Modulação e Codificação Adaptativa em Sistemas OFDM Utilizando Aprendizado por Reforço

João P. Leite, Paulo H. P. de Carvalho, Robson D. Vieira

***Resumo*— Este artigo apresenta uma solução baseada em aprendizado por reforço para o problema de modulação de codificação adaptativas em sistemas OFDM. Embora técnicas de aprendizado de máquina já sejam encontradas na literatura para os problemas de adaptação de enlace, elas são fundamentadas em algoritmos *off-line* de aprendizado supervisionado, o que as torna pouco atrativas para sistemas em tempo real e de grande variabilidade. A solução proposta é capaz de aprender, por meio da interação direta com o ambiente de rádio, qual a melhor combinação de modulação e codificação para uma dada razão sinal-ruído, sem depender de um supervisor ou de muitos recursos computacionais. Os resultados obtidos mostram que, sob condições específicas, a técnica proposta pode superar a solução clássica, que utiliza tabelas de consulta, e pode adaptar-se a características especificas do ambiente de rádio.**

***Palavras-Chave*— Adaptação de enlace, aprendizado por reforço, modulação e codificação adaptativas, OFDM.**

***Abstract*— This paper presents a reinforcement learning approach for link adaptation in OFDM systems through adaptive modulation and coding. Although machine learning techniques have attracted attention for link adaptation, most of the schemes proposed so far are based on off-line training algorithms, what makes them not well suited for real time operation. The proposed solution, based on the reinforcement learning technique, learns the best modulation and coding scheme for a given signal-to-noise ratio by interacting with the radio channel and it does not rely on an off-line training mode. Simulation results show that under specific conditions, the proposed technique can outperform the solution based on look-up tables and it can potentially adapt itself to distinct characteristics of the radio environment.**

***Keywords*— Adaptive Modulation and coding, link adaptation, machine learning, orthogonal frequency-division multiplexing, reinforcement learning.**

I. Introdução

A modulação e codificação adaptativa (AMC, *adaptive modulation and coding*) é uma das diversas técnicas propostas para aumentar a capacidade dos sistemas de comunicação, sobretudo para as tecnologias de terceira e quarta gerações, em que altas taxas de transmissão são requeridas. Para tal, é necessário explorar o conhecimento das condições de canal para modificar os parâmetros de modulação e codificação, de forma a aumentar a vazão efetiva dos sistemas de comunicação. Atualmente, uma das forma de implementação de AMC é por meio das tabelas de consulta (*look-up tables*) [1], que contém os valores limiares de razão sinal-ruído (SNR, *signal-to-noise ratio*) que são utilizados para realizar a mudança entre as diferentes combinações de modulação e codificação. Entretanto, a técnica apresenta algumas desvantagens. Em primeiro lugar, o erro resultante da adaptação pode exibir variância elevada, pois este é sensível à escolha da métrica utilizada para representar a qualidade do enlace [2]. Em segundo lugar, as tabelas de consulta não são determinadas em tempo real: sua obtenção depende de grande tempo de simulação computacional e, em geral, não refletