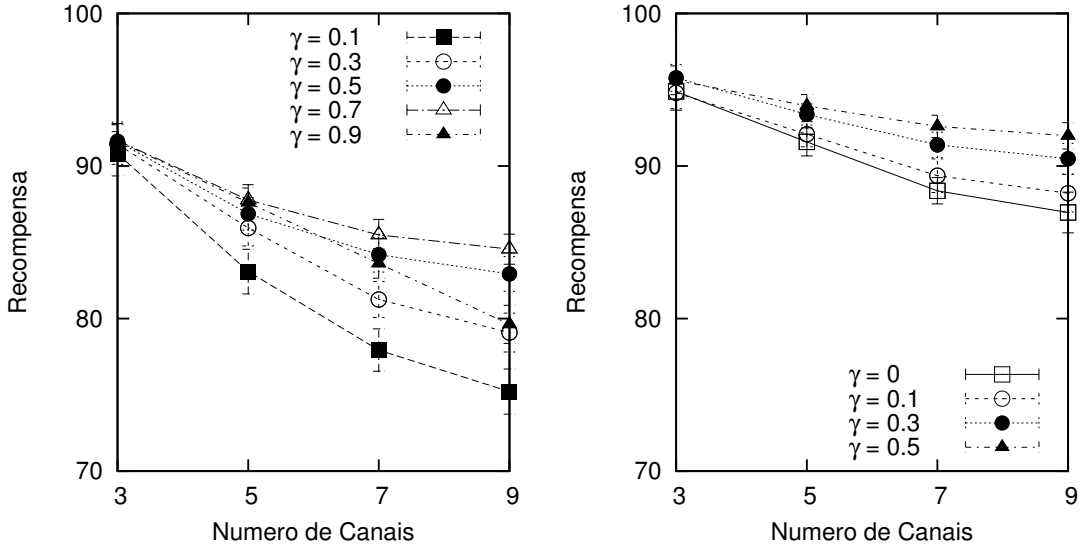
Realizamos alguns experimentos com o objetivo de obter o conjunto de parâmetros do nosso mecanismo baseado no *Q-learning* capaz de maximizar a quantidade de recompensa coletada no cenário dinâmico a que ele está submetido. Os parâmetros a que nos referimos são aqueles descritos na Seção 2.3: γ, α, e os referentes as estratégias, *softmax* e *ε-greedy*, respectivamente, a temperatura t e o fator de investigação ε.

Inicialmente, testamos alguns valores para o parâmetro γ, $\gamma \in [0, 1]$, chamado de fator de desconto. Esse parâmetro determina diretamente o grau de importância da recompensa que será coletada futuramente a partir de uma ação tomada no presente. Portanto, valores maiores de γ indicam que o agente baseia-se mais na recompensa futura que na imediata. Os resultados dessa avaliação são mostrados na Figura 3.10, que apresenta no eixo das ordenadas a recompensa obtida pelo nosso mecanismo conforme o número de canais.

Para esta análise foi necessário escolher uma estratégia de referência e um valor inicial para os demais parâmetros: escolhemos α, do próprio *Q-learning*, e ε, da estratégia *ε-greedy*. Nossa escolha foi baseada nos valores utilizados para esses mesmos parâmetros anteriormente (Seção 3.2.3). Deixamos para frente a análise da

influência do parâmetro β, responsável pelo ajuste dinâmico da taxa de aprendizado do nosso mecanismo.

É possível observar que as abordagens escolhidas (Figuras 3.10(a) e 3.10(b)) apontam para uma relação entre o crescimento de γ e o aumento da recompensa. Entretanto, pode-se notar que existe um limite para o valor desse parâmetro, entre 0.7 e 0.9, que reverte a tendência (Figura 3.10(a)). Uma observação do comportamento do parâmetro para valores mais próximos de 0.7 (Figura 3.10(a)) demonstra que os valores de recompensa coletados podem ser considerados iguais, devido ao intervalo de confiança, se γ estiver no intervalo [0.5,0.7]. Optamos, a priori, pelo valor menor, pois favorece a convergência rápida do mecanismo [27].

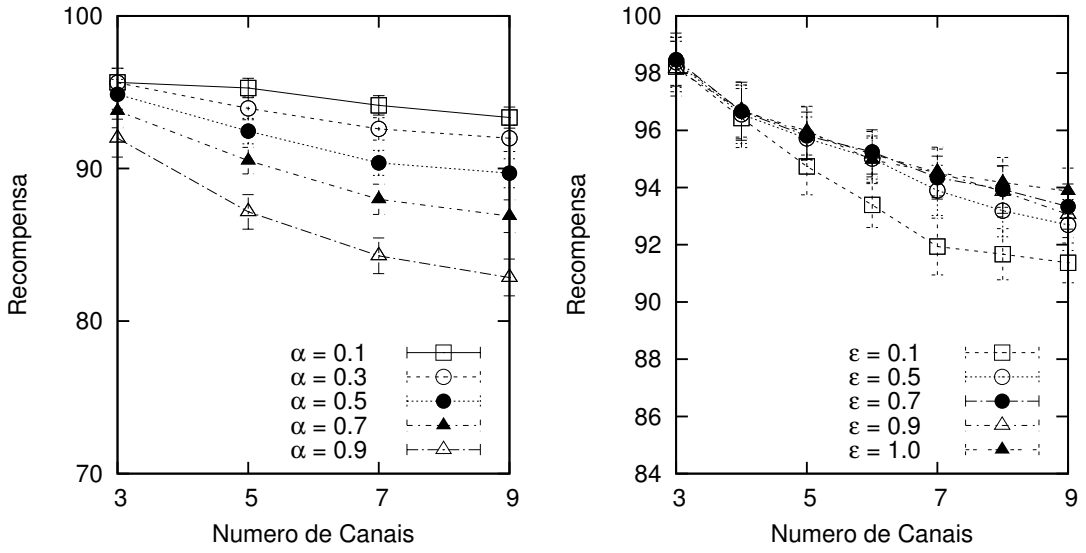


(a) α=0.1 e ε=0.7 (b) α=0.5 e ε=0.7

Figura 3.10: Impacto da variação do fator de desconto γ na nossa proposta (RL) com estratégia ε-*greedy* para 1 usuário secundário.

Em seguida, verificamos o impacto da variação dos parâmetros α e ε (Figura 3.11), onde o eixo das ordenadas também representa o percentual de recompensa obtida pelo nosso mecanismo quando comparado ao valor ótimo obtido por força bruta, conforme o número de canais. Começamos analisando a taxa de aprendizado α,$\alpha \in [0, 1]$. Da mesma forma que na análise anterior, era necessário escolher o valor inicial do demais parâmetros, γ e ε. O valor de γ obtivemos anteriormente e o de ε escolhemos novamente ser 0.7.

Como pode ser visto na Figura 3.11(a), valores menores para α proporcionam um melhor desempenho do mecanismo. Esperávamos que esse resultado demonstrasse a importância da adaptação do mecanismo a dinâmica do ambiente de RF e o quanto isso influenciaria no desempenho geral, porém constatamos que para nossa aplicação é mais importante valorizar o conhecimento adquirido, mantendo o parâmetro α



(a) $\gamma = 0.5$ e $\varepsilon = 0.7$ (b) $\gamma = 0.5$ e $\alpha = 0.1$

Figura 3.11: Impacto da variação do taxa de aprendizagem α e do parâmetro ε na nossa proposta (RL) com estratégia ε-*greedy* e 10 usuários secundários.

baixo.

O parâmetro ε, $\varepsilon \in [0, 1]$, também chamado de fator de investigação, está ligado a estratégia ε-*greedy* e estabelece a probabilidade que direciona o mecanismo para investigar novas ações a partir de um estado. A escolha adequada do fator de investigação depende muito do problema a ser tratado. Se para atingir o objetivo estabelecido o mecanismo necessitar armazenar um "conhecimento" maior, o valor desse parâmetro tende a ser maior. Por outro lado, se na solução do problema existir a execução de outras tarefas paralelamente a de aprendizagem, não parece ser adequado assumir que o mecanismo devesse investigar mais ações e, com isso, um compromisso entre a aprendizagem e o desempenho precisa ser considerado na escolha desse parâmetro.

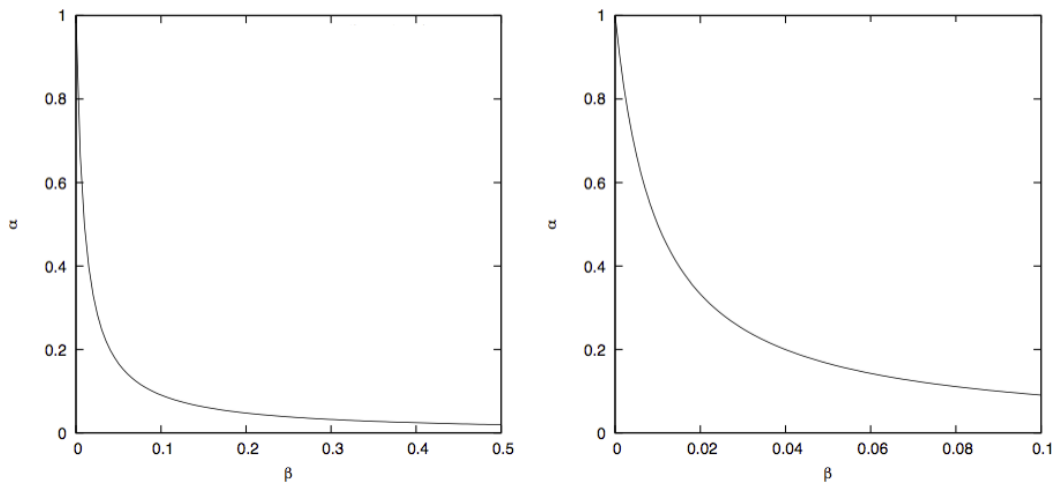
Observando a Figura 3.11(b), podemos notar que o parâmetro ε demonstra possuir maior influência na recompensa quando o seu valor está no intervalo [0.1,0.5]. Para valores maiores que 0.5, considerando o intervalo de confiança, a influência do parâmetro demonstra ser a mesma. Assim, considerando a possibilidade de inter-relacionamento entre os diversos parâmetros, o valor previamente escolhido para ε, 0.7, pode não ser o que possibilita o máximo desempenho do mecanismo.

A estratégia ***softmax*** é muito sensível a escolha adequada do parâmetro temperatura t, responsável pelo controle da investigação de novas ações. Se a escolha recair sobre um valor muito baixo, a investigação torna-se gulosa; e, caso contrário, aleatória.

Além da temperatura t, também é necessário analisar a influência do parâmetro

β, cujo valor está compreendido no intervalo [0,1] e que ajusta a taxa de aprendizagem do nosso mecanismo de modo dinâmico.

Observando a Figura 3.12(a), podemos notar que para β com valores acima de 0.1, a taxa de aprendizado assume valores no intervalo [0,0.1]. Como não desejamos que a taxa de aprendizado fique tão baixa, decidimos ceifar o limite superior do intervalo de valores possíveis para β, reduzindo-o de 1.0 para 0.1 (Figura 3.12(b)).



(a) (b)

Figura 3.12: Curva da taxa de aprendizado segundo o parâmetro β (Equação 3.6).

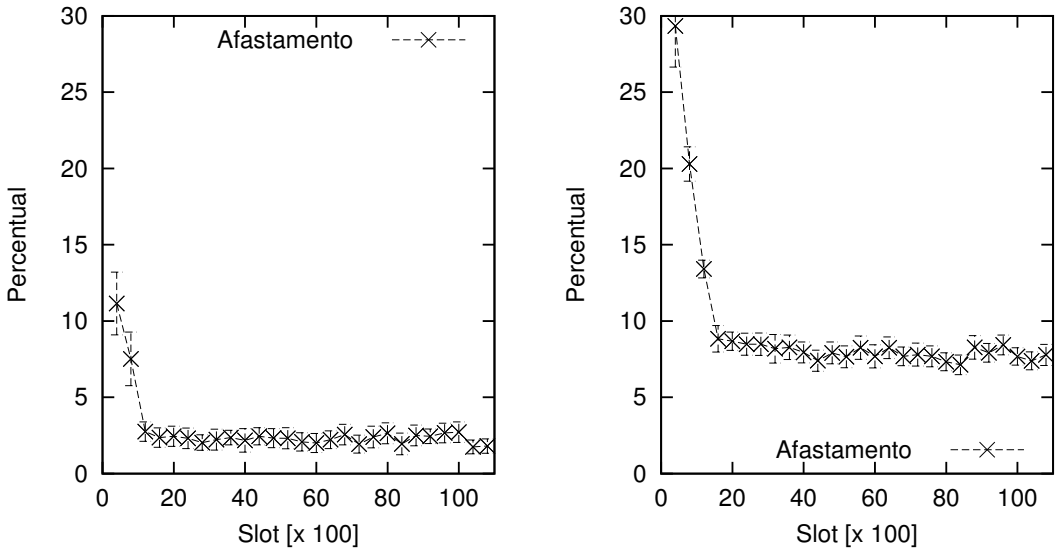
Assim, para chegarmos a um valor de β que obtivesse o melhor desempenho para o nosso mecanismo, em ambas as estratégias, realizamos previamente uma análise detalhada com outros valores possíveis para os demais parâmetros e escolhemos para ε o valor de 0.3 e para t_{ZERO} o valor de 1.000, fixando o valor de t_{FINAL} em 100.

Para auxiliar a busca do melhor valor de β, estabelecemos uma nova métrica, que mede o percentual de afastamento da recompensa obtida pelo nosso mecanismo do valor ótimo obtido por força bruta, a quem chamamos *afastamento*. Com isso, encontramos como melhor valor do parâmetro, $\beta = 0.08$. A Figura 3.13 mostra esse resultado.

Quantidade de Secundários

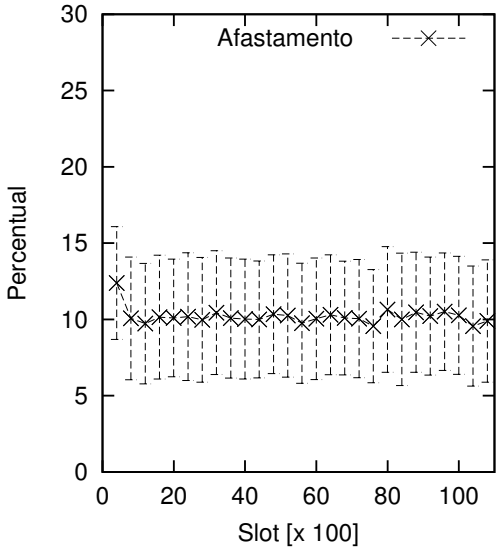
Nessa parte da avaliação, realizamos uma análise do impacto da variação na quantidade de secundários simultaneamente a do fator de utilização dos canais pelos primários (FU), ou seja, equivalente a probabilidade do canal estar indisponível.

Para isso, realizamos uma simulação com 8 canais para cada estratégia, *softmax* e *ε-greedy*. Essa quantidade foi escolhida pressupondo que em um cenário com mais canais devem existir mais "oportunidades" para serem aproveitadas pelo nosso mecanismo, avaliando indiretamente a sua eficiência. As figuras plotadas foram

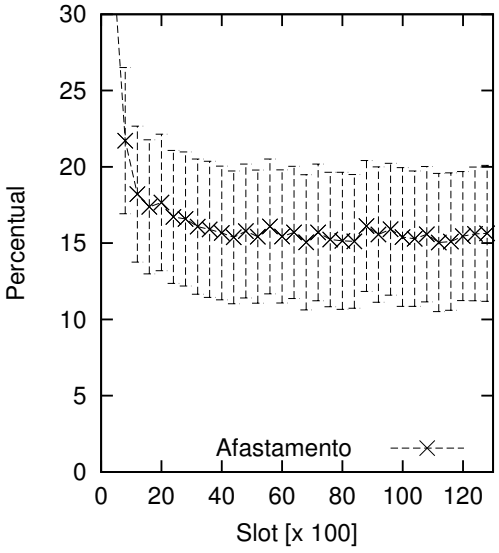


(a) 03 canais, *ε-greedy*.

(b) 10 canais, *ε-greedy*.



(c) 03 canais, *softmax*.

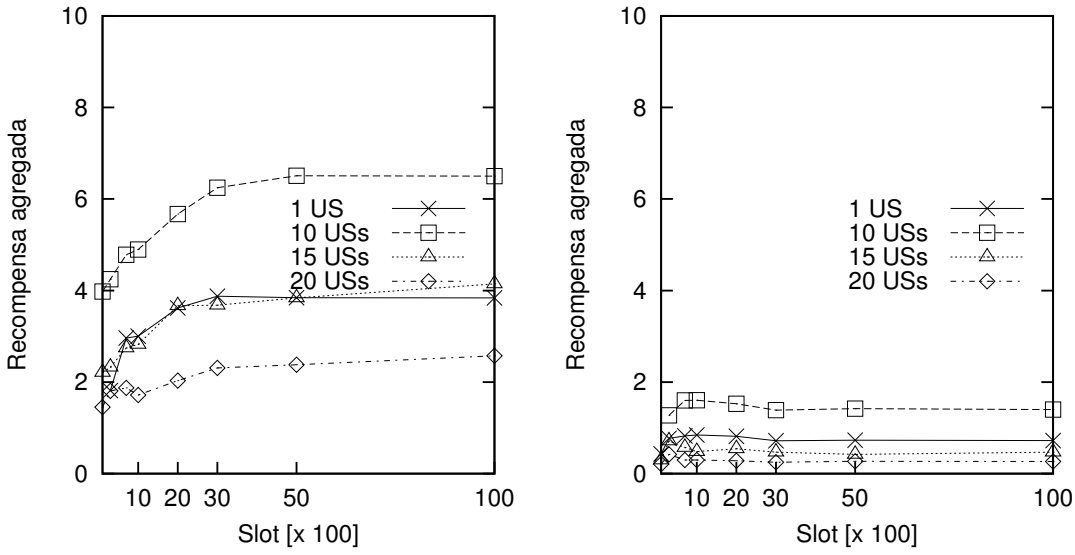


(d) 10 canais, *softmax*.

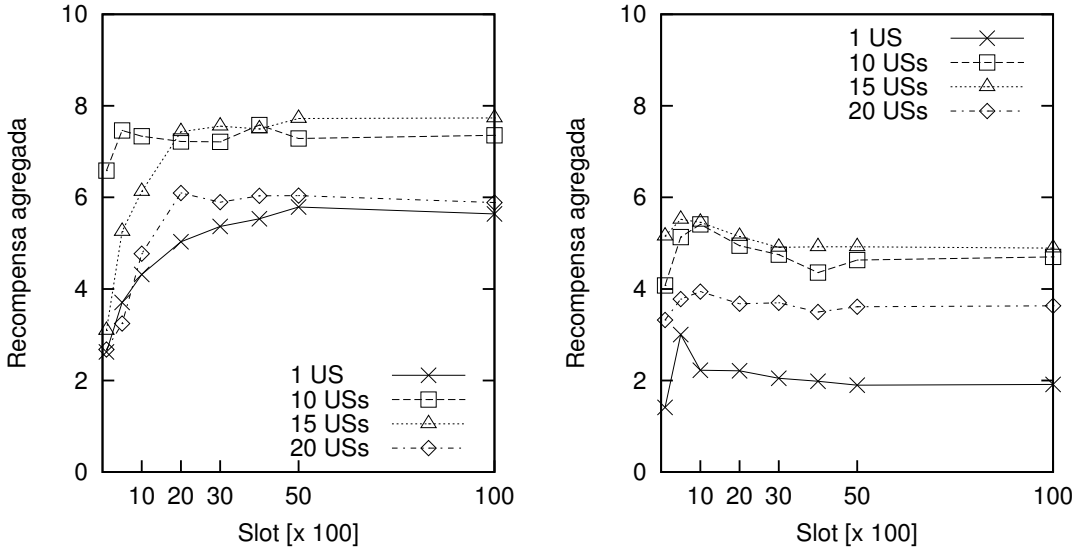
Figura 3.13: Escolha do parâmetro β na nossa proposta (`RL`), para ambas as estratégias.

ceifadas em 20% da quantidade total de *slots* para aproveitar a parte que contém mais informação.

Os resultados dessa avaliação são mostrados na Figura 3.14. Conforme esperado, é possível observar que com baixo fator de utilização do canal pelos primários (FU), a quantidade de recompensa coletada é maior (Figuras 3.14(b) e 3.14(d)), uma consequência direta da maior quantidade de "oportunidades" para os secundários. Neste resultado também é possível observar que existe um desempenho melhor do mecanismo para uma determinada quantidade de secundários, em ambas as estratégias. Podemos entender o fenômeno como resultado da acomodação dos



(a) $FU = 10\%$, *softmax* (b) $FU = 90\%$, *softmax*



(c) $FU = 10\%$ (d) $FU = 90\%$

Figura 3.14: Impacto da variação da quantidade de usuários secundários e do parâmetro fator de utilização (FU), na nossa proposta (RL), para ambas as estratégias, e com 9 canais.

secundários, consequência da maior disputa pelas "oportunidades" disponíveis, que não são suficientes para a demanda, acarretando uma redução do total quando a quantidade de secundários aumenta muito.

Quando o parâmetro fator de utilização do canal (FU) aumenta, espera-se uma redução na recompensa coletada, o que pode ser comprovado através dos resultados nas Figuras 3.14(a) e 3.14(c). Nesse cenário, também nota-se que existe um desempenho melhor do mecanismo para um determinada quantidade de usuários secundários, embora o valor seja diferente daquele existente quando há mais "opor-

tunidades" de uso do canal pelos secundários.

Outra observação importante é que a estratégia *ε-greedy* tem desempenho superior a *softmax* quando aplicada ao nosso problema, independente do percentual de utilização dos canais pelos primários. Isso de certa forma contradiz a teoria de que há uma preponderância da técnica *softmax* [27], demonstrando que em certos cenários e aplicações, a técnica *ε-greedy* pode sim ter um desempenho superior.

Os resultados obtidos com este experimento apontam também para uma necessidade de adaptação do nosso mecanismo a quantidade de secundários, sendo interessante que exista uma capacidade de restrição autônoma dessa quantidade pelo mecanismo, visando atingir uma recompensa maior. Isso será investigado em trabalhos futuros.

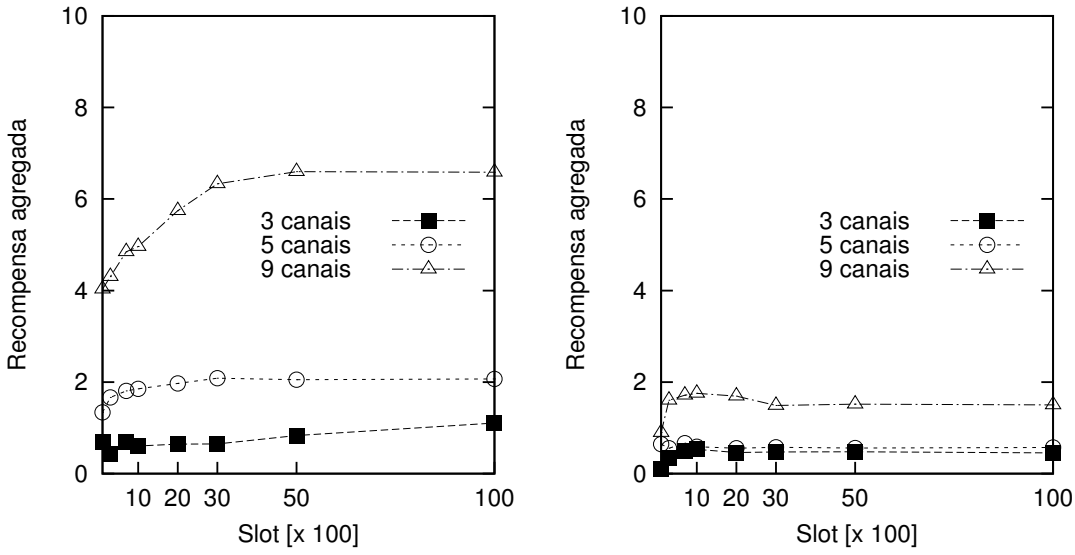
Quantidade de Canais

Nesta etapa foram avaliados os efeitos na recompensa em razão da variação da quantidade de canais simultaneamente a do fator de utilização do canal pelos primários(FU), equivalente a probabilidade do canal estar indisponível.

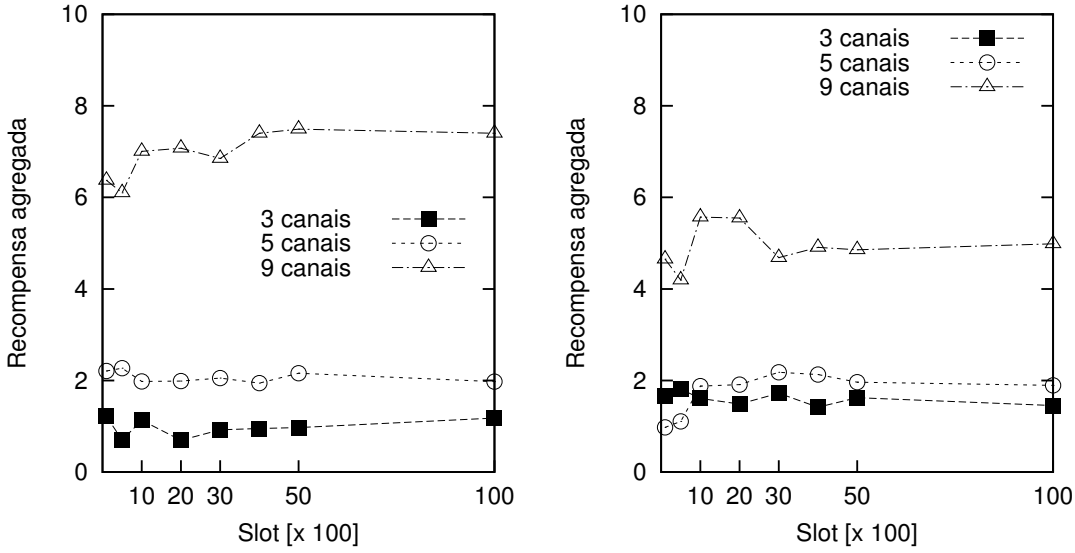
Para isso, realizamos uma simulação com 10 secundários para cada estratégia, *softmax* e *ε-greedy*. As figuras plotadas foram ceifadas em 20% da quantidade total de *slots*, para maior clareza.

Conforme esperado, para ambas as estratégias e para os dois percentuais avaliados do parâmetro fator de utilização (FU), os maiores valores de recompensa foram obtidos com a maior oferta de canais (Figura 3.15). Entretanto, com baixa disponibilidade de canais para os usuários secundários (Figuras 3.15(a) e 3.15(c)), o desempenho do mecanismo ficou semelhante para 3 canais ou 5 canais. A causa deste comportamento é que um percentual alto do parâmetro fator de utilização (FU) com um número reduzido de canais significa que há menos "oportunidades" para serem aproveitadas pelo mecanismo RL, bastando um pequeno aumento dessas "oportunidades" para que o desempenho do mecanismo aumente significativamente.

Para todas as configurações apresentadas, comprovamos que o mecanismo é imune a qualquer perturbação provocada pela variação, seja da quantidade de usuários, seja da quantidade de canais, tanto para uma maior indisponibilidade dos canais (maior FU) quanto para uma menor, e que a medida que o número de *slots* cresce o mecanismo evolui para um ponto de equilíbrio, que para o nosso cenário e aplicação, ficou próximo de 5.000 *slots*.



(a) $FU = 10\%$, *softmax* (b) $FU = 90\%$, *softmax*



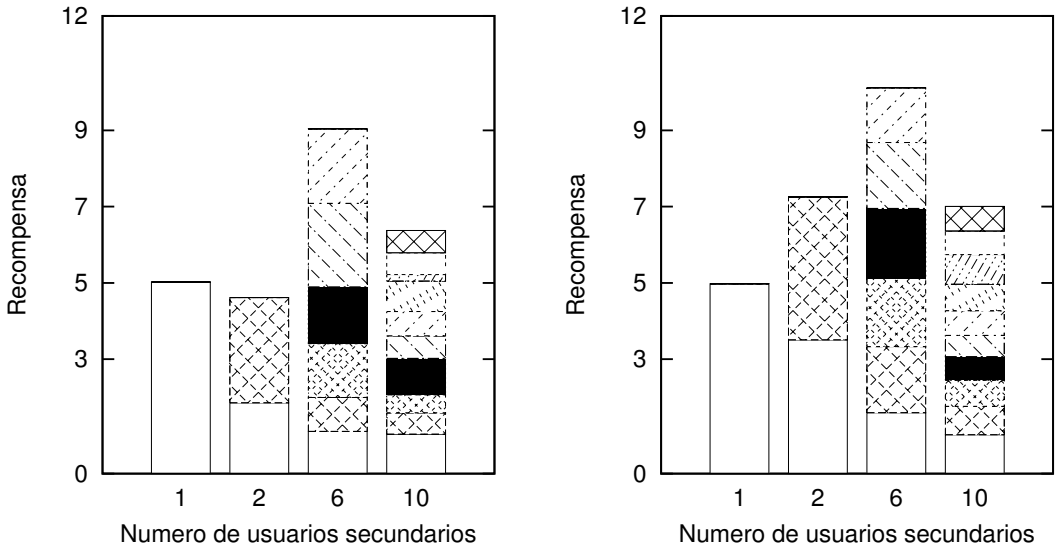
(c) $FU = 10\%$ (d) $FU = 90\%$

Figura 3.15: Impacto da variação da quantidade de canais e do parâmetro fator de utilização (FU), na nossa proposta (RL), com ambas as estratégias, e com 10 secundários.

Medida de Justiça

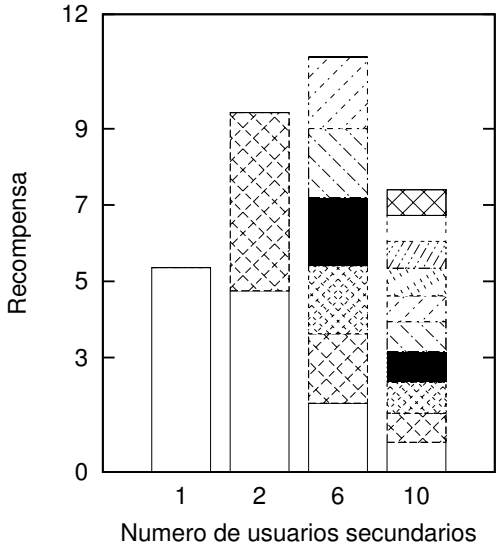
Outra avaliação que realizamos foi a medida de justiça entre os secundários (Figuras 3.16 e 3.17). A importância dessa análise se deveu ao fato de que observamos até agora os valores agregados da recompensa, que poderia estar mascarando algum comportamento egoísta do mecanismo RL na atribuição das "melhores" sequências, favorecendo alguns secundários em detrimento de outros.

Com esse resultado, constatamos que a medida que aumentamos a quantidade de *slots*, a recompensa final de cada secundário tende para uma fração homogê-



(a) Transiente até 100 *slots*

(b) Transiente até 1.000 *slots*



(c) Medição até 10.000 *slots*

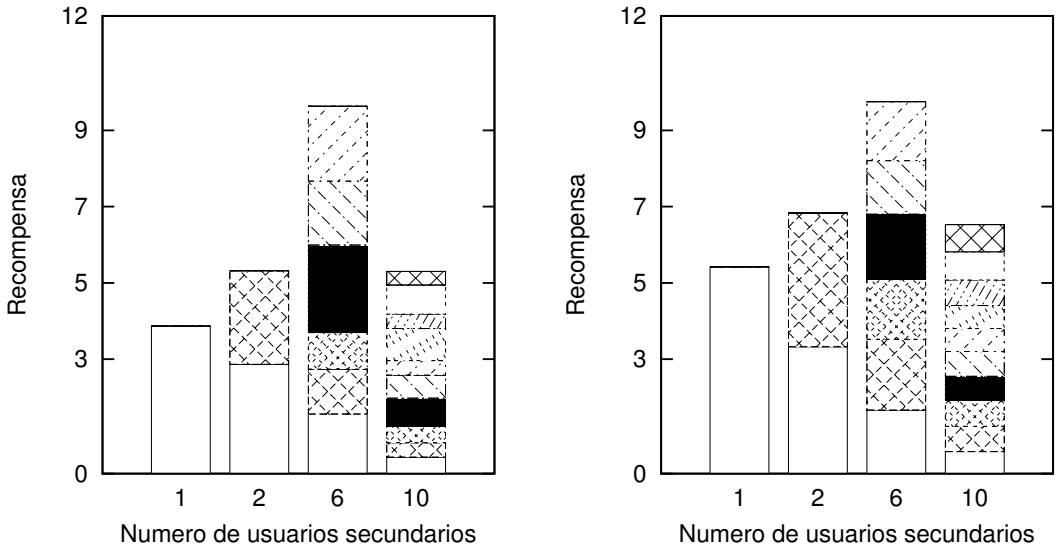
Figura 3.16: Medida de justiça na nossa proposta (RL), com estratégia ε-*greedy*, para 10 secundários, 9 canais e para fator de utilização (FU) de 90%.

nea da recompensa agregada, equivalente à dos demais secundários, não ocorrendo comportamentos egoístas em nenhuma das duas estratégias analisadas.

Resultados Comparativos

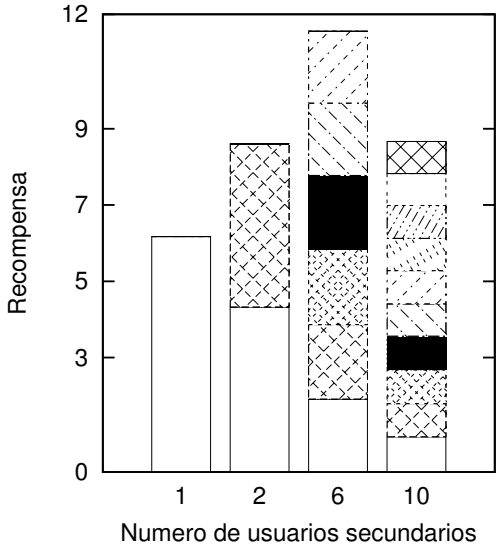
Na Figura 3.18, mostramos um retrato da problemática das colisões na rede secundária, analisando o funcionamento do nosso mecanismo para ambas as estratégias, *softmax* e ε-*greedy*.

Observamos que para o fator de utilização (FU) igual a 10% (Figura 3.18(a)), as colisões aumentam com o crescimento da quantidade de secundários, evidenciando



(a) Transitório até 100 *slots*

(b) Transitório até 1.000 *slots*



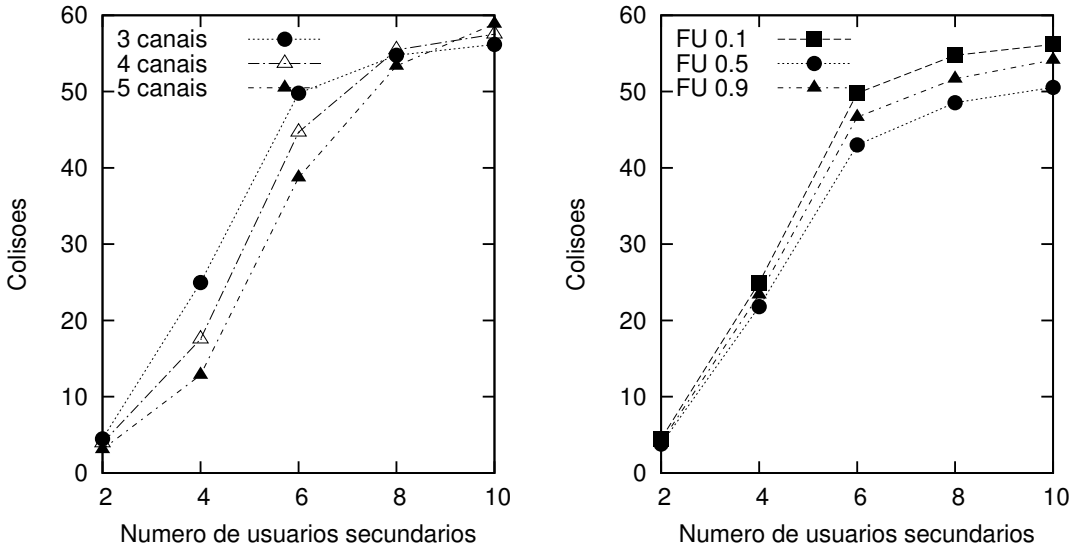
(c) Medição até 10.000 *slots*

Figura 3.17: Medida de justiça na nossa proposta (RL), com estratégia *softmax*, para 10 secundários, 9 canais e para fator de utilização (FU) de 90%.

uma intensificação da disputa pelo uso simultâneo de um canal. Porém, também se observa que as colisões são menores para uma quantidade maior de canais ofertados, o que era esperado, pois houve crescimento das "oportunidades" disponíveis proporcionalmente a quantidade de canais. Outro indicador da disputa na rede secundária pode ser observado na Figura 3.18(b), onde a curva superior é a de menor valor do parâmetro fator de utilização (FU), indicando a maior disponibilidade dos canais para ocupação pelos secundários.

Resumidamente, os resultados da nossa simulação mostraram que a probabilidade de colisão ($\Pr_{COLLISION}$), de fato, sofre influência de variados fatores, como as

probabilidades de disponibilidade dos canais ($\Pr_{CH-AV}$), a quantidade de canais e também, a quantidade de secundários. Além disso, verificamos que a $\Pr_{COLLISION}$ comporta-se de modo inversamente proporcional a quantidade de canais e diretamente proporcional tanto aos valores da $\Pr_{CH-AV}$ quanto a quantidade de secundários.



(a) $FU = 10\%$ (b) 3 canais

Figura 3.18: Problemática das colisões na rede secundária.

A partir do resultado da análise paramétrica do *Q-learning*, ajustamos o nosso mecanismo e realizamos uma comparação da sequência fornecida por ele, *RL*, com as sequências obtidas através de outros mecanismos de ordenação, definidos na Subseção 3.3.2.

Os resultados dessa avaliação são mostrados nas Figuras 3.19 e 3.20. É possível observar para ambas as estratégias testadas, *softmax* e ε-*greedy*, que a sequência fornecida pela nossa proposta, *RL*, apresenta resultados melhores. O desempenho das outras sequências é desfavorável, pois nenhuma delas utiliza uma regra de parada baseada na predição do desempenho estimado de se continuar sensoreando os próximos canais da sequência, ou seja, nos outros mecanismos a regra utilizada é parar no primeiro canal sensoreado como "livre", presumidamente com resultados melhores, segundo a literatura [19]. Desta forma, o nosso *RL*, que utiliza o conhecimento armazenado na *Q-table*, adquirido das experiências passadas, consegue determinar de maneira eficiente se é vantajoso ou não utilizar imediatamente um canal sensoreado como "livre".

Uma observação interessante a respeito das curvas das Figuras 3.20(a), 3.20(b) e 3.19(a), 3.19(b), é que o desempenho da sequência `Prob`, muito próximo ao da `Aleatória`, é inferior ao apresentado pela sequência `Cap`. Isso indica que nesse ce-

nário, a diferenciação entre as capacidades médias dos canais ($\overline{C}$) é mais importante do que a diferenciação entre as suas probabilidades de disponibilidade ($\Pr_{CH-AV}$). Com isso, é melhor ordenar os canais pela ordem decrescente de suas capacidades médias, pois aumenta-se a probabilidade do primeiro canal sensoreado como "livre" ter maior capacidade.

Outro detalhe é que a curva referente a nossa proposta, RL, apresenta um crescimento menor conforme aumenta a quantidade de secundários, indicando que há um limiar de saturação da capacidade disponível da rede secundária, conforme a quantidade de canais existentes.

A partir das Figuras 3.20(c) e 3.19(c), podemos verificar que com poucos secundários, a recompensa obtida se altera pouco com a variação do fator de utilização (FU) e da quantidade de canais. Quando aumentamos a quantidade de secundários (Figuras 3.20(d) e 3.19(d)), observamos um aumento da recompensa coletada, embora possamos notar que a variação no parâmetro fator de utilização (FU) é pequena, mantendo um crescimento proporcional da recompensa com o aumento da quantidade de canais existentes.

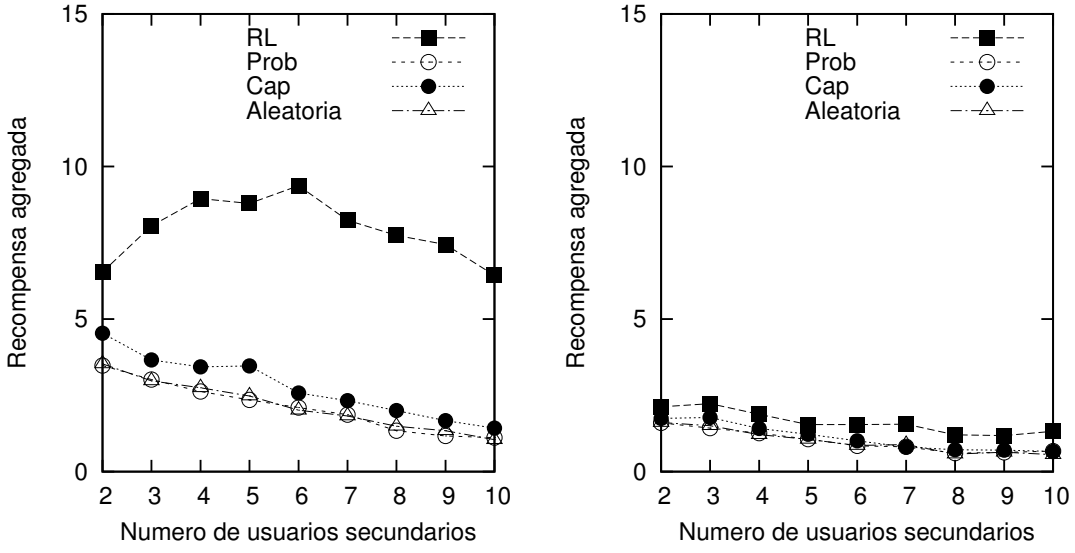
Finalmente, para os cenários avaliados, verificamos que a estratégia ε-*greedy* apresentou um resultado superior a *softmax*, contrariando a expectativa tradicional [27].

3.4 Conclusões

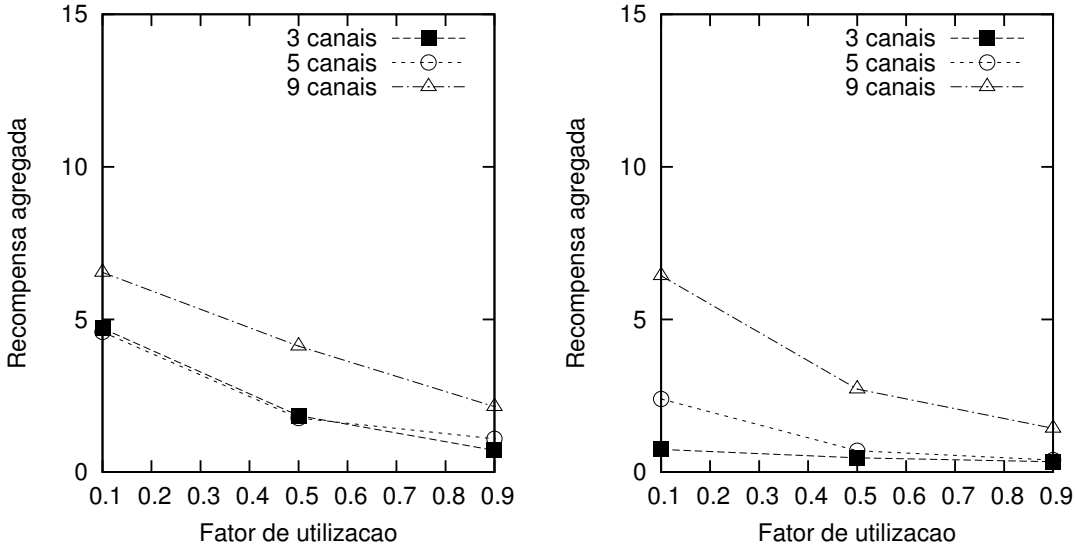
Neste capítulo, analisamos o problema da escolha da ordem de sensoreamento de canais em um sistema multiusuário, onde cada usuário é capaz de realizar o sensoreamento em apenas um canal por vez a fim de detectar oportunidades de uso. A ordem de sensoreamento indica a sequência de canais sensoreados pelos usuários secundários na busca por um canal disponível para ser "ocupado" e efetivamente utilizado. Nesses casos, a sequência utilizada para sensorear os canais pode ter grande impacto no desempenho.

Na nossa abordagem, consideramos uma rede de rádios cognitivos multicanal onde os canais não estão sempre disponíveis, devido as atividades dos primários e propomos uma solução de baixa complexidade que utiliza uma máquina de aprendizado por reforço baseada na técnica *Q-learning*. Essa solução não requer o conhecimento prévio das probabilidades de disponibilidade nem das capacidades médias esperadas em cada canal, podendo se adaptar dinamicamente às variações dessas estatísticas. Além disso, ao possuir baixa complexidade, essa solução torna-se também atrativa para ser embarcada em rádios cognitivos.

Os resultados das simulações mostraram que o mecanismo proposto obtém desempenho superior aos outros tipos de ordenamento que foram avaliados quando



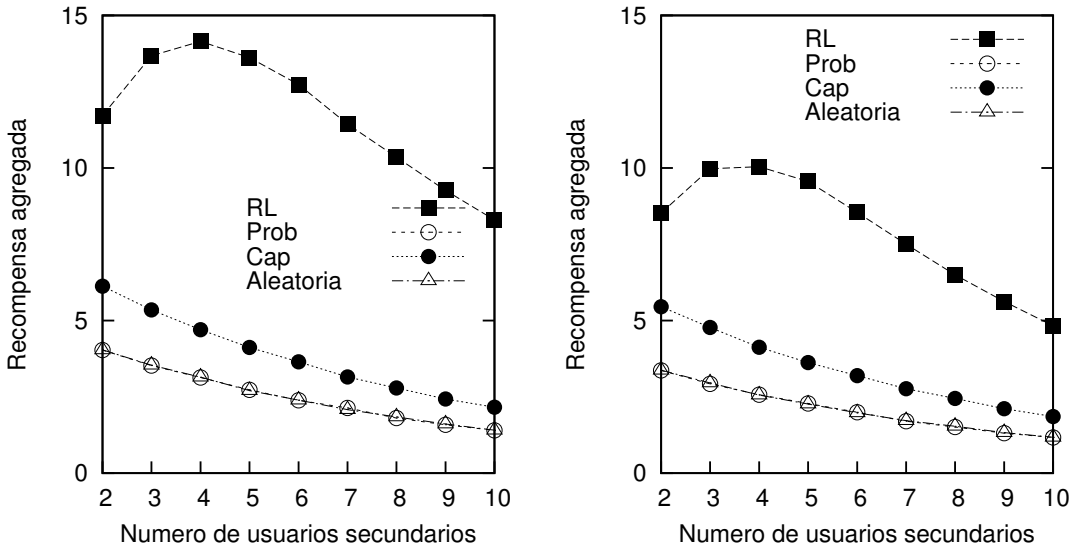
(a) $FU = 10\%$, *softmax* (b) $FU = 90\%$, *softmax*



(c) 2 secundários, *softmax* (d) 10 secundários, *softmax*

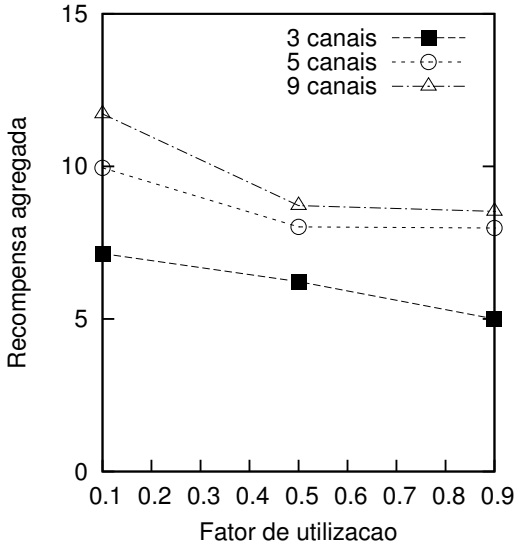
Figura 3.19: Resultados para a sequência obtida através da nossa proposta (`RL`) com estratégia *softmax*, comparada com as sequências seguindo a ordem decrescente das probabilidades de disponibilidade dos canais (`Prob`), a ordem decrescente de capacidades médias (`Cap`) e a ordem aleatória (`Aleatória`).

variamos o número de usuários secundários, para diferentes valores do fator de utilização dos canais pelos usuários primários. Outra observação importante mostra a inexistência de secundários gananciosos e que, ao variarmos a quantidade de canais e aumentarmos a probabilidade de indisponibilidade dos canais, há um limiar máximo da recompensa que pode ser coletada, onde mesmo que se aumente a quantidade de canais ofertados, o valor da recompensa obtida não acompanha na mesma proporção.

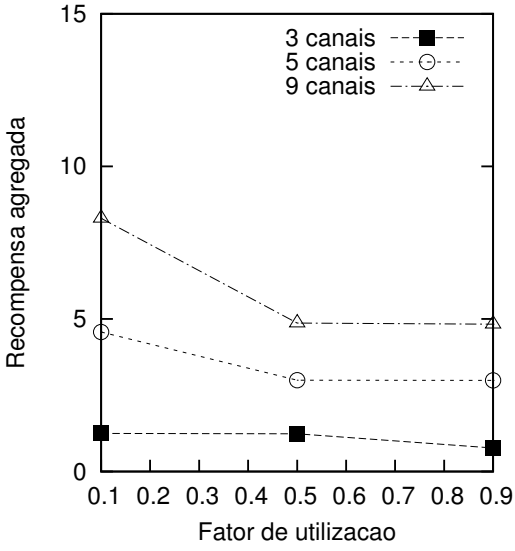


(a) $FU = 10\%$

(b) $FU = 90\%$



(c) 2 secundários



(d) 10 secundários

Figura 3.20: Resultados para a sequência obtida através da nossa proposta (`RL`) com estratégia *ε-greedy*, comparada com as sequências seguindo a ordem decrescente das probabilidades de disponibilidade dos canais (`Prob`), a ordem decrescente de capacidades médias (`Cap`) e a ordem aleatória (`Aleatória`).

Capítulo 4

Novas Estratégias e Análise da Convergência do Mecanismo Proposto

Neste capítulo, propomos novas estratégias para o balanceamento do dilema investigação-exploração e, em seguida, avaliamos e comparamos a nossa proposta com outras da literatura [78]. Logo após, apresentamos a análise da convergência do nosso mecanismo multiusuário, descrito no Capítulo 3, bem como, realizamos recomendações para a seleção e o ajuste dos seus parâmetros de modo a melhorar a sua convergência.

4.1 Novas Estratégias

Um dos principais desafios que se apresentam no desenvolvimento das técnicas de aprendizado é o compromisso entre investigação (*exploration*) e exploração (*exploitation*) [27].

Em geral, o processo de investigação das possíveis ações que podem ser alcançadas a partir de um estado conduzem a um aumento do conhecimento sobre a importância individual de cada ação para a coleta de maiores recompensas no longo prazo, mas aumentam também o risco de serem coletadas recompensas menores, face ao desconhecimento inerente ao processo. Em contrapartida, lembrando que o modelo preciso do ambiente pode ser desconhecido (Seção 2.3), a exploração do conhecimento adquirido talvez possa conduzir para uma seleção sub-ótima da próxima ação a partir de um estado, caso a função de utilidade que leva para a ação ótima esteja subestimada. Como consequência, surge o dilema entre investigação (*exploration*) e exploração (*exploitation*), que demonstra que nem a investigação de novas ações nem a exploração das "melhores" ações conhecidas devem ser realizadas

exclusivamente, visando o melhor desempenho no longo prazo.

No nosso trabalho, implementamos duas estratégias clássicas para mitigar esse dilema. A primeira é a *ε-greedy* [27], que tenta fazer com que todas as ações, e seus efeitos, sejam experimentados igualmente. A outra estratégia clássica, é chamada *softmax* [27], onde a probabilidade de escolha de uma determinada ação varia conforme o valor correspondente do *Q-value*. Ambas as estratégias estão descritas na Seção 2.3.

Uma desvantagem da estratégia *ε-greedy* é que a probabilidade de escolha das ações durante a investigação segue uma distribuição uniforme, permitindo que uma ação (ruim) com recompensa baixa seja escolhida com a mesma probabilidade que uma ação (boa) com alta recompensa. O *softmax* vem resolver esse problema, atribuindo uma ponderação para cada ação a partir do estado corrente, conforme o seu valor estimado de recompensa, obtido do seu *Q-value*. Assim, uma ação é escolhida de fato a partir desse valor ponderado, minimizando a probabilidade de escolha de uma ação "ruim" em face de uma ação"boa", o que é desejado, especialmente quando essa escolha puder causar uma consequência desastrosa.

Comparando essas duas estratégias, não podemos afirmar qual seria a melhor em uma análise de longo prazo [27], pois a natureza das tarefas necessárias a solução do problema é que vai determinar como cada estratégia influencia o aprendizado (e o desempenho da métrica desejada), e dependendo do problema a ser resolvido, existem também fatores subjetivos envolvidos.

Um outro aspecto existente e pouco comentado refere-se utilização de uma distribuição de probabilidades com comportamento caótico (*chaotic exploration*) ao invés da forma mais usual, baseada em uma distribuição uniforme, para a realização do processo de investigação das ações, em qualquer estratégia, conforme mostrado no trabalho em [111] com resultados promissores. Porém, essa discussão deixamos para os trabalhos futuros.

Nesta seção, em complemento ao trabalho realizado, e decorrente dos resultados obtidos nas avaliações das propostas dos nossos mecanismos, são apresentados os conceitos utilizados na elaboração de duas variações para a estratégia *softmax* e da proposta de uma nova estratégia para o balanceamento do dilema investigação-exploração existente. Em seguida, realizamos uma avaliação comparativa com outras soluções obtidas da literatura.

Softmax Investigativo - SI

Na estratégia *softmax* [27], a probabilidade de escolha de uma determinada ação varia conforme o valor correspondente do *Q-value*, seguindo uma distribuição de probabilidades, a qual normalmente é a de *Boltzmann*, ou *Gibbs* (Equação 2.2).

Essa distribuição possui o parâmetro t, chamado temperatura, que estabelece a forma de escolha das ações. Quando t é alto, as ações são escolhidas com semelhantes probabilidades, que tendem a serem iguais, quando $t \to \infty$. No caso contrário, as ações com maior valor estimado de recompensa possuem maiores probabilidades de escolha e, no limite, quando $t \to 0$, a melhor ação é sempre a escolhida (como na estratégia ε-*greedy*).

Assim, no *softmax* clássico, a quantidade de "investigação" que será praticada já está embutida no próprio mecanismo, sendo um compromisso assumido através do parâmetro temperatura. Desta forma, não é possível ter um comportamento gananciosa simultaneamente a manter a investigação equiprovável (imitando a estratégia ε-*greedy*).

Para realizar esse ajuste mais refinado do mecanismo, criamos a estratégia híbrida, chamada "*softmax* investigativo", na qual introduzimos um novo parâmetro que funciona como um limiar para determinar os instantes de decisão onde as ações a serem tomadas seguirão uma distribuição uniforme de probabilidades, mesmo se t assumir um valor baixo.

Embora tal modificação aparentemente descaracterize a principal vantagem do *softmax* clássico, de fato, isso não ocorreu para a nossa aplicação do mecanismo, onde os cenários de emprego possuem canais de comunicação, que variam dinamicamente suas capacidades instantâneas (em razão dos efeitos de sombreamento, por exemplo), e também, variam suas disponibilidades (em razão do comportamento do primário), dentro de cada *slot* (Seção 2.3).

Softmax Ganancioso - SE

Uma estratégia puramente gananciosa (*greedy*) escolheria sempre a ação a que maximizasse $Q(s, a)$.

Nesse trabalho, implementamos a estratégia ε-*greedy*, que regula esse comportamento rígido através de um limiar (ε), permitindo que em algumas ocasiões, sejam escolhidas outras ações além da gananciosa. E de forma intuitiva, permitimos um ajuste desse limiar após a progressão do mecanismo ao longo dos episódios, realizando no início mais investigação das ações em busca de melhores estimativas de recompensa e assumindo as ações mais gananciosas no final.

Um problema da estratégia gananciosa é que todas as ações, exceto a melhor ($max_{a \in \mathbb{A}} Q(s, a)$), são tratadas da mesma forma. Assim, se existissem duas ações cujas recompensas coletadas fossem semelhantemente altas e as demais reconhecidamente com menor retorno, seria razoável que a seleção recaísse entre as duas ações com maior retorno, direcionando o esforço para identificar e escolher a melhor dentre essas duas ações, ao invés de realizar a investigação sobre todo o conjunto, incluindo

as muitas de menor retorno.

Após a avaliação realizada na Subseção 3.3.3, observamos que a estratégia gananciosa apresentou um desempenho superior a do *softmax*, especialmente em relação a métrica recompensa.

Assim, na tentativa de explorar esse resultado e melhorar o desempenho geral do nosso mecanismo, criamos uma estratégia híbrida, chamada "*softmax* ganancioso", fruto da adaptação da nossa implementação da estratégia gananciosa (*ε-greedy*), permitindo uma investigação das ações seguindo uma política regulada pela estratégia *softmax* ao invés de uma escolha equiprovável.

WinTable and Epsilon-greedy tied Strategy - StEaW

As estratégias clássicas e as suas variações, permitem um ajuste mais flexível do nosso mecanismo na medida que temos um conjunto de estratégias a serem experimentadas e também comparadas, para que seja utilizada a melhor delas. Contudo, observamos que alguns mecanismos simples de aprendizado, também possuem um desempenho elevado, conforme mostrado no trabalho em [112].

Diante dessa observação, passamos a analisar o comportamento do mecanismo de janela aplicado ao nosso cenário, cujo funcionamento é o seguinte: inicialmente, um vetor com o número de elementos equivalente ao de canais de comunicação existentes é criado e, em seguida, ele é utilizado para armazenar a taxa de sucesso na utilização do referido canal ao longo de uma janela cujo tamanho equivale a um valor pré-definido de *slots*. Por fim, esse vetor é ordenado ao final do período de duração da janela, segundo a ordem decrescente dos valores nele armazenados.

Repare que o nosso mecanismo (Subseções 3.2.2 e 3.3.2) não utiliza sempre o primeiro canal "livre" de primários, pelo contrário, a cada instante de decisão, ele estabelece uma comparação entre a possível *recompensa* em permanecer nesse canal "livre", com uma estimativa da *recompensa* (armazenada em sua *Q-table*) a ser coletada futuramente para as *ações* a partir daquele *estado*, ou seja, prosseguindo para outros canais correspondentes as demais posições na ordem de sensoreamento. Desta forma, no mecanismo de janela anteriormente descrito, o vetor não é um simples contador, pois armazena a taxa de sucessos como uma função de fatores diretos, como a taxa de transmissão efetiva, e também indiretamente contabilizados, como a disponibilidade dos canais.

Assim, criamos uma estratégia híbrida, chamada "*WinTable and Epsilon-greedy tied Strategy - StEaW*", onde a partir da comparação com um limiar, escolhe-se o canal com a maior taxa de sucessos ou então prossegue-se em uma estratégia gananciosa baseada na *ε-greedy*.

Algoritmo e Complexidade

O funcionamento do mecanismo completo é descrito em detalhes no Algoritmo 3.

Os passos 3 e 4 deste algoritmo correspondem a *fase de inicialização*, onde todos os pares *estado-ação* da *Q-table* são completados com zeros. Realizada a inicialização, começa a *fase de aprendizado*, que é repetida durante todo o período de funcionamento do mecanismo no episódio (equivalente a um *slot*).

No passo 8, ampliado no Algoritmo 4, é realizada a seleção da estratégia a ser adotada, conforme descrito na Subseção 4.1.

Após a execução da *ação*, o mecanismo torna-se capaz de calcular a *recompensa* obtida e atualizar o correspondente *Q-value*.

Em seguida, no passo 12, apresenta-se uma característica importante da nossa proposta, que diz respeito ao uso dos canais sensoreados como livres. A nossa proposta utiliza um critério de parada que consiste em comparar a *recompensa* atual, r_i, com o melhor *Q-value* das *ações* possíveis a partir daquele *estado* (passo 16) (Equação 3.2). Assim, é possível estimar se a *recompensa* do canal "livre" atual é superior a esperada se tomada a melhor *ação* existente. Repare que mesmo no caso onde o canal "livre" não é utilizado, o *Q-value* referente aquela *ação* também é atualizado

Caso o canal esteja "ocupado", o fator de perda δ é empregado para reduzir o *Q-value* referente aquela *ação* (passos 22 a 25).

Complexidade

Determinamos a eficiência do nosso algoritmo a partir da análise teórica da complexidade de pior caso para o tempo de execução ($T(n)$) e para a utilização de recursos ($S(n)$).

- *Eficiência Temporal*: observando o Algoritmo 4, é possível notar a repetição dos passos de 6 a 26 durante todo o funcionamento do mecanismo, correspondente a *fase de aprendizado*.

  Esse laço guiado pelo valor da variável *fim_do_sensoreamento* pode se repetir, no pior caso, até $N - 1$ vezes, onde N é a quantidade máxima de canais.

  Dentro desse laço, existem operações simples, que consomem 1 unidade de tempo de execução cada, tornando a complexidade de tempo constante, $\mathcal{O}(1)$.

  Desta forma, nosso algoritmo manteve a complexidade de $T(n) = \mathcal{O}(N)$.

- *Eficiência Espacial*: o maior consumo de recursos continua relacionado ao armazenamento da *Q-table*, que é uma matriz de dimensões $N^2 \times N$ ($estados \times$

ações).

Contudo, dentro do laço existente entre os passos 6 e 26, precisamos ter armazenado em memória apenas as ações possíveis a partir de um estado, ou seja, no pior caso, $N - 1$ ações. Assim, $S(n) = \mathcal{O}(N)$.

Algoritmo 3: Mecanismo proposto baseado em aprendizado por reforço.

```
1  fim_do_sensoreamento = 0;
2  /* inicialização da Q-table */
3  foreach s ∈ S, a ∈ A do
4      Q(s,a) = 0;
5  sorteia número aleatório x entre 0 e 1;
6  while ! fim_do_sensoreamento do
7      /* aprendizado */
8      switch ESTRATEGIA do
9          /* seleção da estratégia para a */
10         /* escolha da ação a */
11         /* a partir do estado s */
12     if (canal livre) then
13         /* canal c_a correspondente a ação a */
14         calcula recompensa r_t(s,a);
15         Q_{t+1}(s,a) ← (1 − α)Q(s,a) + α [r(s,a) + γ max_{a∈A} Q(s_{t+1},a)[;
16         if r_t(s,a) > max_{a∈A} Q(s',a') then
17             /* usa canal c_a */
18             fim_do_sensoreamento = 1;
19         else
20             /* não usa canal c_a */
21             continua sensoreamento;
22     else
23         /* canal c_a ocupado */
24         Q_{t+1}(s,a) ← δQ_t(s,a);
25         continua sensoreamento;
26     s_t = s_{t+1};
```

4.1.1 Implementação

Afora a parte referente ao acréscimo das "novas" estratégias, aproveitamos a implementação já realizada para o nosso simulador, conforme descrito nas Subseções 3.2.2 e 3.3.2.

Simulador Próprio

Algoritmo 4: Seleção de estratégias do mecanismo proposto.

```
1  /* aprendizado */
2  switch ESTRATEGIA do
3      case ε−greedy
4          if (x < ε) then
5              /* investigação → seleciona uma ação a aleatoriamente */
6          else
7              /* exploração → escolhe a ação a que possua o maior Q-value para
                   o estado atual s */
8      case softmax
9          /* calcula a probabilidade da ação a usando a distribuição de
               Boltzmann */
10         /* ordena as probabilidades */
11         /* seleciona a ação de maior probabilidade a partir do estado atual s
               */
12     case Softmax Investigativo - SI
13         /* seleciona uma ação a */
14         if (x < LIMIAR) then
15             /* Aleatoriamente → Investigação */
16         else
17             /* segundo a softmax */
18     case Softmax Ganancioso - SE
19         /* seleciona uma ação a */
20         if (x < LIMIAR) then
21             /* segundo a softmax */
22         else
23             /* exploração → escolhe a ação a que possua o maior Q-value para
                   o estado atual s */
24     case WinTable and Epsilon-greedy tied Strategy - StEaW
25         /* seleciona uma ação a */
26         if (x < LIMIAR) then
27             /* a ação a com maior taxa de sucessos */
28         else
29             /* segundo a ε−greedy */
```

Através do nosso simulador, onde cada um dos usuários da rede secundária utiliza sequências de sensoreamento individuais, foram avaliadas as seguintes ordens de sensoreamento, obtidas da literatura, e adaptadas para o nosso modelo de sistema (Seção 3.1):

- `FB`: ordem ótima dos canais obtida por força bruta, apenas para o caso de um usuário secundário [39];
- `RL`: ordem adaptativa dos canais fornecida pela nossa proposta;

- `Prob` $\times$ `Cap`: ordem dos canais na sequência decrescente do produto de suas capacidades médias ($\overline{C}$) pelas suas respectivas probabilidades de disponibilidade (Pr_{CH-AV});

- `Prob`: ordem de canais na sequência decrescente das probabilidades de disponibilidade de cada canal (Pr_{CH-AV});

- `Cap`: ordem de canais na sequência decrescente das suas capacidades médias ($\overline{C}$);

- `Aleat`: ordem de canais na sequência definida através de sorteio, utilizando uma distribuição uniforme;

- `Adapt`: ordem adaptativa dos canais que acompanha, conforme o resultado da comparação de uma variável aleatória com um limiar, a sequência `Aleat` ou realiza a investigação de novos canais [112];

- `Win`: ordem adaptativa dos canais que acompanha, conforme o resultado da comparação de uma variável aleatória com um limiar, a sequência daqueles com a maior ocorrência do estado "livre" em uma janela deslizante de x *slots* ou realiza a investigação de novos canais [112];

- `LS`: ordem de canais na ordem fornecida por um *quadrado latino* [1] [78]; e,

- `GP`: ordem de canais fornecida por um mecanismo, chamado *Gamma Persistent*, que analisa os sucessos (ocorrência do estado "livre") e falhas (ocorrência do estado "ocupado") dos canais nos *slots* anteriores, criando uma lista ordenada na sequência decrescente daqueles com maior incidência de sucessos [78].

Em destaque, a sequência *Prob*, que é considerada ótima para o cenário com somente um secundário [19] e sem a utilização de modulação adaptativa, e a sequência *Cap*, que ao seguir a ordem decrescente das capacidades médias dos canais, intuitivamente conduz a uma boa expectativa do seu desempenho.

A sequência `Aleat` está presente apenas como referência, devendo ser o limite inferior de desempenho para qualquer mecanismo.

Vale ressaltar que todas as sequências acima, com exceção das seguintes: `RL`, `Adapt`, `Win` e `GP`, são *estáticas*, ou seja, não mudam durante toda a simulação. No caso da sequência `RL`, devido ao próprio aprendizado por reforço, ela pode variar durante a simulação.

[1] O quadrado latino (*latin square*) é uma matriz quadrada preenchida com diferentes números (ou símbolos quaisquer), que ocorrem apenas uma vez em cada linha ou coluna.

Finalmente, todas as sequências, exceto a `RL`, `Adapt`, `Win`, `GP` e `Aleat`, assumem o conhecimento a priori das capacidades médias de cada canal ($\overline{C}$) e/ou de suas probabilidades de disponibilidade (Pr_{CH-AV}).

Modelo de Simulação

Para essa análise, resolvemos estabelecer mais 2 cenários diferenciados para as características físicas dos canais, com a finalidade de realizar uma maior aproximação com a realidade, em adição ao cenário já avaliado anteriormente (cenário I). Os cenários são:

I) Os parâmetros conhecidos, FHC (*Fator de Homogeneidade* dos Canais) e FVA (*Fator de Variabilidade* do Ambiente), introduzidos na Subseção 3.2.2, controlam a variação das capacidades médias ($\overline{C}$) e instantâneas (C_{INST}) de cada canal, segundo as Equações 4.1.

$$\begin{aligned} \overline{C} &= \overline{C}_{MAX}\ (FHC + ((1 - FHC\ rand()))) \\ C_{INST} &= \overline{C}\ (1 + FVA\ (0.5 - rand())) \end{aligned} \tag{4.1}$$

II) A *capacidade média* de cada canal ($\overline{C}$) é sorteada segundo uma distribuição uniforme dentro do intervalo $]0, C_{MAX}]$, onde $\overline{C}_{MAX}$ é a *capacidade média máxima* do canal, e a *capacidade instantânea* do canal (C_{INST}) é sorteada utilizando-se uma distribuição normal, com média igual a $\overline{C}_{MAX}$ e desvio padrão dado por σ. Esta configuração é mostrada através das Equações 4.2.

$$\begin{aligned} \overline{C} &\approxeq C_{MAX}\ rand(),\ \overline{C} \geq 0 \\ C_{INST} &= \mathcal{N}(\overline{C}_{MAX}, \sigma] \end{aligned} \tag{4.2}$$

III) A *capacidade média* de cada canal ($\overline{C}$) é equivalente a sua própria *capacidade média máxima* ($\overline{C}_{MAX}$), e a capacidade instantânea de cada canal (C_{INST}) é sorteada utilizando-se uma distribuição uniforme dentro do intervalo $[\overline{C} \times (1 - \frac{FVA}{2}), \overline{C} \times (1 + \frac{FVA}{2})]$, onde FVA é o conhecido *Fator de Variabilidade* do ambiente e $\overline{C}$ é a *capacidade média* do canal. Esta configuração é mostrada através das Equações 4.3.

$$\begin{aligned} \overline{C} &\equiv C_{constante} \\ C_{INST} &= \overline{C}\ (1 + FVA\ (0.5 - rand())) \end{aligned} \tag{4.3}$$

Essas 3 variações agrupam, respectivamente, os principais casos de uso relativo as características dos canais, conforme discutido na Subseção 2.5.2: heterogêneo, controlado por FHC e FVA; totalmente heterogêneo; e, homogêneo, controlado por FVA.

E a ocupação dos canais (modelo de comportamento para o usuário primário) manteve-se conforme descrito na Subseção 3.2.2, ou seja, "livre" ou "ocupado", variando a cada *slot* segundo:

- uma distribuição uniforme; ou,
- um modelo *ON-OFF* exponencial.

A cada *slot* T, o simulador calcula a recompensa obtida por cada uma das sequências implementadas, correspondente a taxa de transmissão efetiva (Seção 3.1), utilizando-se dos mesmos estados e das mesmas capacidades instantâneas dos canais (critério de justiça).

Uma rodada de simulação consiste na execução de X *slots*. Ao final de cada rodada, o simulador fornece a recompensa média obtida por cada uma das sequências em todos os X *slots*.

4.1.2 Avaliação

Os detalhes da parametrização para a avaliação do nosso mecanismo, ampliado para a inclusão das "novas" estratégias, mantiveram-se em grande parte conforme descrito no início da Subseção 3.3.3. Passaremos a comentar sobre as modificações realizadas.

Para auxiliar na avaliação, introduzimos algumas métricas adicionais, todas com valores limitados no intervalo [0,1]:

- **Índice de aproveitamento das "oportunidades" - IAO**: medida efetiva da utilização das "oportunidades" criadas pela desocupação do canal pelo primário. Esse índice mede a sensibilidade na percepção da criação das "oportunidades" e uma correspondente eficácia no seu aproveitamento. É desconhecido pelos mecanismos, sendo calculado a cada *slot* através da Equação 4.4, e ao final de todos os *slots*, obtida a sua média:

$$\mathsf{IAO} = \begin{cases} \frac{\#_{opUsed}}{\#_{SEC}} & se\ \#_{opCreated} \geq \#_{SEC} \\ \frac{\#_{opUsed}}{\#_{opCreated}} & se\ \#_{opCreated} < \#_{SEC} \end{cases} \tag{4.4}$$

  Onde $\#_{SEC}$ corresponde à quantidade de secundários e, $\#_{opCreated}$ e $\#_{opUsed}$, correspondem, respectivamente, às quantidades de "oportunidades" criadas e de "oportunidades" aproveitadas. Essas últimas, são obtidas através de

contadores, que estão no nível apenas do simulador, e são desconhecidas pelos mecanismos.

Desta forma, se a quantidade de "oportunidades" criadas for superior a quantidade de secundários, para ser justo, o **IAO** se referencia a quantidade de secundários; do contrário, a referência passa a ser a quantidade de "oportunidades" criadas.

- **Índice do consumo de energia - ICE**: é um indicador do dispêndio de energia dos mecanismos, contudo, propositalmente desconhecido por eles. É também calculado a cada *slot*, e ao final de todos os *slots*, obtida a sua média.

  Aqui cabe uma pequena discussão sobre o que entendemos dessa métrica. A nossa intenção ao estabelecermos esse indicador foi avaliar, de modo aproximado, o consumo energético dos mecanismos. Para isso, relaxamos um pouco a precisão e consideramos apenas os seguintes fatores como fortemente impactantes nesse consumo:

  - Processo de mudança de canal: podemos considerar que o chaveamento entre canais tem um gasto energético fixo entre os mecanismos e, portanto, a quantidade de "chaveamentos" é quem determina o gasto total;
  - Processo de sensoreamento de canal: embora seja predominantemente "passivo", como no detector por energia, há um gasto energético fixo do circuito eletrônico que é empregado na função. Assim, podemos considerar que se esse processo é repetido em mais canais, o gasto total aumenta; e,
  - Processo de transmissão (de dados): podemos estimar o gasto energético desse processo indiretamente através da efetividade da transmissão, calculada pela fórmula $e_i = 1 - \frac{i \times \tau}{T}$ (Subseção 3.2.2). Quanto maior a efetividade, maior o gasto total.

  O cálculo rigoroso do gasto energético dos mecanismos deve considerar com precisão, além desses fatores, aspectos de camada física, como potência de transmissão, SNR, modulação empregada, eficiência dos circuitos eletrônicos, etc., e está além do escopo pretendido nesta análise.

  Repare que os dois primeiros processos listados são complementares com o último, de modo que, se há uma quantidade maior de "chaveamentos" entre canais, que, por sua vez, exigem um sensoreamento para se certificar que estão de fato "livres", antes de serem utilizados, a efetividade da transmissão diminui. Além disso, observando os processos listados, é razoável desprezar o

gasto energético referente as mudanças de canal e ao sensoreamento, face ao gasto com a transmissão. Assim, se mantivermos o processo de transmissão na contabilização do gasto energético, estaremos, indiretamente, realizando uma medida da efetividade da transmissão, tornando o mecanismo mais "efetivo", também o mais voraz no tocante ao consumo.

Assim, para manter o propósito da métrica e reduzir a correlação existente entre os processos, contabilizaremos apenas a quantidade de repetições dos processos de mudança e de sensoreamento de canal no cálculo do indicador do dispêndio de energia dos mecanismos (**ICE**), que é ainda normalizado pelo pior caso, caracterizado pela necessidade de visitar e sensorear todos os canais existentes, em todos os *slots*.

- **Índice de *Fairness* -** $f(x)$: criado pelo Professor Raj. Jain [113], corresponde a uma medida quantitativa da "equidade" na alocação da recompensa disponível entre os secundários, sendo largamente empregado em diversas situações, sempre com essa mesma finalidade. É obtido através da Equação 4.5, onde U corresponde à quantidade de secundários e r_i a recompensa coletada pelo i-esimo secundário.

$$f(r) = \frac{(\sum_{i=1}^{U} r_i)^2}{\sum_{i=1}^{U} r_i^2} \ , \ r_i \ \geq \ 0, \ \sum_{i=1}^{U} r_i^2 \ > \ 0 \tag{4.5}$$

Foram feitas 200 rodadas de simulação para cada conjunto de parâmetros, para os 3 cenários anteriormente descritos, onde comparamos o desempenho das sequências detalhadas na Subseção 4.1.1. Os parâmetros usados, bem como os valores que eles assumem, estão resumidos na Tabela 4.1.

Como era de se esperar, observando as Figuras 4.1 a 4.3, o desempenho da sequência `Aleat` foi o pior para os 3 cenários avaliados. Além disso, a diferença entre essa sequência e as demais aumenta com o aumento da quantidade de secundários. Isso também era esperado, pois uma escolha aleatória de canais tende a ser menos efetiva a medida que o conjunto de opções cresce. Contudo, para os cenários com maior contenção, por exemplo quando 7 ou 10 secundários disputam 3 canais, o desempenho da sequência `Aleat` é similar ao das demais sequências. A razão para isso recai sobre a baixa quantidade de "oportunidades" de canais para utilização efetiva, quando a escolha aleatória gera uma boa distribuição entre os secundários.

Nosso mecanismo apresenta um desempenho melhor em todos os cenários. Esse comportamento é devido a própria construção dos mecanismos. No `RL`, os canais são sensoreados a uma taxa baixa porque os episódios possuem duração T fixa e a escolha efetiva dos melhores canais acontece somente depois do transiente, quando a *Q-table*

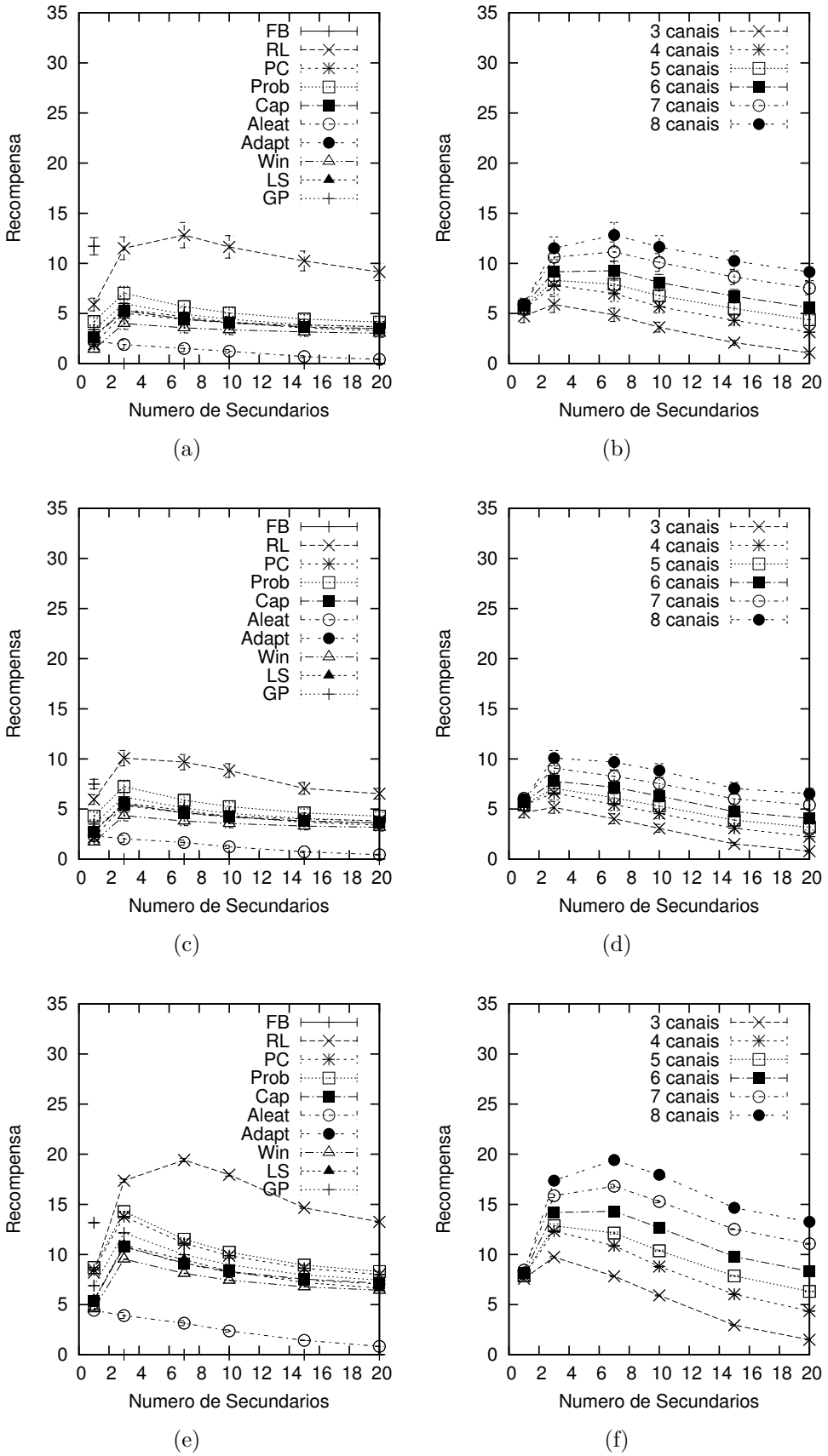


Figura 4.1: Recompensa coletada para comportamento Uniforme do primário, para o cenário 1 (heterogêneo, controlado por FHC e FVA) ((a) e (b)), o cenário 2 (totalmente heterogêneo) ((c) e (d)) e o cenário 3 (homogêneo, controlado por FVA) ((e) e (f)), para 8 canais e estratégia $StEaW$.

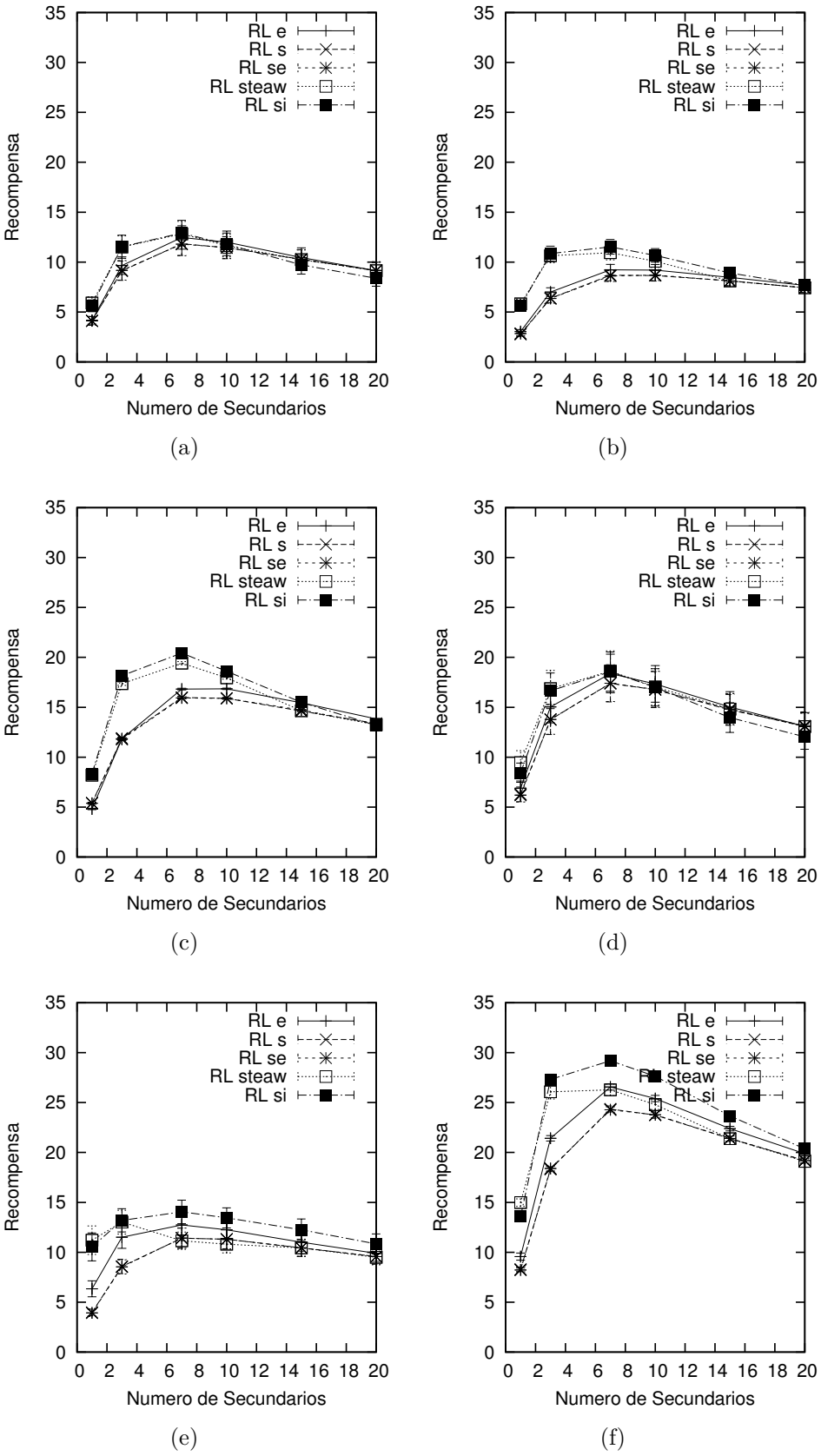


Figura 4.2: Recompensa coletada para o comportamento Uniforme ((a), (b) e (c)) e *ON-OFF* ((d), (e) e (f)) do primário, para o cenário 1 (heterogêneo, controlado por FHC e FVA) ((a) e (d)), o cenário 2 (totalmente heterogêneo) ((b) e (e)) e o cenário 3 (homogêneo, controlado por FVA) ((c) e (f)), para 8 canais e estratégias ε-*greedy*, *softmax*, *softmax* investigativo *se*, $StEaW$ e *softmax* ganancioso *si*.

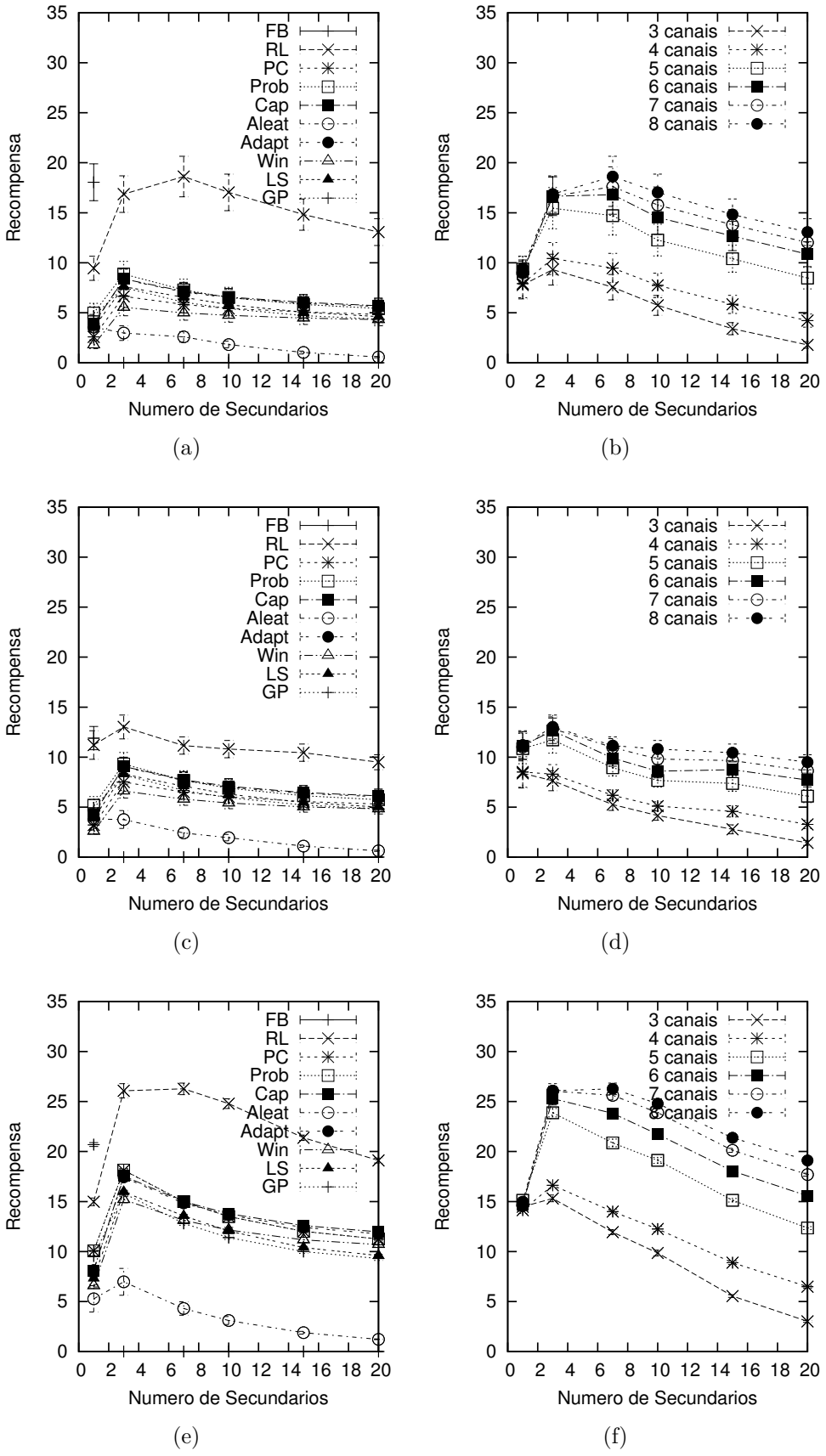


Figura 4.3: Recompensa coletada para comportamento *ON-OFF* do primário, para o cenário 1 (heterogêneo, controlado por FHC e FVA) ((a) e (b)), o cenário 2 (totalmente heterogêneo) ((c) e (d)) e o cenário 3 (homogêneo, controlado por FVA) ((e) e (f)), para 8 canais e estratégia $StEaW$.

Nome	Valor	Conteúdo
$\#_{SEC}$	20	número máximo de secundários
ε	0.3, $\varepsilon \in [0,1]$ (0.7 no transiente)	fator de investigação do `RL`
W_m	8	parâmetro para prob. de colisão [101]
β	0.08, $\beta \in [0,1]$	meta-parâmetro para α
$\overline{C}, C_{INST}$	conforme o cenário	capacidade média e instantânea do canal
Modelo de Canal	Uniforme ou *ON-OFF* Exponencial	comportamento do primário
μ_{OFF}	200	duração média da ocupação do canal
δ	0.95, $\delta \in [0,1]$	fator de perda do `RL`
γ	0.0, $\gamma \in [0,1]$	fator de desconto do `RL`
FHC	$FHC \in [0,1]$	fator de homogeneidade dos canais
FVA	$FVA \in [1,2]$	fator de variabilidade do ambiente
$\overline{C}_{MAX}$	10	capacidade média máxima do canal
σ	5	desvio padrão para $\overline{C}_{MAX}$
T	$2 \times \#_{CH} \times \tau$	tamanho do *slot*
$\mathrm{Pr}_{MISDETECTION}$	0.0	prob. de falha na detecção do primário
$\mathrm{Pr}_{FALSEALARM}$	0.0	prob. de alarme falso de primário
$\#_{SLOTS}$	50.000	número de *slots*
$\#_{CH}$	3 a 8	número de canais
$\#_{RUNS}$	200	número de rodadas

Tabela 4.1: Parâmetros da simulação e avaliação da proposta multiusuário, com o acréscimo das "novas" estratégias (destaque, na parte superior, dos parâmetros variados em relação à avaliação realizada para o caso de um secundário (Tabela 3.1)).

passa a espelhar a consciência do ambiente, na visão do mecanismo. Nos cenários onde as "oportunidades" ofertadas pelos primários variam pouco, esse processo evolui rapidamente, com a *Q-table* sendo representativa das mudanças do ambiente, permitindo, inclusive, que o transiente possa ser encurtado. Por outro lado, naqueles cenários onde o surgimento de "oportunidades" apresenta muita variação, o processo de investigação fornecido pelas estratégias, especialmente a $StEaW$, auxilia na busca pelos canais mais vantajosos.

O nível de contenção entre os secundários exerce uma importante influência nos mecanismos. Quando apenas um secundário realiza a sua busca por um canal, a nossa sequência `RL` demonstra o melhor desempenho, não importa a quantidade de canais. Quando a quantidade de secundários aumenta para 7, a nossa sequência ainda apresenta o melhor desempenho. Entretanto, com 20 secundários, o ganho no desempenho da nossa sequência relativo as demais se reduz.

Observamos através das Figuras 4.1(b), 4.1(d), 4.1(f) e 4.3(b)), 4.3(d), 4.3(f),

que a quantidade de recompensa coletada demonstra um crescimento proporcional a quantidade de secundários, enquanto essa quantidade é menor que a de canais existentes, sendo o contrário também verdadeiro (tendência de redução na recompensa quando a quantidade de secundário torna-se maior que a de canais). Uma explicação para isso está na saturação do sistema. Enquanto não há saturação, quanto maior a quantidade de secundários, maior a recompensa agregada. Entretanto, quando há mais secundários que canais, há maior incidência de colisões, degradando o desempenho, como veremos mais adiante na discussão a respeito da problemática das colisões na rede secundária. Também observamos que o incremento na recompensa agregada cresce com a quantidade de canais, em razão de uma baixa $\Pr_{COLLISION}$ e no aumento da oferta de "oportunidades".

Nesta etapa, avaliamos a métrica *índice de aproveitamento das "oportunidades"* - *IAO*. A oferta de "oportunidades" pelos primários para utilização oportunística dos canais não significa, de forma direta, que um secundário seguindo qualquer ordem de sensoreamento conseguirá aproveitá-las plenamente. Para isso, há dependência intrínseca do mecanismo que fornece a referida sequência ordenada de canais. Da mesma forma, a simples disponibilidade do canal não demonstra o seu real potencial para auferir a melhor recompensa para o secundário. Essas considerações indiretas precisam ser contabilizadas pelo mecanismo para que não haja um desperdício dos canais ofertados, reduzindo a sua utilização oportunística efetiva no longo prazo, porém sem empregar demasiado esforço, inclusive energético (que será analisado mais adiante), por um benefício pequeno (baixo custo-benefício) que levará a um pior desempenho.

Observando as Figuras 4.4(a), 4.4(d) e 4.4(g), podemos verificar o resultado apresentado pelas sequências nessa métrica nos 3 cenários, para o comportamento do primário modelado por uma distribuição ON-OFF exponencial (a diferença para a modelagem uniforme não foi significante). Visivelmente, o nosso mecanismo com qualquer das estratégias se comporta melhor, utilizando mais "oportunidades", mesmo quando a densidade de secundários aumenta. Contudo, observa-se que a medida que aumenta a densidade de secundários, o *IAO* auferido pelo nosso mecanismo se reduz, afastando-se do máximo. Uma explicação para esse desempenho recai sobre a avaliação que é realizada sobre a vantagem de se utilizar no presente o canal considerado "livre" ou prosseguir para o sensoreamento dos próximos canais na sequência. Assim, quando se prossegue na busca, de fato, há um descarte daquela oportunidade, cuja consequência, é a redução do *IAO*.

A contenção sofrida pelos secundários também se reflete sobre essa métrica. Nas Figuras 4.4(b), 4.4(e) e 4.4(h), com o crescimento da quantidade de secundários, há uma forte redução no valor do *IAO*, quando a quantidade de canais é menor. Outra observação, a partir das Figuras 4.4(c), 4.4(f) e 4.4(i), é que as diferentes

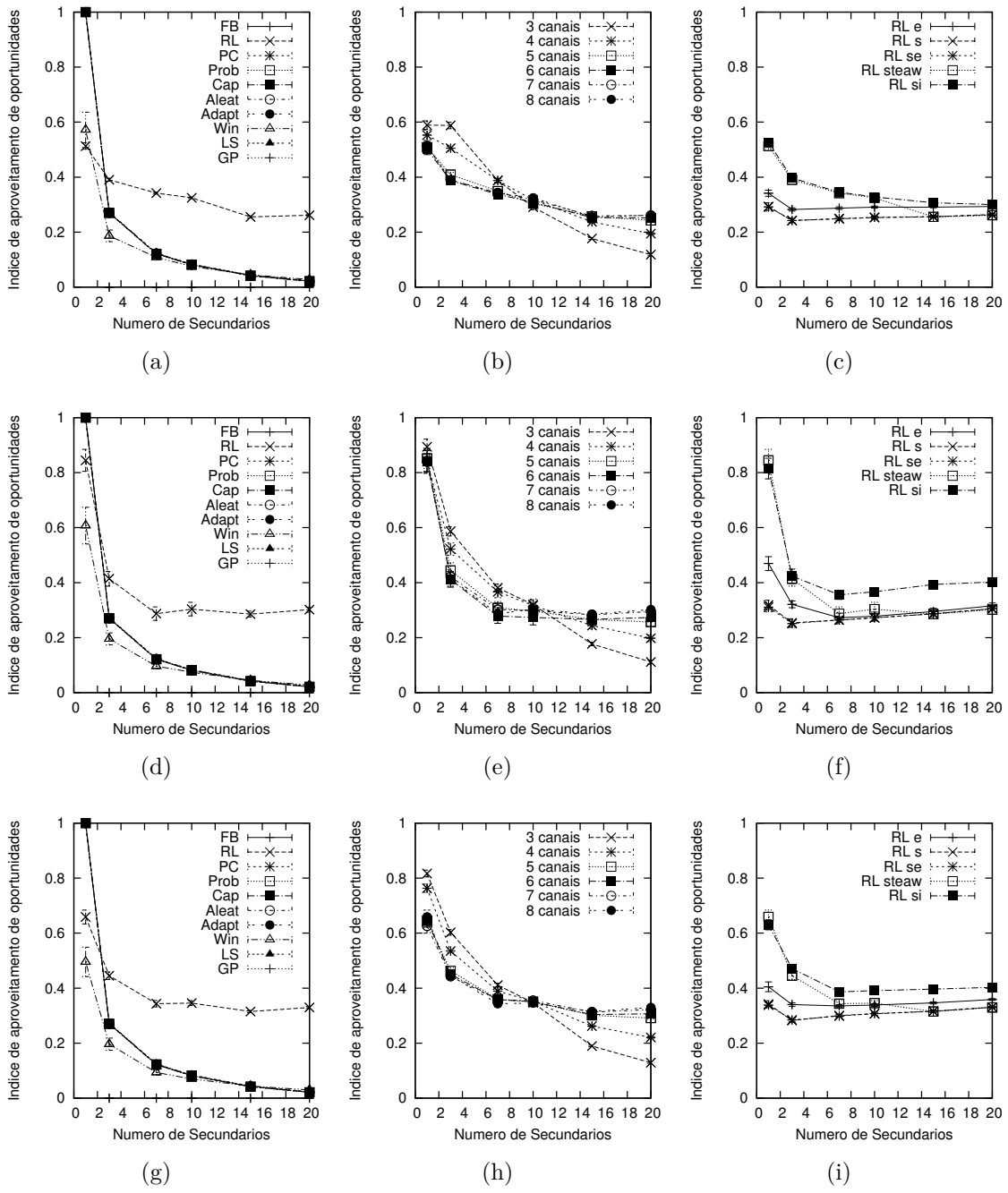


Figura 4.4: Resultados para o comportamento *ON-OFF* do primário do índice de aproveitamento de oportunidades - **IAO**, entre mecanismos para 8 canais; do nosso mecanismo `RL` com estratégia $StEaW$, por canais; e, do `RL`, por estratégia. Para o cenário 1 (heterogêneo, controlado por FHC e FVA) ((a), (b) e (c)), o cenário 2 (totalmente heterogêneo) ((d), (e) e (f)) e o cenário 3 (homogêneo, controlado por FVA) ((g), (h) e (i)).

estratégias empregadas pelo nosso mecanismo modificam o seu desempenho nessa métrica. Isso é devido, principalmente, ao aprendizado efetuado pelo mecanismo, que permite capturar as tendências de modificação da disponibilidade dos canais, mesmo ao variarmos os cenários.

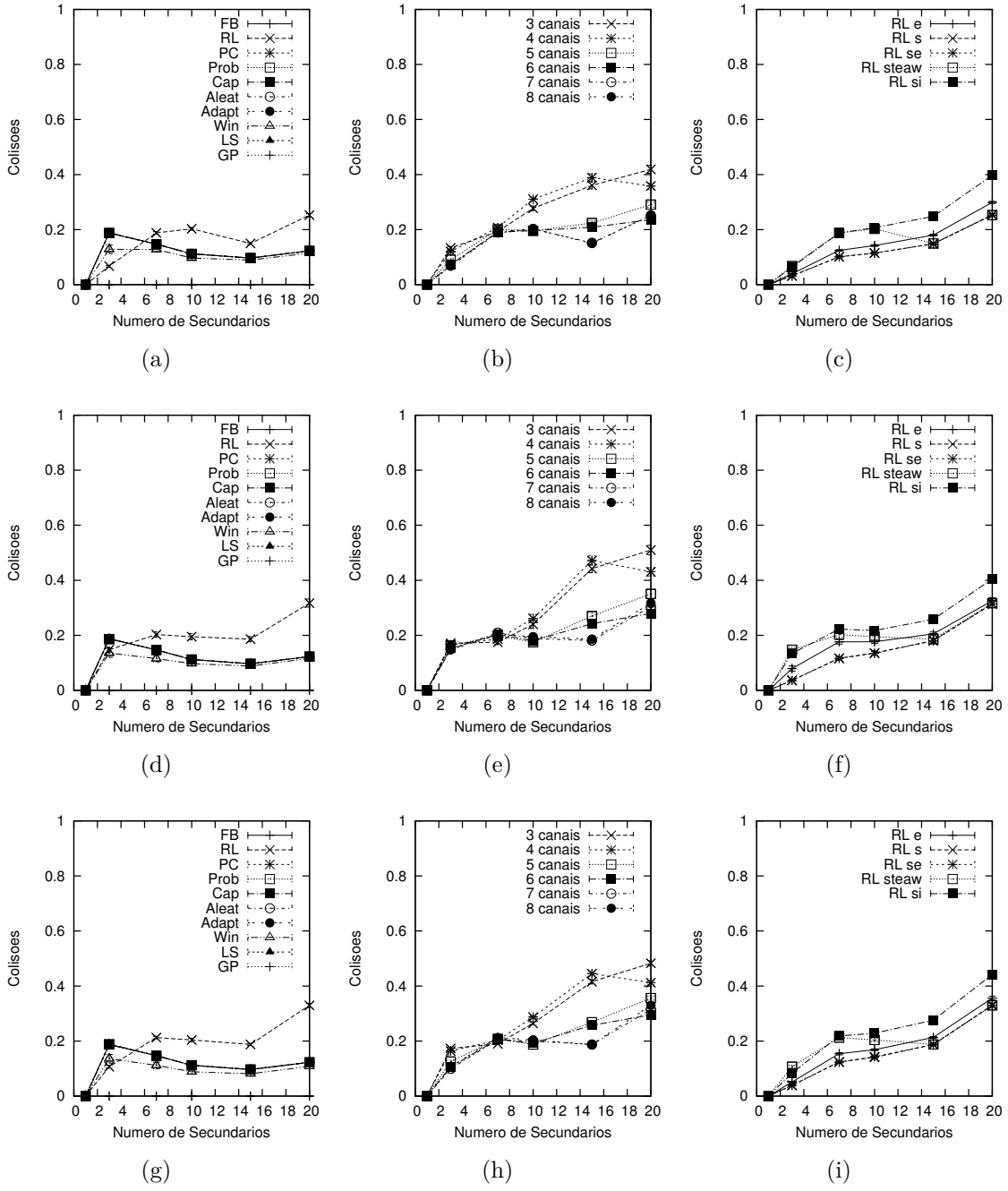


Figura 4.5: Resultados, para o comportamento *ON-OFF* do primário, das colisões na rede secundária, entre mecanismos para 8 canais; do nosso mecanismo `RL` com estratégia $StEaW$, por canais; e, do `RL`, por estratégia. Para o cenário 1 (heterogêneo, controlado por FHC e FVA) ((a), (b) e (c)), o cenário 2 (totalmente heterogêneo) ((d), (e) e (f)) e o cenário 3 (homogêneo, controlado por FVA) ((g), (h) e (i)).

A Figura 4.5 resume os resultados dos mecanismos para a problemática das colisões na rede secundária, nos 3 cenários, para o comportamento *ON-OFF* do primário (a diferença para a modelagem uniforme não foi significante).

Considerando um cenário onde os canais apresentam probabilidades de disponibilidade (Pr_{CH-AV}) semelhantes, o percentual de colisões decresce com a quantidade

de canais. E para uma determinada quantidade de canais, o percentual de colisões cresce proporcionalmente com o crescimento da $\Pr_{CH-AV}$. Essa conclusão é fruto da constatação de que na medida que a quantidade de canais cresce, a chance de dois secundários sensorearem o mesmo canal e o considerarem "livre" de primário se reduz, levando a uma baixa $\Pr_{COLLISION}$. Por outro lado, as "oportunidades" aumentam com o aumento da $\Pr_{CH-AV}$ e, fixando a quantidade de canais, o aumento das "oportunidades" acarreta um crescimento da $\Pr_{COLLISION}$ na medida em que aumenta a possibilidade de que ao menos dois secundários tentem ocupar o mesmo canal, em razão da natureza investigativa/exploratória embutida no nosso mecanismo.

Como esperado, a $\Pr_{COLLISION}$ cresce proporcionalmente a quantidade de secundários, sendo esse crescimento ainda mais abrupto, se a quantidade de canais for pequena, comparado ao caso onde a quantidade de canais é grande. Esse fenômeno pode ser atribuído à maior possibilidade de ocorrência de tentativas de ocupação do mesmo canal simultaneamente, quando há poucos canais. Além disso, quando a quantidade de secundários se equivale a de canais, o nosso mecanismo tem um desempenho melhor em relação as colisões na rede secundária. Contudo, todos os fatores relacionados e as condições necessárias para que isso aconteça ainda precisam de maior investigação.

Por outro lado, quando a quantidade de canais é maior que a de secundários (Figuras 4.5(b), 4.5(e) e 4.5(h)), o desempenho do nosso mecanismo relativo as colisões fica próximo dos demais, o que era esperado. E, quando as $\Pr_{CH-AV}$ dos canais é pequena, o percentual de colisões também se reduz, em razão, principalmente, da baixa quantidade de "oportunidades" ofertadas.

Como podemos verificar na Figura 4.6, o comportamento do nosso mecanismo `RL` em relação a análise da energia consumida, a partir do seu indicador *ICE* se mostrou favorável para todas as estratégias.

Quando estabelecemos o indicador *ICE* postulamos que ele, predominantemente, acompanharia a quantidade de repetições dos processos de mudança e de sensoreamento de canal, sendo esses dois processos, fortemente repetidos pelo nosso mecanismo na sua busca dinâmica pelos "melhores" canais. E com esse pensamento, esperávamos um desempenho inferior ao que ele apresentou nessa métrica.

Uma explicação para isso é que de fato o mecanismo `RL` realiza um intenso processo de busca de canais, porém apenas enquanto há uma investigação muito aleatória das ações (que para a nossa modelagem, são representadas pelos canais). Quando se atinge o equilíbrio e a investigação é reduzida, há uma forte tendência de que a *Q-table* esteja espelhando a realidade encontrada nos cenários, antecipando através da sua exploração, verdadeiramente, os canais mais vantajosos, evitando a necessidade de repetição dos referidos processos, e limitando o consumo de energia.

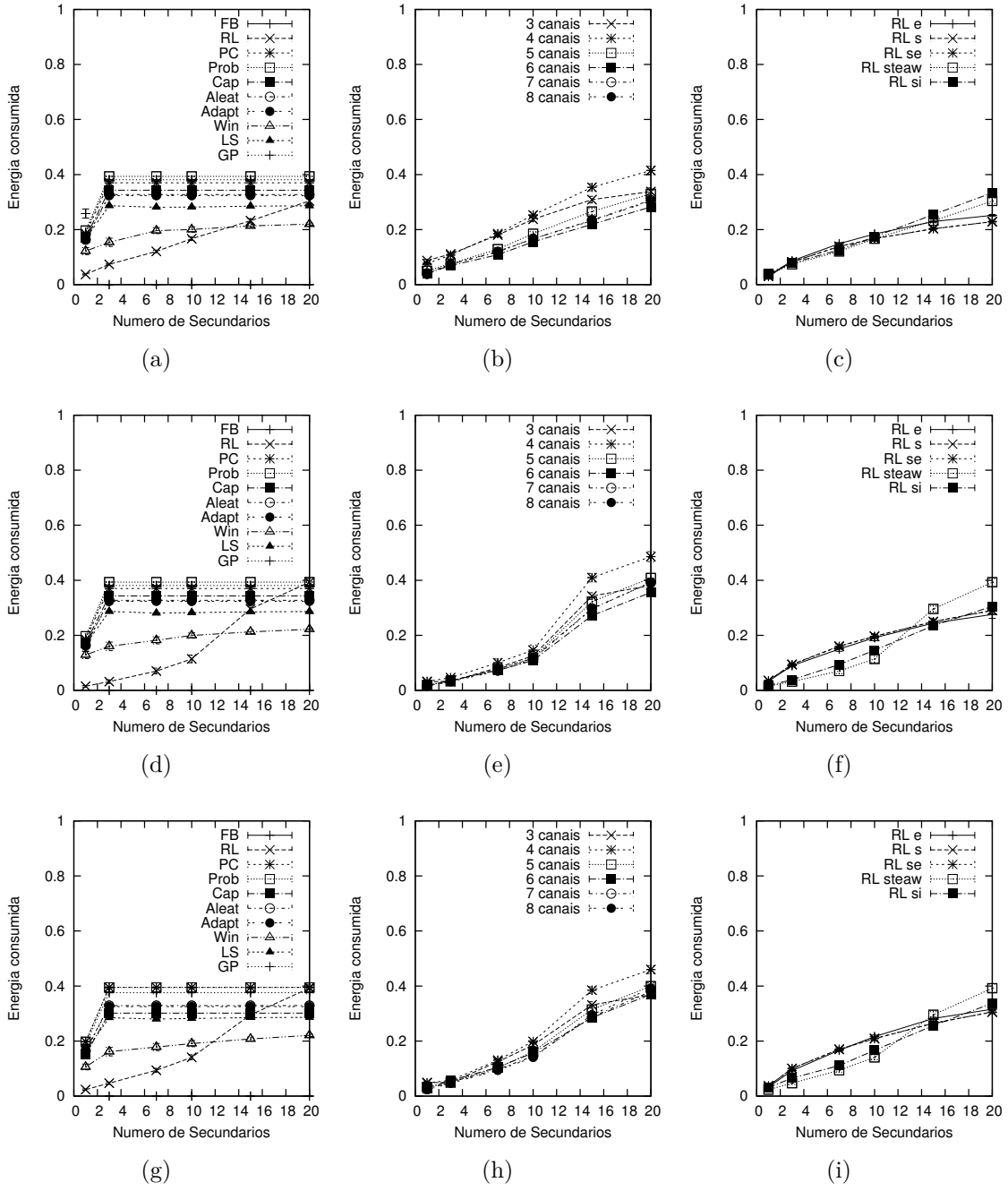


Figura 4.6: Resultados, para o comportamento *ON-OFF* do primário, do índice de consumo de energia - **ICE**, entre mecanismos para 8 canais; do nosso mecanismo `RL` com estratégia $StEaW$, por canais; e, do `RL`, por estratégia. Para o cenário 1 (heterogêneo, controlado por FHC e FVA) ((a), (b) e (c)), o cenário 2 (totalmente heterogêneo) ((d), (e) e (f)) e o cenário 3 (homogêneo, controlado por FVA) ((g), (h) e (i)).

Em ralação ao índice de *fairness* - $f(r)$, cujo resultado podemos observar na Figura 4.7, nenhuma surpresa no desempenho do nosso mecanismo, que já apresentava um excelente comportamento na divisão individual da recompensa agregada durante a evolução no seu desenvolvimento, evitando comportamentos egoístas, comprovado

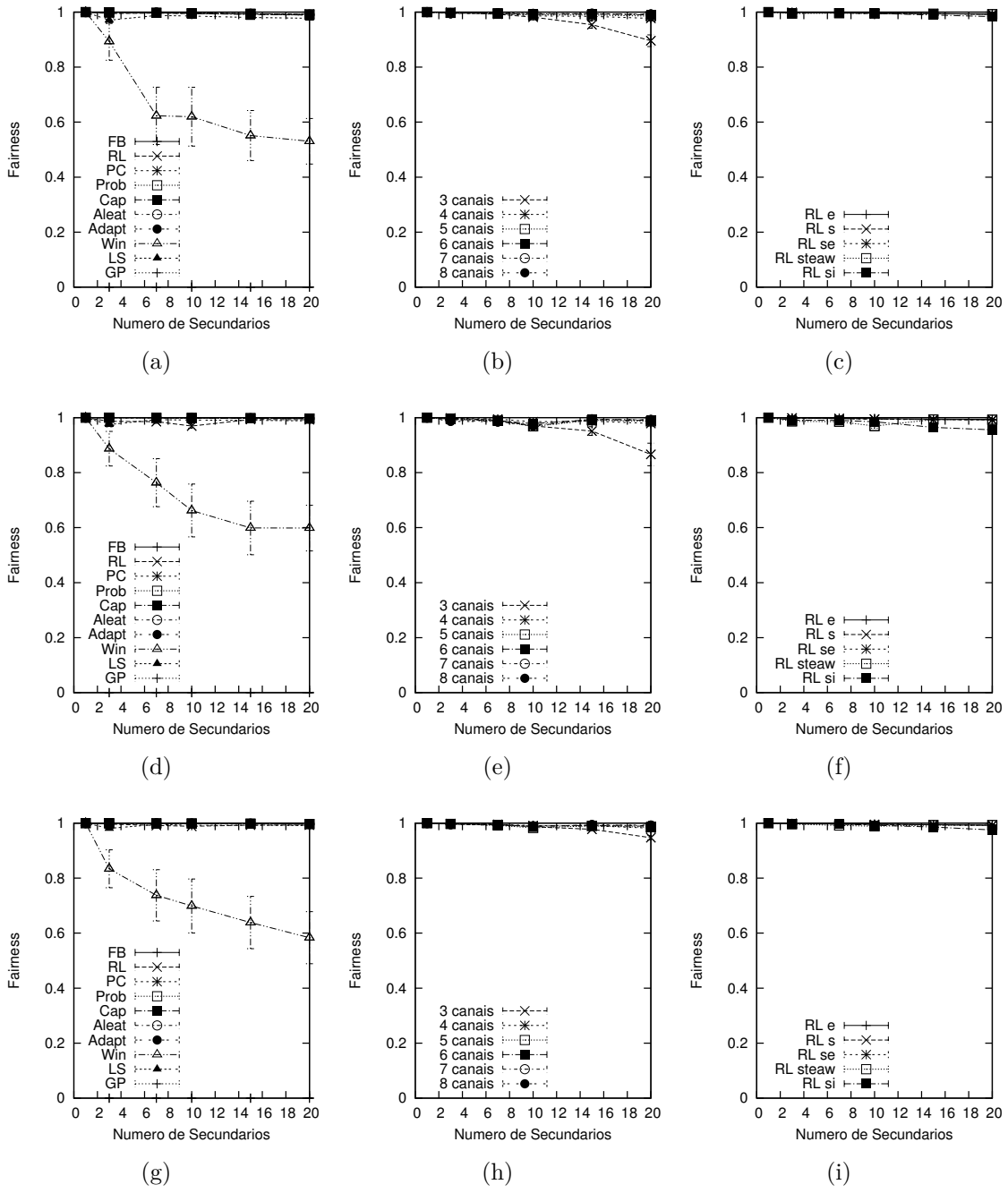


Figura 4.7: Resultados, para o comportamento *ON-OFF* do primário, do índice de *Fairness* - $f(r)$, entre mecanismos para 8 canais; do nosso mecanismo `RL` com estratégia $StEaW$, por canais; e, do `RL`, por estratégia. Para o cenário 1 (heterogêneo, controlado por FHC e FVA) ((a), (b) e (c)), o cenário 2 (totalmente heterogêneo) ((d), (e) e (f)) e o cenário 3 (homogêneo, controlado por FVA) ((g), (h) e (i)).

pelos histogramas apresentados na avaliação discutida na Subseção 3.3.3.

Contudo, quando a quantidade de canais é pequena e há muitos secundários, apenas uma pequena parcela deles coleta toda a recompensa agregada contabilizada, deixando alguns secundários com nenhuma recompensa. Isso é refletido pela redução da métrica, conforme verificamos nas Figuras 4.7(b), 4.7(e) e 4.7(h).

Idealmente, a repartição da recompensa agregada é melhor quando a quantidade de secundários é próxima da quantidade de canais, porém ao custo de um crescimento da $\Pr_{COLLISION}$. E nesse caso, a regra que estabelece a quantidade de canais e de secundários, mantendo o índice de *fairness* da rede e a $\Pr_{COLLISION}$ baixos poderia ser aplicada para evoluir o nosso mecanismo, mitigando o problema das colisões. Entretanto, essa regra ainda não conseguimos obter a partir das análises e avaliações que fizemos, deixando essa tarefa para os trabalhos futuros.

4.2 Análise da Convergência do Mecanismo

Esta seção discute a convergência do nosso mecanismo multiusuário, detalhado no Capítulo 3, aproveitando a implementação já realizada do nosso simulador, conforme descrito nas Subseções 3.2.2 e 3.3.2.

O modelo de simulação manteve-se conforme descrito na Subseção 3.3.2, lembrando que uma rodada de simulação consiste na execução de X *slots* e, ao final de cada rodada, o simulador fornece a recompensa média obtida pela sequência gerada pelo nosso mecanismo, em todos os X *slots*.

Os detalhes da parametrização para essa avaliação, mantiveram-se em grande parte conforme descrito na Subseção 3.3.3. Foram feitas 200 rodadas de simulação para cada conjunto de parâmetros, cujos valores assumidos estão resumidos na Tabela 4.2. A barra de erros corresponde ao intervalo de confiança de 95%.

Inicialmente, assumimos que a convergência do mecanismo indica que houve uma solução para o problema, que é próxima da solução ótima, e que esse resultado não se modifica com o passar do tempo. Quando o problema é dependente do tempo, a convergência do mecanismo indica também que houve uma solução conforme já comentado, porém relativa a um determinado instante.

Assim, o nosso caso é um problema restrito ao *slot*. Se adotarmos uma "janela" de observação (em *slots*) demasiadamente pequena, o mecanismo pode até convergir, mas a acurácia da solução pode ficar comprometida.

A discussão acerca da convergência do mecanismo reúne também aspectos do processo iterativo para se chegar até a solução. Assim, podem ser necessários vários *loops* de iteração dentro do próprio mecanismo para isso, inclusive, cada qual com seus próprios critérios para a convergência.

A "eficiência" da convergência (ou o tempo para convergir) geralmente está relacionada com os valores assumidos pelos parâmetros do mecanismo, que no nosso caso são: a taxa de aprendizado α, o fator de desconto γ, o fator de investigação ε da estratégia ε-*greedy* e o valor do parâmetro temperatura t da estratégia *softmax*.

Uma métrica importante para essa discussão é a medida do arrependimento (*regret*). Esse termo tem origem na análise econômica e corresponde a uma medida

Nome	Valor	Conteúdo
ε	$\varepsilon \in [0.2, 0.7]$	fator de investigação da ε-*greedy*
t	$t \in [100, 1000]$	temperatura do *softmax*
$\#_{CH}$	3 a 10	número de canais
β	0.08, $\beta \in [0, 1]$	meta-parâmetro para α
Modelo de Canal	Uniforme ou *ON-OFF* Exponencial	comportamento do primário
μ_{OFF}	200	duração média da ocupação do canal
δ	0.95, $\delta \in [0, 1]$	fator de perda do `RL`
γ	0.0, $\gamma \in [0, 1]$	fator de desconto do `RL`
FHC	$FHC \in [0, 1]$	fator de homogeneidade dos canais
FVA	$FVA \in [1, 2]$	fator de variabilidade do ambiente
$\overline{C}_{MAX}$	10	capacidade média máxima do canal
σ	5	desvio padrão para $\overline{C}_{MAX}$
$\overline{C}$	$U(FHC * C_{MAX}, C_{MAX})$	capacidade média do canal
C_{INST}	$U(\overline{C} * (1 - FVA/2), \overline{C} * (1 + FVA/2))$	capacidade instantânea do canal
T	$2 \times \#_{CH} \times \tau$	tamanho do *slot*
$\Pr_{MISDETECTION}$	0.0	prob. de falha na detecção do primário
$\Pr_{FALSEALARM}$	0.0	prob. de alarme falso de primário
$\#_{SLOTS}$	até 50.000	número de *slots*
$\#_{RUNS}$	200	número de rodadas
$\#_{SEC}$	1	número de secundários

Tabela 4.2: Parâmetros da simulação e avaliação da convergência do mecanismo multiusuário, com o acréscimo das "novas" estratégias (destaque, na parte superior, dos parâmetros variados em relação a avaliação realizada para o caso de um secundário (Tabela 3.1)).

da divergência entre o resultado exato (ou ótimo) e aquele que de fato foi obtido. Assim, a ocorrência de uma grande diferença entre a recompensa coletada a partir de uma ação tomada e aquela (provável) obtida a partir da ação ótima gera um "arrependimento" elevado, em função do que deixou de ser coletado.

$$\rho(L) = \sum_{j=1}^{L} (r^*(j) - r_\pi(j)) \tag{4.6}$$

A Equação 4.6 calcula a medida do arrependimento, $\rho(L)$, após L, $L \in \mathbb{Z}$, *slots* transcorridos, onde $r^*(j)$ é a recompensa coletada no *slot* j, $j \in \mathbb{Z}$, segundo a estratégia ótima, e, $r_\pi(j)$ é a recompensa coletada no mesmo *slot*, conforme a estratégia adotada π.

A partir da medida do arrependimento, derivamos outra métrica, que utilizamos na nossa discussão, chamada medida do arrependimento médio por *slot*, $\frac{\overline{\rho(L)}}{L}$. Desta

forma, se $\lim_{L\to\infty} \frac{\overline{\rho(L)}}{L} \to 0$ o mecanismo é dito com "arrependimento zero". Intuitivamente, um mecanismo com essa propriedade converge para a solução ótima, após transcorridos um número suficiente de *slots*.

Calculamos o valor dessa métrica como uma média móvel dentro de um intervalo variante de 400 *slots*. Essa métrica é, então, observada ao longo de toda a simulação, que considera também diferentes intervalos para o período do transiente e de aprendizado, em número de *slots*, que chamamos "janelas", as quais atribuímos os valores de 1.000, 3.000 ou 5.000 *slots*, seguidos por mais 10.000 *slots* a fim de se verificar a estacionariedade.

As Figuras apresentadas resumem os resultados do mecanismo, nos 3 cenários (Subseção 4.1.1), para o comportamento *ON-OFF* do primário (a diferença para a modelagem uniforme não foi significante).

Iniciamos, então, nossa discussão, pelos resultados do arrependimento médio por *slot*. Podemos observar na Figura 4.8 que o mecanismo mantém o valor dessa métrica quase constante na passagem dos *slots*, nas 3 "janelas", para as estratégias *ε-greedy* e *softmax*, e suas derivadas.

Uma vez que esse valor não é decrescente (tendência de "arrependimento zero"), nada podemos afirmar sobre a convergência do mecanismo, apenas com essa análise. Estendendo até 50.000 o número de *slots* transcorridos, ainda assim o valor da métrica se mantém quase constante.

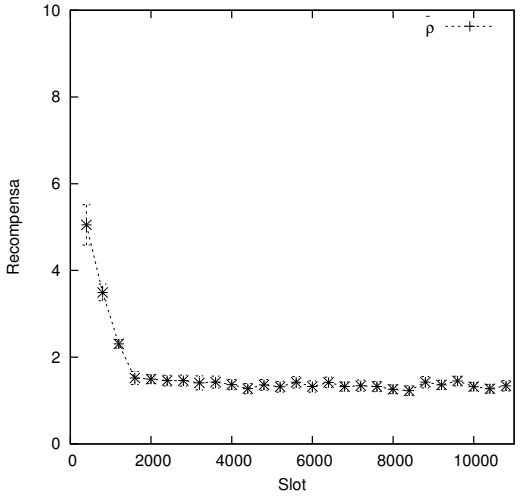
De fato, em razão dessa constância, podemos admitir que o mecanismo seja, no mínimo, sub-ótimo e, também, que há um forte indicador para a sua convergência. Esse comportamento sub-ótimo já havia sido sugerido pelos resultados obtidos na análise da proposta realizado na Subseção 3.2.3, quando nosso mecanismo obteve resultados muito próximos do ótimo obtido por força bruta.

Para auxiliar na discussão sobre a convergência do nosso mecanismo, decidimos criar mais duas métricas:

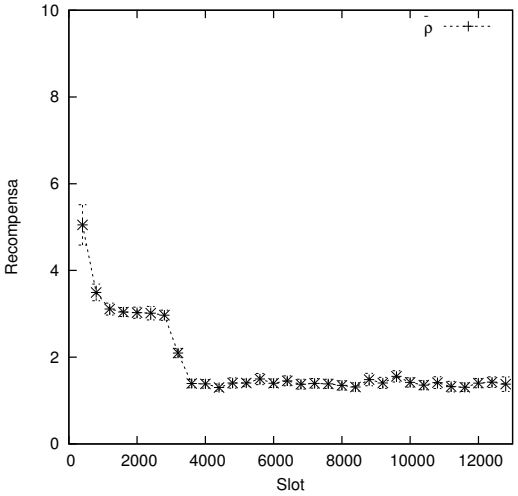
1. Afastamento do ótimo ($AFAST$); e,
2. Quantidade de "janelas" até a convergência ($J2CONV$).

A primeira métrica, $AFAST$, mede o percentual do afastamento da recompensa obtida pelo nosso mecanismo do valor ótimo obtido por força bruta, versus o valor do parâmetro característico de cada estratégia, ε ou t. Essa observação é também realizada dentro de diferentes intervalos para o período do transiente e de aprendizado, em número de *slots*, ou seja, nas "janelas" de 1.000, 3.000 ou 5.000 *slots*, seguidos por mais 10.000 *slots* a fim de se verificar a estacionariedade.

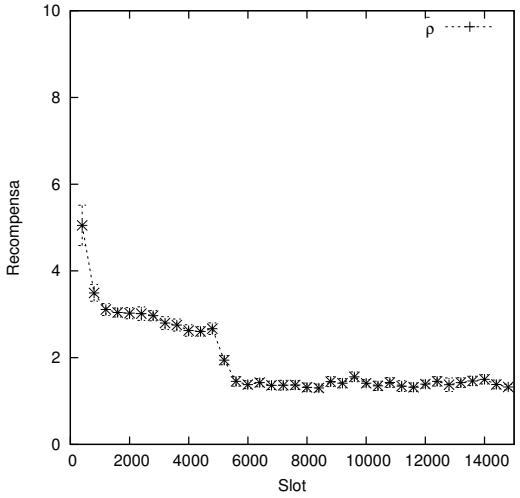
A Figura 4.9 resume os resultados para essa métrica referente as estratégias ***ε-greedy***, e sua derivada $StEaW$. Com 03 canais, e para todas as "janelas", o valor escolhido para o parâmetro $\varepsilon = 0.3$ mostrou-se adequado, mantendo o $AFAST$



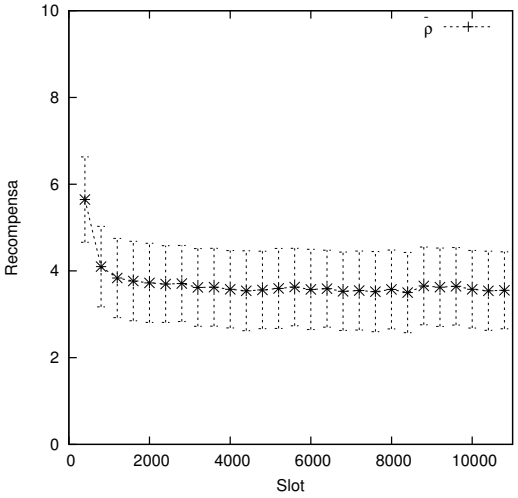
(a) *Janela* de 1.000 *slots*.



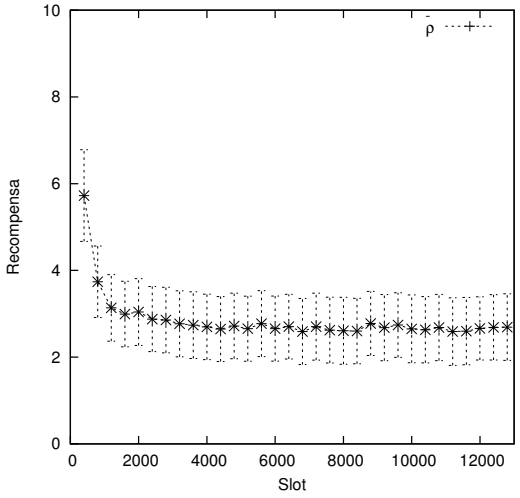
(b) *Janela* de 3.000 *slots*.



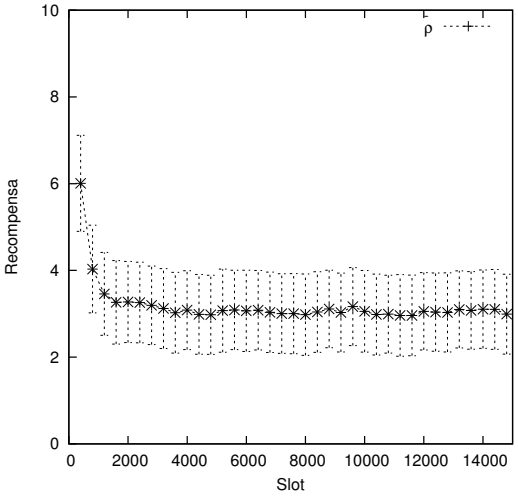
(c) *Janela* de 5.000 *slots*.



(d) *Janela* de 1.000 *slots*.

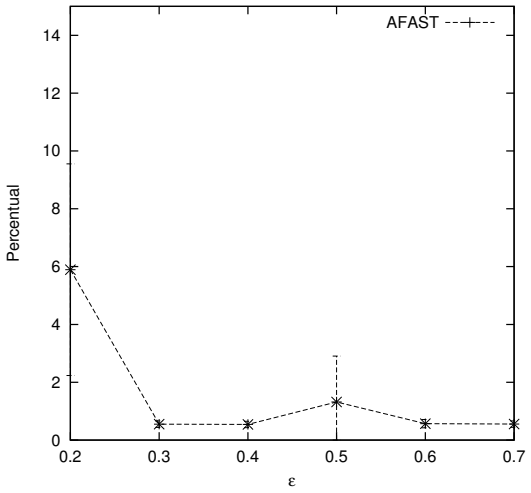


(e) *Janela* de 3.000 *slots*.

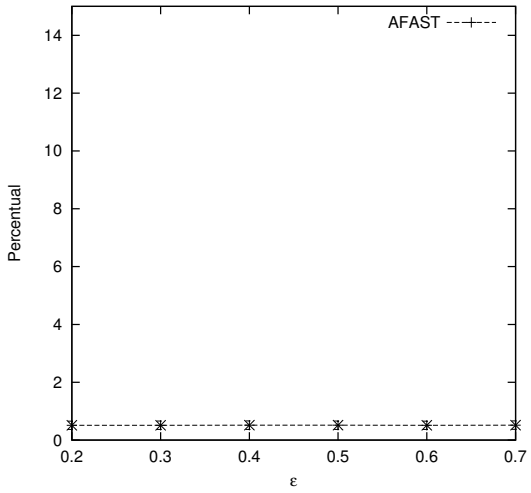


(f) *Janela* de 5.000 *slots*.

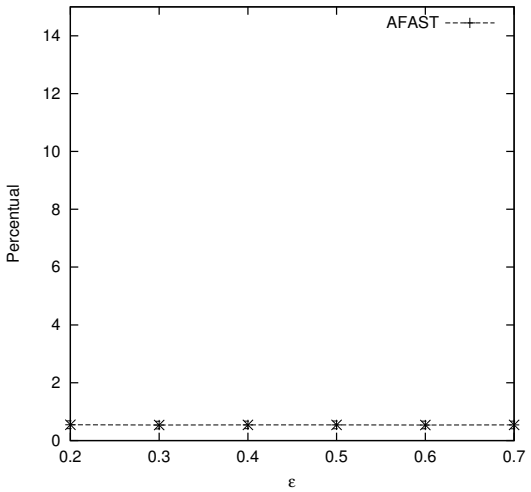
Figura 4.8: Impacto da métrica arrependimento médio por *slot*, $\overline{\rho}$, para as estratégias ε-*greedy* ((a), (b) e (c)) e *softmax* ((d), (e) e (f)), e suas derivadas, e 10 canais.



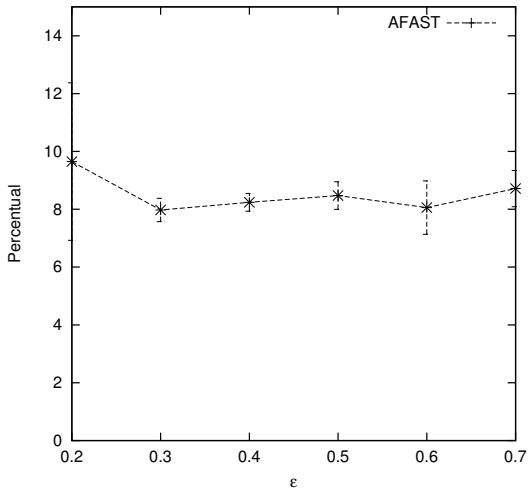
(a) *Janela* de 1.000 *slots*, 03 canais.



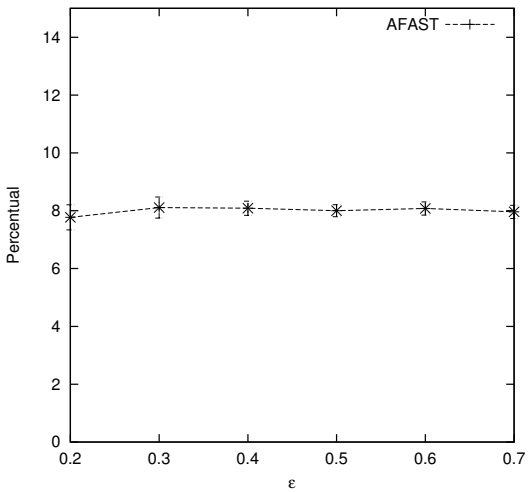
(b) *Janela* de 3.000 *slots*, 03 canais.



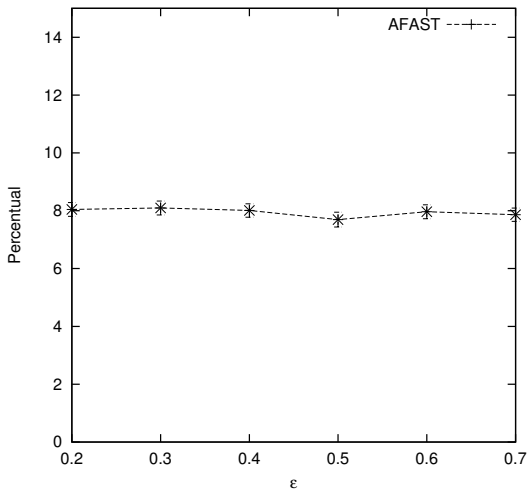
(c) *Janela* de 5.000 *slots*, 03 canais.



(d) *Janela* de 1.000 *slots*, 10 canais.



(e) *Janela* de 3.000 *slots*, 10 canais.



(f) *Janela* de 5.000 *slots*, 10 canais.

Figura 4.9: Impacto da métrica $AFAST$ versus o valor do parâmetro ε.

abaixo de 1%. Com mais canais, e novamente para todas as "janelas", o valor do parâmetro também produziu o $AFAST$ mínimo, em torno de 8%.

Para as estratégias ***softmax***, e suas derivadas: SE e SI (Figura 4.10), tanto para 3 quanto para 10 canais, o valor escolhido para o parâmetro $t_0 = 1.000$ produziu o menor $AFAST$, em todas as "janelas", respectivamente, 5% e 10%.

A segunda métrica, $J2CONV$, mede o período de tempo decorrido até a estacionariedade em quantidade de "janelas" transpassadas, versus o valor do parâmetro característico de cada estratégia, ε ou t. Ou seja, um indicador da "velocidade" da convergência, conforme o parâmetro da estratégia.

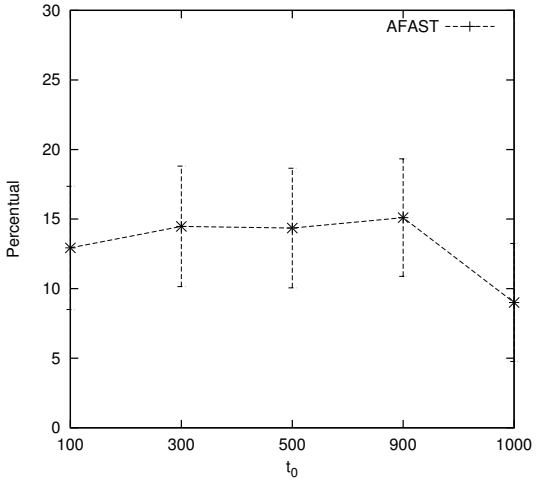
Assim, na "janela" de 1.000 *slots*, para as estratégias ***ε-greedy***, e sua derivada $StEaW$, a Figura 4.11 mostra que o valor 0.3 escolhido para o parâmetro ε não obteve a maior velocidade de convergência, sendo até 50% menos "veloz". Esse resultado demonstra a existência de um compromisso entre manter o $AFAST$ baixo ou convergir mais rapidamente (menor $J2CONV$) e que precisa ser considerado dependendo da aplicação do mecanismo. Para as demais "janelas", o valor escolhido para ε não influenciou na velocidade da convergência.

Outra observação importante é em relação a inexistência de influência do parâmetro ε para a velocidade de convergência, nas "janelas" a partir de 3.000 *slots*. Em outras palavras, nessa condição, a modificação do fator de investigação afeta somente o $AFAST$, ou seja, o afastamento percentual da recompensa obtida pelo nosso mecanismo do valor ótimo.

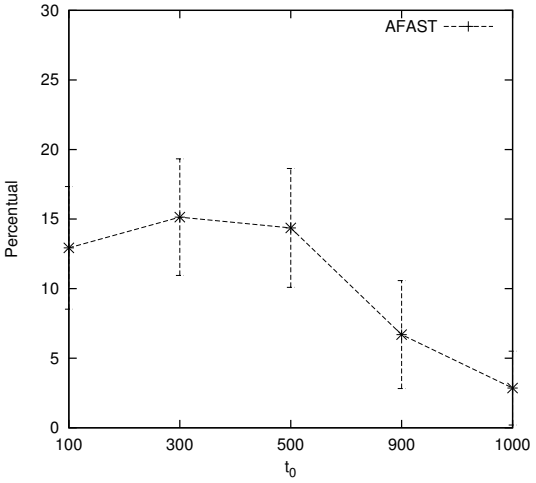
Para as estratégias ***softmax***, e suas derivadas: SE e SI (Figura 4.12), o valor 1.000 escolhido para t_0 também não é o mais "veloz", em todas as "janelas", principalmente na de 1.000 *slots*, onde há a maior diferença. Contudo, diferente do *ε-greedy*, a modificação da inclinação da reta que define o parâmetro *temperatura* (Figura 3.9), através da modificação dinâmica do valor de t dentro do intervalo [100,1000] influencia na velocidade de convergência do mecanismo, para alguns valores, em até 100%.

Se observarmos a evolução da recompensa coletada por *slot*, para as estratégias *ε-greedy* e *softmax* (Figura 4.13), e suas derivadas, é possível notar a tendência assintótica dos valores coletados de recompensa a partir de um determinado *slot*, um forte indicador de que o mecanismo atingiu a estacionariedade.

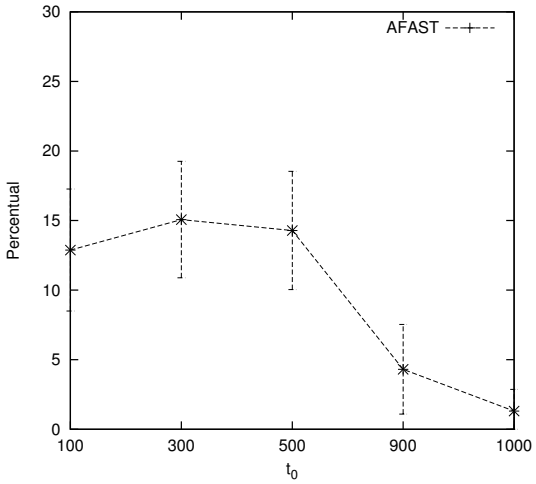
Como conclusão, a partir da análise feita para os 3 cenários, podemos considerar que o arrependimento médio constante por *slot*, visto na Figura 4.8; o valor de $AFAST$ limitado (e por vezes, decrescente), visto nas Figuras 4.9 e 4.10; o valor também limitado de $J2CONV$, visto nas Figuras 4.11 e 4.12; e a tendência assintótica do mecanismo, visto na Figura 4.13, são condicionantes para a convergência do mecanismo; e, que as estratégias ***softmax***, e suas derivadas: SE e SI, convergem para o resultado mais rapidamente (aproximadamente duas vezes mais rápido)



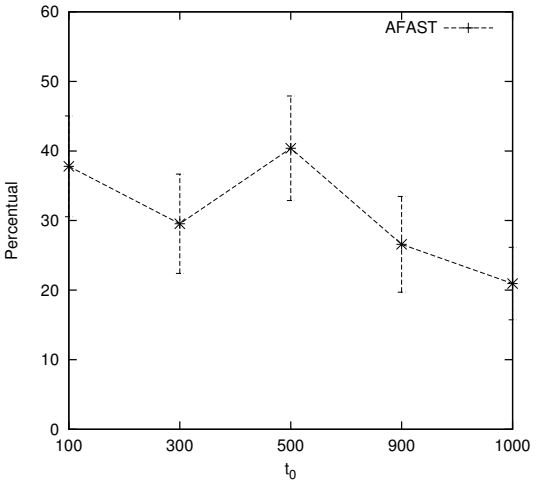
(a) *Janela* de 1.000 *slots*, 03 canais.



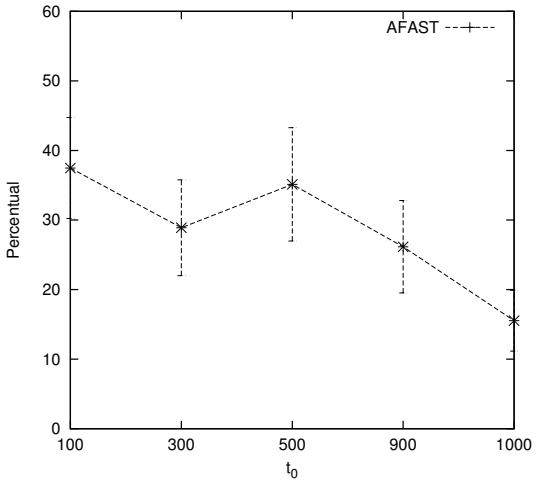
(b) *Janela* de 3.000 *slots*, 03 canais.



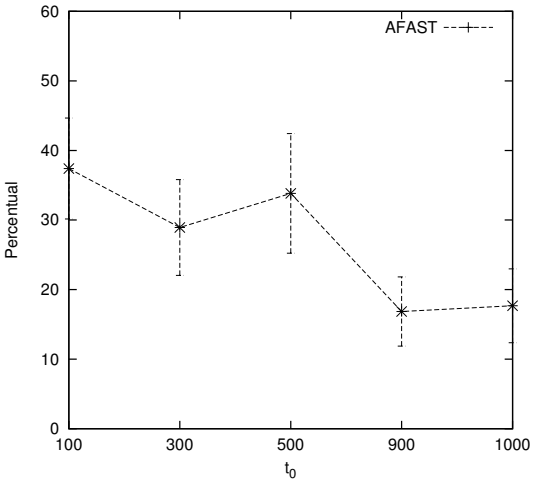
(c) *Janela* de 5.000 *slots*, 03 canais.



(d) *Janela* de 1.000 *slots*, 10 canais.

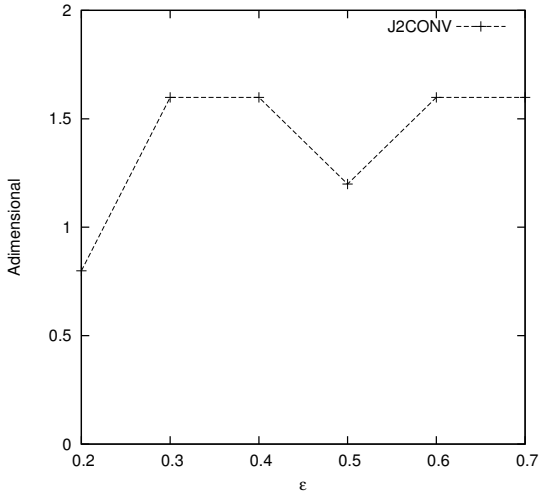


(e) *Janela* de 3.000 *slots*, 10 canais.

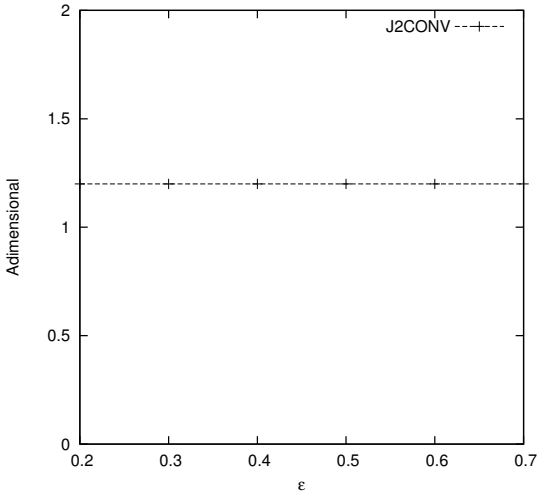


(f) *Janela* de 5.000 *slots*, 10 canais.

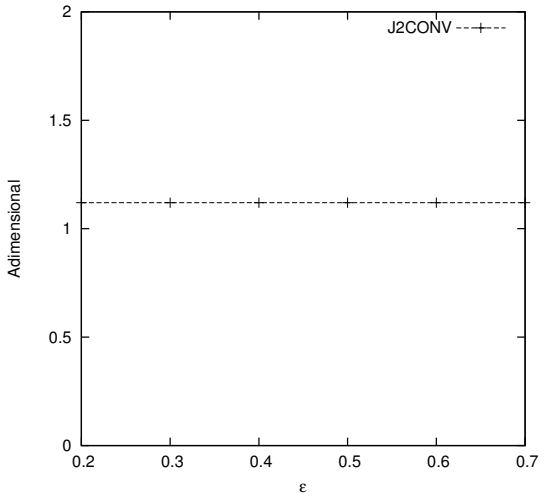
Figura 4.10: Impacto da métrica $AFAST$ versus o valor do parâmetro t_0.



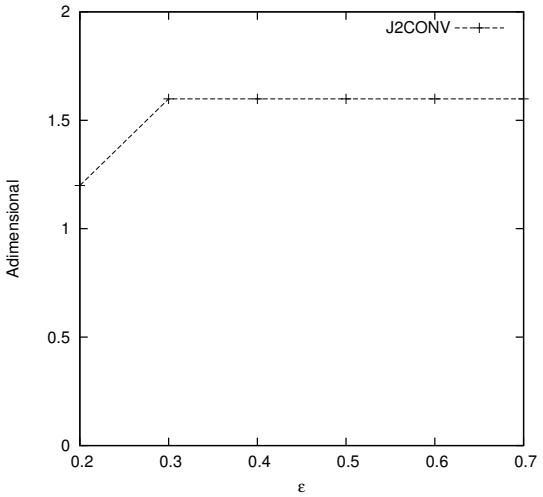
(a) *Janela* de 1.000 *slots*, 03 canais.



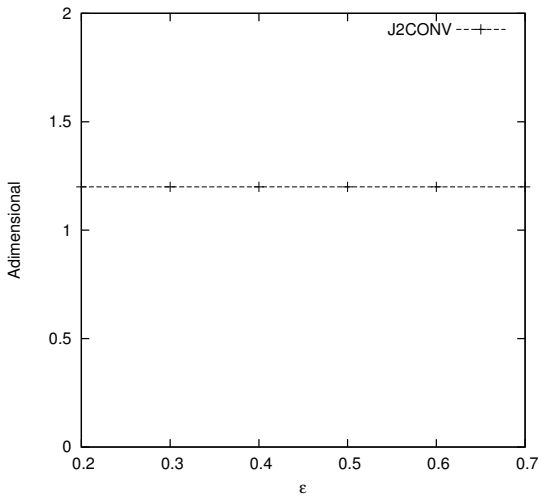
(b) *Janela* de 3.000 *slots*, 03 canais.



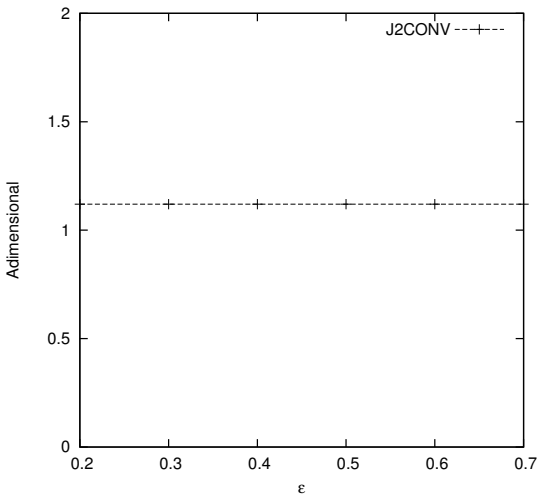
(c) *Janela* de 5.000 *slots*, 03 canais.



(d) *Janela* de 1.000 *slots*, 10 canais.

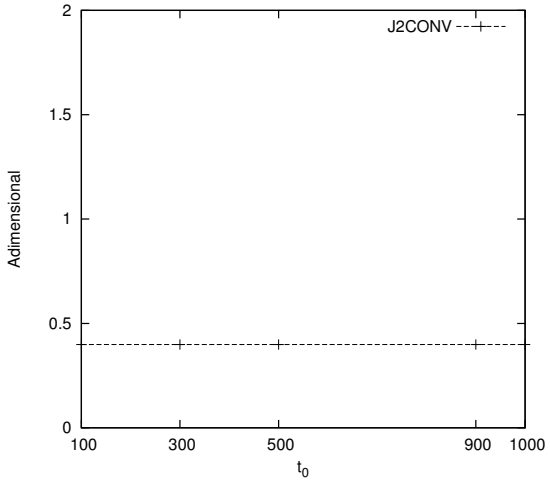


(e) *Janela* de 3.000 *slots*, 10 canais.

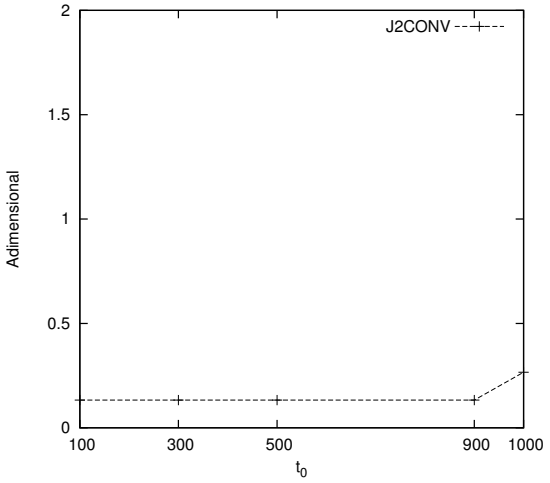


(f) *Janela* de 5.000 *slots*, 10 canais.

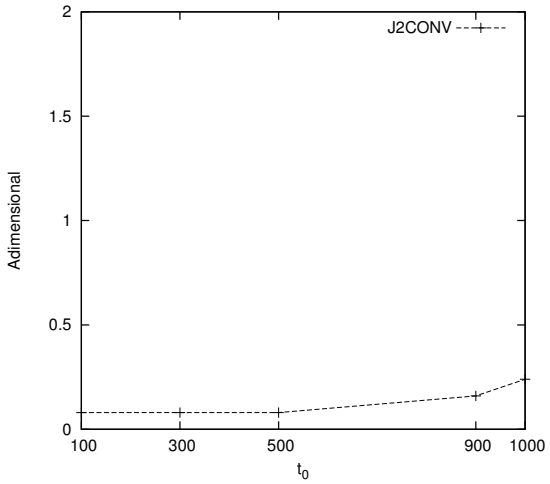
Figura 4.11: Impacto da métrica $J2CONV$ versus o valor do parâmetro ε.



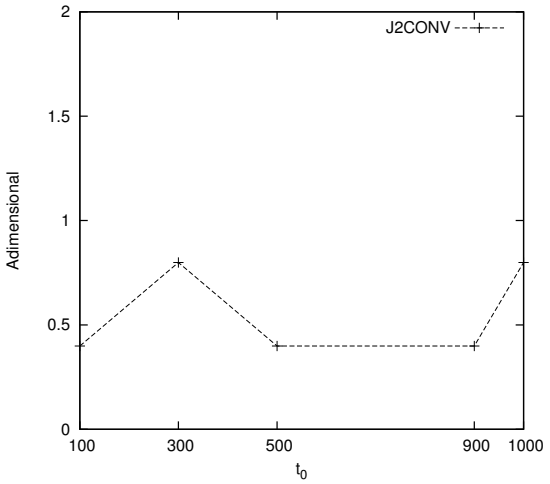
(a) *Janela* de 1.000 *slots*, 03 canais.



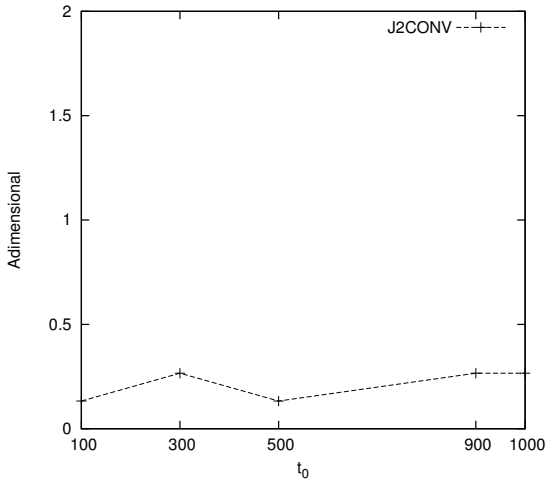
(b) *Janela* de 3.000 *slots*, 03 canais.



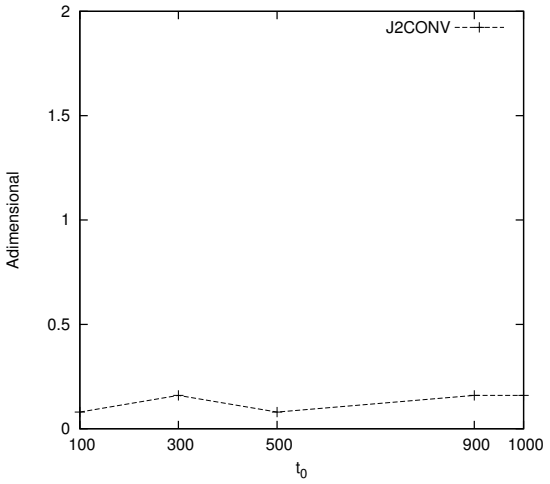
(c) *Janela* de 5.000 *slots*, 03 canais.



(d) *Janela* de 1.000 *slots*, 10 canais.

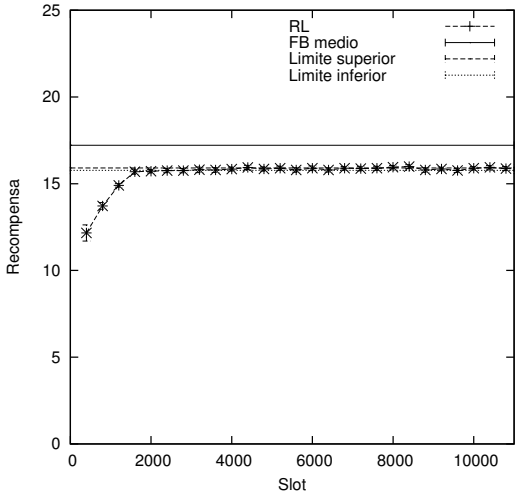


(e) *Janela* de 3.000 *slots*, 10 canais.

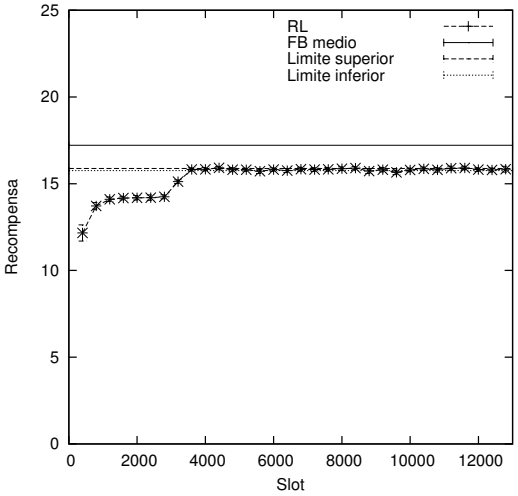


(f) *Janela* de 5.000 *slots*, 10 canais.

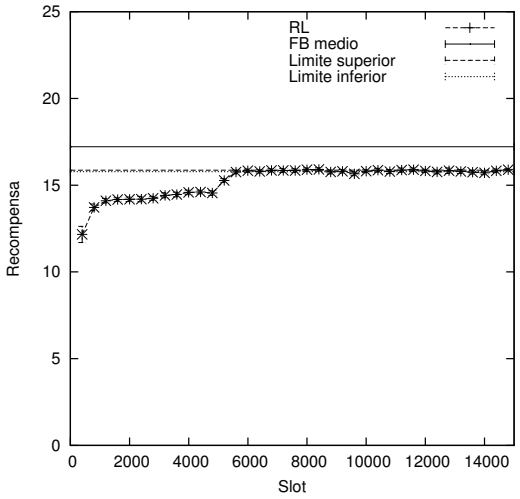
Figura 4.12: Impacto da métrica $J2CONV$ versus o valor do parâmetro t_0.



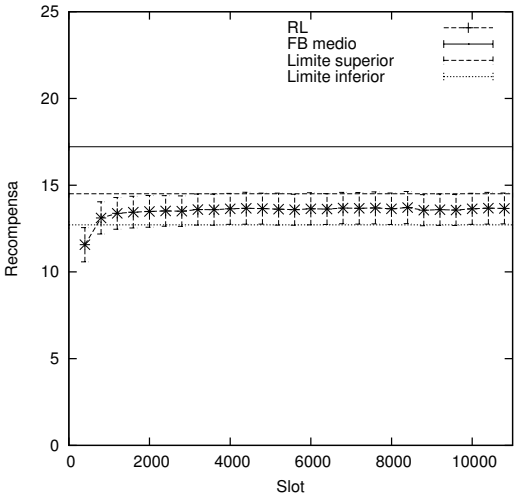
(a) *Janela* de 1.000 *slots*.



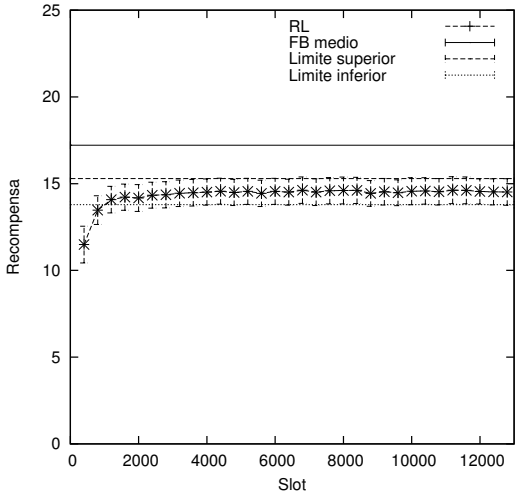
(b) *Janela* de 3.000 *slots*.



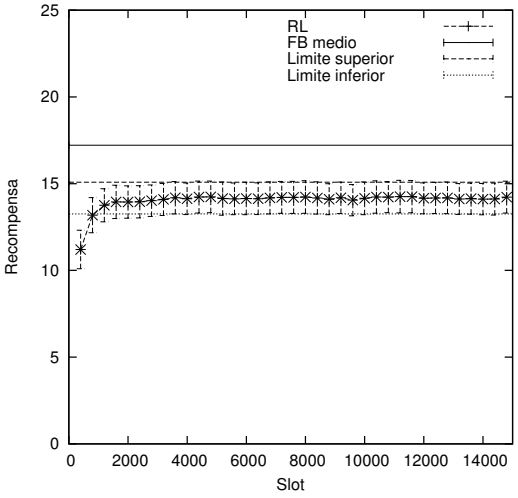
(c) *Janela* de 5.000 *slots*.



(d) *Janela* de 1.000 *slots*.



(e) *Janela* de 3.000 *slots*.



(f) *Janela* de 5.000 *slots*.

Figura 4.13: Evolução da recompensa coletada por *slot*, para as estratégias ε-*greedy* ((a), (b) e (c)) e *softmax* ((d), (e) e (f)), e suas derivadas, e 10 canais.

que as estratégias ε-***greedy***, e sua derivada $StEaW$, mas aproximam-se menos do ótimo, quando comparadas com as derivadas da estratégia ε-***greedy***, afastando-se, na média, aproximadamente o dobro.

4.3 Conclusões

Neste capítulo, propomos novas estratégias para o balanceamento do dilema investigação-exploração existente no nosso mecanismo baseado na técnica *Q-learning* e realizamos a sua avaliação a partir de simulações e da comparação com outras soluções obtidas da literatura [78], obtendo resultados melhores e promissores.

Em uma segunda etapa, realizamos uma discussão sobre a convergência do nosso mecanismo multiusuário, descrito no Capítulo 3, a partir do resultado de simulações, incluindo recomendações para a seleção e o ajuste dos seus parâmetros de modo a melhorar a sua convergência, concluindo, ao final, que há condicionantes fortes para a convergência do mecanismo.

Capítulo 5

Conclusões e Trabalhos Futuros

Os trabalhos apresentados nos capítulos anteriores tratam a questão da ordem de sensoreamento de canais pelos usuários secundários, tanto em relação a fatores internos a rede secundária, como o quantitativo de usuários e a justiça entre eles, quanto em relação a fatores externos, como o quantitativo de canais existentes e o fator de utilização desses mesmos canais pelos usuários primários, com a finalidade de minimizar o tempo de busca e acesso a um canal livre para ser, efetivamente, utilizado.

O foco do trabalho, de forma geral, não foi desenvolver algoritmos ou mecanismos determinísticos com objetivo puramente de alcançar resultados ótimos, mas sim mecanismos que possam atender a requisitos funcionais para seleção e acesso efetivo aos canais disponíveis para uso, tratando o problema de forma probabilística, por entender que questões físicas, por exemplo, relacionadas a propagação do sinal no ambiente de RF levam à incerteza quanto ao real estado da rede em cada instante, portanto, oferecendo um caráter probabilístico a própria rede secundária.

Dentro deste enfoque, nas próximas seções deste capítulo apresentamos as considerações sobre o trabalho, as contribuições desta tese e os trabalhos futuros esperados.

5.1 Considerações

O atual processo regulatório para licenciamento de uso do espectro tem resultado numa grande divergência entre a reserva do espectro e a sua real utilização. Novos desenvolvimentos tecnológicos trarão a expectativa de que este espectro subutilizado possa ser disponibilizado, porém não se vislumbra grandes mudanças políticas e/ ou regulatórias sem uma garantia de não-interferência nos usuários legados (ou primários).

Seguindo neste caminho, surge como uma possibilidade de solução desse impasse, o rádio cognitivo, um dispositivo sem fio capaz de ajustar suas transmissões para

usar apenas bandas subutilizadas, com potencial de ser a chave para aumentar a eficiência de uso do espectro numa região geográfica, formando uma rede secundária.

A tese de Mitola III [11] discutiu os vários níveis operacionais de um rádio cognitivo, mas o formalismo apresentado por ele se concentra quase que exclusivamente sobre a camada de aplicação e acima.

Adicionalmente, os métodos tradicionais de aprendizado "artificial" exigem significantes recursos computacionais, o que poderia limitar a utilidade de um rádio cognitivo capaz de "aprender", mas que consome energia e memória muitas vezes mais que os rádios com tecnologia anterior.

Ao invés disso, esta tese trata as funções do rádio cognitivo como inerentes ao funcionamento das camadas física e de controle de acesso ao meio (MAC) e a cognição embutida resultante não exige um modelo robusto e oferece uma estrutura que funciona dentro da capacidade de computação disponível nas plataformas sem fio atuais, preocupando-se, em paralelo, com a eficiência no aproveitamento das "oportunidades" criadas pelo rádio licenciado e com o consumo energético.

Por conta do esforço para a convergência de redes e serviços, da necessidade de reuso do espectro e da crescente abrangência e popularização do acesso a banda larga entre outros fatores, imaginamos que é questão de pouco tempo para que tenhamos, de fato, redes secundárias maciçamente em operação.

Muitos trabalhos têm se focado no estudo do comportamento das redes secundárias e alguns poucos na sua implementação prática. Essa situação tende a mudar com a iniciativa do FCC para regulamentação do acesso oportunístico a banda de TV analógica nos EUA [114].

5.2 Contribuições

As contribuições que destacamos do nosso trabalho podem ser divididas em duas partes:

- Em *primeiro* lugar, propomos um mecanismo que fornece uma ordem dinâmica de sensoreamento de canais para usuários não-licenciados (secundários), capaz de decidir se deve parar em um canal e utilizá-lo, visando maximizar os ganhos de uma métrica de interesse, ou continuar a busca, mesmo se o canal estiver livre de primários. Além disso, nossa abordagem para a solução do problema de exploração do espectro de RF (*spectrum exploitation*) não exige um conhecimento a priori das capacidades médias e/ou das probabilidades de disponibilidade de cada canal, sendo esta considerada um indicador da atividade do primário;

  - Construímos um *simulador* que emula o funcionamento de uma rede secundária, onde cada um de seus usuários utiliza sequências de sensoreamento arbitrárias.
  - Propomos um *algoritmo* de baixa complexidade ($\mathcal{O}(N)$, onde N é o número de canais disponíveis), que utiliza uma máquina de aprendizado por reforço (*Q-learning*), para o estabelecimento da ordem dinâmica de sensoreamento de canais em ambiente multiusuário e multicanal que:
    * Adapta-se dinamicamente as variações das probabilidades de disponibilidade e da capacidade média esperada em cada canal;
    * É imune as possíveis mudanças nas probabilidades de disponibilidade dos canais, que podem ocorrer devido a alteração no padrão de atividade dos usuários primários, e as possíveis mudanças na qualidade dos canais (SNRs médias), que podem ocorrer devido a mobilidade e aos efeitos de desvanecimento de larga escala;
    * Não favorece a existência de usuários secundários gananciosos;
    * Variando a quantidade de canais e reduzindo a probabilidade de disponibilidade dos canais atinge-se o *limiar máximo de recompensa*: ou "limiar de saturação" da capacidade disponível da rede secundária conforme o número de canais disponíveis, onde mesmo que se aumente o número de canais, a recompensa não acompanha na mesma proporção; e,
    * Possui complexidade computacional baixa, tornando-se atrativa para ser embarcada no rádio cognitivo.

- Em *segundo* lugar, nós comparamos nossas propostas para a solução do problema de exploração do espectro de RF (*spectrum exploitation*) com um conjunto de outras soluções para o mesmo problema, obtidas da literatura. Os resultados da nossa avaliação baseada em simulação mostraram que o nosso mecanismo:
  - Forneceu uma ordem dinâmica de sensoreamento promissora, que manteve um desempenho superior, mesmo quando houve variação no nível de atividade dos primários, o que acarreta variação das oportunidades disponíveis para os secundários; e,
  - Obteve desempenho superior aos outros tipos de ordenamento que foram avaliados, mesmo variando o número de secundários, para diferentes valores do fator de utilização dos canais pelos primários;

5.3 Trabalhos Futuros

Uma deficiência do nosso trabalho é desconsiderarmos como possíveis erros no sensoreamento, por exemplo devido aos efeitos de sombreamento e multicaminho, podem afetar o mecanismo proposto. Essa deficiência pode ser sanada considerando essa análise como trabalhos futuros.

Como complemento para o trabalho, estão a implementação de novas estratégias para o balanceamento do dilema investigação-exploração visando reduzir a quantidade de parâmetros necessários para a configuração do mecanismo, por exemplo, através de meta parâmetros, e uma evolução do nosso mecanismo de forma que ele tenha capacidade de restrição autônoma da quantidade de usuários, visando atingir uma recompensa maior, evitando a saturação da rede, conforme comentado na Seção 3.3.3.

Até o momento, assumimos que o alcance dos primários se estende sobre todos os secundários, ou seja, a visão da rede é a mesma para todos na rede secundária. Isso nem sempre é verdade em decorrência de uma série de fatores, como, por exemplo, a informação imperfeita do canal, erros de detecção e acerca do alcance de influência do usuário primário, entre outros. Com isso, também como trabalho futuro, se encontra em estudo a possibilidade dos usuários secundários utilizarem diferentes subconjuntos das sequências possíveis, motivado por suas visões individuais da rede, visando a formação de *clusters*. Como resultado, essa abordagem pode reduzir o número de colisões na rede secundária e, também, o percentual de indisponibilidade dos canais devido a sua utilização pelos usuários primários.

Em uma outra abordagem, pode-se realizar uma modificação do nosso mecanismo para que um mapa de utilização de canais pelos usuários primários seja construído dinamicamente. Esse mapa poderia servir para efetivamente realizar a previsão, através de controles probabilísticos, do próximo canal a ser ocupado por um usuário primário e, com isso, evitá-lo, reduzindo a possibilidade de interferência curta com esse mesmo usuário primário, entre outros benefícios.

A implementação da arquitetura em ambiente de simulação abre um grande conjunto de possibilidades de trabalhos futuros. Implementamos um simulador em linguagem *Tcl*, com a adoção de uma camada de enlace específica, existindo portanto contenção. Uma evolução natural seria a avaliação do roteamento de pacotes, através de protocolos devidamente implementados, e o detalhamento necessário para o desenvolvimento de um mecanismo completo, da camada de rede até a física.

Uma vez que o nosso mecanismo é independente do conhecimento prévio e preciso das estatísticas dos canais, poderia ser adaptado para que se torne um bom estimador da estatística probabilidade de disponibilidade de canais e, com isso, aplicar esse mecanismo em um protocolo de roteamento do tipo estado do enlace, que

se aproveita dos canais com maior disponibilidade, a fim de maximizar uma métrica de interesse (por exemplo, a vazão).

E, por fim, entendemos que uma avaliação rigorosa baseada em um simulador amplamente reconhecido pela comunidade, como o NetSim [108], o ns-2 [109] e o ns-3 [110], se faz necessária como complemento do trabalho realizado. Assim, a implementação do mecanismo em um desses simuladores e a sua avaliação podem ser incluídas também como trabalhos futuros.

Referências Bibliográficas

[1] FCC. "Home | FCC.gov". 2011. `http://www.fcc.gov` - último acesso em 3/11/2011.

[2] ANATEL. "Portal Anatel". 2011. `http://www.anatel.gov.br` - último acesso em 3/11/2011.

[3] M. A. MCHENRY. "NSF Spectrum Occupancy Measurements Project", Shared Spectrum Company Report, Aug. 2005, 2005.

[4] 802.11. "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications", IEEE Standard, 1999.

[5] HAYKIN, S. "Cognitive radio: brain-empowered wireless communications", *IEEE J. Select. Areas Commun*, v. 23, n. 2, pp. 201–220, fev. 2005.

[6] WALKO, J. "Cognitive Radio", *IEE Review*, v. 51, n. 5, pp. 34–37, maio 2005. ISSN: 0953-5683.

[7] AKYILDIZ, I. F., LEE, W.-Y., VURAN, M. C., et al. "NeXt Generation/Dynamic Spectrum Access/Cognitive Radio Wireless Networks: A Survey", *Computer Networks*, v. 50, n. 13, pp. 2127 – 2159, 2006. ISSN: 1389-1286. doi: 10.1016/j.comnet.2006.05.001.

[8] ZHAO, Q., SADLER, B. M. "A Survey of Dynamic Spectrum Access: Signal Processing, Networking, and Regulatory Policy", *IEEE Signal Processing Magazine*, pp. 79–89, maio 2007.

[9] FCC. *FCC-03-322 - NOTICE OF PROPOSED RULE MAKING AND ORDER*. Relatório técnico, Federal Communications Commission, 30 dez. 2003.

[10] MITOLA III, J., MAGUIRE, G. Q. "Cognitive Radio: Making Software Radio more Personal", *IEEE Pers. Communications*, v. 6, n. 4, pp. 13–18, ago. 1999.

[11] MITOLA, J. *Cognitive radio: An integrated agent architecture for software defined radio*. Tese de Doutorado, Royal Institute of Technology, maio 2000.

[12] AKYILDIZ, I. F., LEE, W.-Y., VURAN, M. C., et al. "Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey", *Computer Networks: The Int. J. of Comp. and Telecom. Networking*, v. 50, pp. 2127–2159, September 2006.

[13] REN Y., D. P., P., K. "Analysis and Implementation of Reinforcement Learning on a GNU Radio Cognitive Radio Platform]". In: *5th International Conference on Cognitive Radio Oriented Wireless Networks and Communications, (CROWNCOM 2010), Cannes (France)*, <undefined> 2010.

[14] NIYATO, D., HOSSAIN, E. "Cognitive Radio Networks". pp. 179–214, CRC, 2008.

[15] P. DEMESTICHAS, G. DIMITRAKOPOULOS, J. STRASSNER. "Introducing Reconfigurability and Cognitive Networks Concepts in the Wireless World". In: *IEEE Vehicular Tech. Magazine*, 2006.

[16] KHALIFE, H., MALOUCH, N., FDIDA, S. "Multihop Cognitive Radio Networks: To Route or Not to Route", *IEEE Network*, v. 23, n. 4, pp. 20–25, jul. 2009.

[17] CHOW, Y. S., ROBBINS, H., SIEGMUND, D. *Great Expectations: The Theory of Optimal Stopping*. Boston, Houghton Mifflin Company, 1971.

[18] H. KIM, K. G. SHIN. "Fast Discovery of Spectrum Opportunities in Cognitive Radio Networks". In: *IEEE DySPAN*, 2008.

[19] H. JIANG, L. LAI, R. FAN, et al. "Optimal Selection of Channel Sensing Order in Cognitive Radio", *IEEE Transactions in Wireless Communications*, jan. 2009.

[20] RAPPAPORT, T. "Wireless Communications: Principles and Practice". Prentice-Hall, 2001.

[21] HALPERN, J., MOSES, Y. "Knowledge and common knowledge in a distributed environment", *Journal of the ACM*, 1990.

[22] DURFEE, E. H., LESSER, V. R., CORKILL, D. D. "Coherent cooperation among communicating problem solvers", *IEEE IEEE Transactions on Computers*, 1984.

[23] MITOLA, J. “Cognitive radio for flexible mobile multimedia communications”. In: *MoMuC ’99: IEEE International Workshop on Mobile Multimedia Communications*, pp. 3–10, nov. 1999.

[24] AKYILDIZ, I. F., LEE, W.-Y., VURAN, M. C., et al. “A Survey on Spectrum Management in Cognitive Radio Networks”, *IEEE Communications Magazine*, v. 46, n. 4, pp. 40–48, april 2008. ISSN: 0163-6804. doi: 10.1109/MCOM.2008.4481339.

[25] SALAMEH, H. A. B., KRUNZ, M. M., YOUNIS, O. “MAC Protocol for Opportunistic Cognitive Radio Networks with Soft Guarantees”, *IEEE Transactions on Mobile Computing*, v. 8, n. 10, pp. 1339–1352, out. 2009.

[26] CESANA, M., CUOMO, F., EKICI, E. “Routing in Cognitive Radio Networks: Challenges and Solutions”, *Ad Hoc Networks*, v. In Press, Corrected Proof, pp. –, 2010. ISSN: 1570-8705. doi: DOI:10.1016/j.adhoc.2010.06.009. Disponível em: <http://www.sciencedirect.com/science/article/B7576-50G6W7W-2/2/d5b461fdc68da442089f14324c0d948d>.

[27] SUTTON, R. S., BARTO, A. G. *Reinforcement Learning: an Introduction*. Cambridge, MP, 1998. Cambridge Univ Press.

[28] WATKINS, C. J. *Learning from Delayed Rewards*. Tese de Doutorado, King’s College, maio 1989.

[29] YAU, K. A., KOMISARCZUK, P., TEAL, P. D. “Applications of Reinforcement Learning to Cognitive Radio Networks”. In: *IEEE International Conference in Communications (ICC)*, jul. 2010.

[30] WATKINS, C. J., DAYAN, P. “Q-learning”, *Machine Learning*, v. 8, n. 3-4, pp. 279–292, 1992. Springer.

[31] YAU, K.-L. A., KOMISARCZUK, P., TEAL, P. D. “Enhancing Network Performance in Distributed Cognitive Radio Networks using Single-Agent and Multi-Agent Reinforcement Learning”. In: *IEEE LCN*, <undefined> 2010.

[32] KOK, J., VLASSIS, N. “Collaborative Multiagent Reinforcement Learning by Payoff Propagation”, *J. Mach. Learn. Res.*, v. 7, pp. 1789–1828, December 2006. ISSN: 1532-4435. Disponível em: <http://portal.acm.org/citation.cfm?id=1248547.1248612>.

[33] DAW, N. D., O’DOHERTY, J. P., DAYAN, P., et al. “Cortical substrates for exploratory decisions in humans”, *Nature*, 2006.

[34] NEIMAN, T., LOEWENSTEIN, Y. "Reinforcement learning in professional basketball players", *Nature*, 2011.

[35] S. GUHA, K. MUNAGALA, S. SARKAR. "Approximation Schemes for Information Acquisition and Exploitation in Multichannel Wireless Networks". In: *44th. Allerton Conference*, 2006.

[36] J. JIA, Q. ZHANG, X. SHEN. "HC-MAC: a Hardware Constrained Cognitive MAC for Efficient Spectrum Management", *IEEE JSAC*, jan. 2008.

[37] HAN HAN, JIN-LONG WANG, QI-HUI WU, et al. "Optimal Wideband Spectrum Sensing Order Based on Decision-making Tree in Cognitive Radio", *International Conference on Wireless Communications and Signal Processing (WCSP)*, 2010.

[38] ZHANG, Y., ZHANG, Q., CAO, B., et al. "Model Free Dynamic Sensing Order Selection for Imperfect Sensing Multichannel Cognitive Radio Networks: a Q-Learning Approach". In: *IEEE International Conference on Communication Systems*, 2014.

[39] R. FAN, H. JIANG. "Channel sensing order setting in cognitive radio networks: a two user case", *IEEE Transactions on Vehicular Technology*, 11 2009.

[40] ZHAOWEI, Q., RONG, C., QIZHU, S., et al. "Predictive spectrum sensing strategy based on reinforcement learning", *IEEE Communications China*, 2014.

[41] CLAUS, C., BOUTILIER, C. "The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems", *National Conference on Artificial Intelligence*, 1998.

[42] TAN, M. "Multi-agent Reinforcement Learning: Independent vs. Cooperative Agents". In: *Tenth International Conference on Machine Learning*, 1993.

[43] LAUER, M., RIEDMILLER, M. "An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-agent Systems". In: *ICML*, 2000.

[44] KAPETANAKIS, S., KUDENKO, D. "Improving on the Reinforcement Learning of Coordination in Cooperative Multi-agent Systems". In: *AAMAS*, 2002.

[45] LAUER, M., RIEDMILLER, M. "Reinforcement Learning for Stochastic Cooperative Multiagent Systems". In: *AAMAS*, 2004.

[46] VERBEECK, K., NOWÉ, A., PARENT, J., et al. "Exploring Selfish Reinforcement Learning in Repeated Games with Stochastic Rewards". In: *JAAMAS*, 2006.

[47] YAU, K. L. A., KOMISARCZUK, P., TEAL, P. D. "Reinforcement Learning for context awareness and intelligence in wireless networks: review, new features and open issues", *Journal of Network and Computer Applications*, 2012.

[48] SYED, A. R., YAU, K.-L. A., MOHAMAD, H., et al. "Channel selection in multi-hop cognitive radio network using reinforcement learning: an experimental study". In: *International Conference on Frontiers of Communications, Networks and Applications*, 2014.

[49] SALEEM, Y., YAU, K.-L. A., MOHAMAD, H., et al. "Joint channel selection and cluster-based routing scheme based on reinforcement learning for cognitive radio networks". In: *International Conference on Computer, Communications, and Control Technology*, 2015.

[50] KAKALOU, I., PAPADIMITRIOU, G. I., NICOPOLITIDIS, P., et al. "A Reinforcement learning-based cognitive MAC protocol". In: *IEEE International Conference on Communications*, 2015.

[51] ELIAS, J., MARTIGNON, F., CAPONE, A., et al. "Non-cooperative spectrum access in cognitive radio networks: a game theoretical model", *IEEE Computer Networks*, 2011.

[52] BOWLING, M., VELOSO, M. "Rational and Convergent Learning in Stochastic Games", *17th International Conference on Artficial Intelligence (IJCAI 01)*, 2001.

[53] BOWLING, M., VELOSO, M. "Multiagent Learning Using a Variable Learning Rate", *Artificial Intelligence 136 (2)*, 2002.

[54] BOWLING, M. "Convergence and No-Regret in Multiagent Learning", *Advances in neural information processing systems*, v. 17, pp. 209–216, 2005.

[55] JAFARI, A., GREENWALD, A., GONDEK, D., et al. "On No-Regret Learning, Fictitious Play and Nash Equilibrium", *18th Inter. Conference on Machine Learning*, 2001.

[56] ZAPECHELNYUK, A. *Limit Behavior of No-Regret Dynamics*. Relatório técnico, School of Economics, Kyiv, Ucraine, 2009.

[57] LESLIE, D., COLLINS, E. "Generalised Weakened Fctitious Play", *Games and Economic Behavior 56 (2)*, 2006.

[58] BROWN, G. *Some Notes on Computation of Games Solutions*. Research memoranda rm-125-pr, RAND Corporation, Santa Monica, California, 1949.

[59] XU, Y., WU, Q., WANG, J., et al. "Robust Multiuser Sequential Channel Sensing and Access in Dynamic Cognitive Radio Networks: Potential Games and Stochastic Learning", *IEEE Transactions on Vehicular Technology*, 2015.

[60] ALNWAIMI, G., VAHID, S., MOESSNER, K. "Dynamic Heterogeneous Learning Games for Opportunistic Access in LTE-Based Macro/Femtocell Deployments", *IEEE Transactions on Wireless Communications*, 2015.

[61] BERRY, D., FRISTEDT, B. "Bandit problems: sequential allocation of experiments", *Monographs on Statistics and Applied Probability*, 1985.

[62] MOTAMEDI, A., BAHAI, A. "Optimal channel selection for spectrum-agile low-power wireless packet switched networks in unlicensed band". In: *EURASIP*, 2008.

[63] GITTINS, J. C. "Bandit processes and dynamic allocation indices", *Journal of the Royal Statistical Society, Series B*, 1979.

[64] L. LAI, H. EL GAMAL, H. JIANG, et al. "Cognitive Medium Access: Exploration, Exploitation and Competition", *IEEE Transactions on Networking*, out. 2007.

[65] AUER, P., CESA-BIANCHI, N., FISCHER, P. "Finite time analysis of the multi-armed bandit problem", *Machine Learning*, v. 47, 2002.

[66] LIU, K., ZHAO, Q. "Distributed learning in multi-armed bandit with multiple players", *IEEE Transactions on Signal Processing*, 2010.

[67] LI, B., YANG, P., WANG, J., et al. "Almost Optimal Dynamically-Ordered Channel Sensing and Accessing for Cognitive Networks", *IEEE Transactions on Mobile Computing*, v. 13, n. 10, 10 2014.

[68] Q. ZHAO, L. TONG, A. SWAMI, et al. "Decentralized Cognitive MAC for Opportunistic Spectrum Access in ad hoc Networks: a POMDP Framework", *IEEE JSAC*, abr. 2007.

[69] YANG, L., CAO, L., ZHENG, H. "Proactive channel access in dynamic spectrum networks", *Physical Communication*, 2008.

[70] DO, C., TRAN, N., HONG, C. S. "Throughput maximization for the secondary user over multi-channel cognitive radio networks". In: *International Conference on Information Networking*, 2012.

[71] A. SABHARWAL, A. KHOSHNEVIS, E. KNIGHTLY. "Opportunistic Spectral Usage: Bounds and a Multi-band CSMA/CA Protocol", *IEEE Trans. Networking*, jun. 2007.

[72] CHENG, H. T., ZHUANG, W. "Simple Channel Sensing Order in Cognitive Radio Networks", *IEEE Journal on Selected Areas in Communications*, abr. 2011.

[73] CHANG, N. B., LIU, M. "Optimal channel probing and transmission scheduling for opportunistic spectrum access". In: *ACM MobiCom*, 2007.

[74] SHU, T., KRUNZ, M. "Throughput efficient sequential channel sensing and probing in cognitive radio networks under sensing errors". In: *Mobicom*, 2009.

[75] SAHAI, A., HOVEN, N., TANDRA, R. "Some Fundamental Limits on Cognitive Radio", *Wireless Foundations EECS, Univ. of California, Berkeley*, 2005.

[76] QUAN, Z., CUI, S., SAYED, A. H., et al. "Optimal multiband joint detection for spectrum sensing in cognitive radio networks", *IEEE Transaction in Signal Processing*, 2009.

[77] ZHAO, J., WANG, X. "Channel sensing order in multi-user cognitive radio networks". In: *IEEE DySPAN*, 2012.

[78] KHAN, Z., LEHTOMAKI, J. J., DASILVA, L., et al. "Autonomous Sensing Order Selection Strategies Exploiting Channel Access Information", *IEEE Transactions on Mobile Computing*, v. 12, 2013.

[79] CHEN, Y., CHEN, J., XU, Y., et al. "Multiuser Opportunistic Spectrum Access in Cognitive Radio Networks: An Optimal Stopping Approach with Spectrum Partition". In: *International Conference on Wireless Communications e Signal Processing*, 2013.

[80] YAO, C., WU, Q., XU, Y., et al. "Sequential Channel Sensing in Cognitive Small Cell Based on User Traffic", *IEEE Communications Letters*, 2015.

[81] RASTEGARDOOST, N., JABBARI, B. "On Channel Selection Schemes for Spectrum Sensing in Cognitive Radio Networks". In: *IEEE Wireless Communications and Networking Conference*, 2015.

[82] SHARMA, K. K., TRIVEDI, A. "An Opportunistic Channel Access Scheme with Channel Ordering for Cognitive Radio Network". In: *IEEE International Conference on Communication Systems and Network Technologies*, 2015.

[83] VUCEVIC, N., AKYILDIZ, I. F., PEREZ-ROMERO, J. "Dynamic cooperator selection in cognitive radio networks", *Ad Hoc Networks*, 2012.

[84] ZHAO, J., ZHENG, H., YANG, G. H. "Distributed coordination in dynamic spectrum allocation networks". In: *DySPAN*, 2005.

[85] SHU, T., KRUNZ, M. "Coordinated Channel Access in Cognitive Radio Networks: A Multi-Level Spectrum Opportunity Perspective". In: *INFOCOM*, 2009.

[86] CHEN, X., HUANG, J., LI, H. "Adaptive channel recommendation for dynamic spectrum access". In: *DySPAN*, 2011.

[87] DUAN, L., GAO, L., HUANG, J. "Contract-based cooperative spectrum sharing". In: *DySPAN*, 2011.

[88] EMRE, M., GUR, G., ALAGOZ, S. B. F. "CooperativeQ: Energy-efficient channel access based on cooperative reinforcement learning". In: *IEEE International Conference on Communication Workshop*, 2015.

[89] UNNIKRISHNAN, J., VEERAVALLI, V. "Algorithms for dynamic spectrum access with learning for cognitive radio", *IEEE Transactions on Signal Processing*, v. 58, n. 2, pp. 750–760, 2010.

[90] GEIRHOFER, S., TONG, L., SADLER, B. M. "A Measurement-Based Model for Dynamic Spectrum Access in WLAN Channels". In: *MILCOM '06: IEEE Military Communications Conference*, pp. 1–7, out. 2006.

[91] GEIRHOFER, S., TONG, L., SADLER, B. M. "COGNITIVE RADIOS FOR DYNAMIC SPECTRUM ACCESS - Dynamic Spectrum Access in the Time Domain: Modeling and Exploiting White Space", *Communications Magazine, IEEE*, v. 45, n. 5, pp. 66–72, may 2007. ISSN: 0163-6804. doi: 10.1109/MCOM.2007.358851.

[92] STABELLINI, L. "Quantifying and Modeling Spectrum Opportunities in a Real Wireless Environment". In: *WCNC '10: IEEE Wireless Communications and Networking Conference*, pp. 1–6, april 2010. doi: 10.1109/WCNC.2010.5506438.

[93] CHEN, D., YIN, S., ZHANG, Q., et al. "Mining spectrum usage data: a large-scale spectrum measurement study". In: *MobiCom '09: International conference on Mobile computing and networking*, pp. 13–24, set. 2009.

[94] WELLENS, M., RIIHIJÄRVI, J., MÄHÖNEN, P. "Empirical time and frequency domain models of spectrum use", *Physical Communication*, v. 2, n. 1-2, pp. 10–32, 2009. ISSN: 1874-4907.

[95] WILLKOMM, D., MACHIRAJU, S., BOLOT, J., et al. "Primary Users in Cellular Networks: A Large-Scale Measurement Study". In: *DySPAN '08: IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pp. 1–11, out. 2008. doi: 10.1109/DYSPAN.2008.48.

[96] YUCEK, T., ARSLAN, H. "A Survey of Spectrum Sensing Algorithms for Cognitive Radio Applications", *IEEE Communications Surveys & Tutorials*, v. 11, n. 1, pp. 116–130, 2009.

[97] XIN, C., MA, L. "Path-Centric Channel Assignment in Cognitive Radio Wireless Networks". In: *2nd International Conference on Cognitive Radio Oriented Wireless Networks and Communications, CrownCom*, 2007.

[98] BIAN, K., PARK, J. "Segment-Based Channel Assignment in Cognitive Radio Ad Hoc Networks". In: *2nd International Conference on Cognitive Radio Oriented Wireless Networks and Communications, CrownCom*, 2007.

[99] HOYHTYA, M., POLLIN, S., MAMMELA, A. "Performance improvement with predictive channel selection for cognitive radios". In: *First International Workshop on Cognitive Radio and Advanced Spectrum Management, CogART*, 2008.

[100] 802.22. "Standard for Cognitive Wireless Regional Area Networks (RAN) for Operation in TV Bands", IEEE Standard, 07 2011.

[101] VU, H. L., SAKURAI, T. "Collision Probability in Saturated IEEE 802.11 Networks". In: *Australian Telecommunication Networks and Applications Conference*, 2006.

[102] 802.11B. "Wireless LAN MAC and PHY Specifications: Higher-Speed Physical Layer Extension in the 2.4GHz Band", IEEE Standard, 1999.

[103] MENDES, A. C., SILVA, M. W. R. D., GUEDES, R. M., et al. "Seleção da Ordem de Sensoreamento de Canais em uma Rede Cognitiva Oportunista". In: *WRA*, jun. 2011.

[104] MENDES, A. C., AUGUSTO, C. H. P., SILVA, M. W. R. D., et al. "Channel Sensing Order for Cognitive Radio Networks using Reinforcement Learning". In: *IEEE LCN*, 2011.

[105] CAELEN, O., BONTEMPI, G. "Improving the exploration strategy in bandit algorithms". In: *Second International Conference, LION 2007 II*. Springer, 2008.

[106] OUSTERHOUT, J. K., JONES, K. *Tcl and the Tk toolkit*, v. 227. Massachussets, Addison-Wesley Reading, 1994.

[107] MENDES, A. C., SILVA, M. W. R. D., GUEDES, R. M., et al. "Ordem de Sensoreamento de Canais em Redes de Rádios Cognitivos Multi-usuário". In: *WPERFORMANCE*, 2012.

[108] TETCOS. "NetSim Academic Version". 2015.

[109] NS-2. "The Network Simulator - ns-2". .

[110] NS 3. "ns-3". 2015.

[111] MORIHIRO, K., MATSUI, N., NISHIMURA, H. "Effects of Chaotic Exploration on Reinforcement Maze Learning", *KES LNCS*, 2004.

[112] YAU, K. A., KOMISARCZUK, P., TEAL, P. D. "Learning Mechanisms for Achieving Context Awareness and Intelligence in Cognitive Radio Networks". In: *4th. IEEE Workshop on Wireless and Internet Services*, 2011.

[113] JAIN, R., CHIU, D. M., HAWE, W. R. "A quantitative measure of fairness and discrimination for resource allocation in shared computer systemsf Fairness And Discrimination For Resource Allocation In Shared Computer Systems", *ACM Transaction on Computer Systems*, 1984.

[114] FCC. *FCC-08-260A1 - SECOND REPORT AND ORDER*. Relatório técnico, Federal Communications Commission, 14 nov. 2008.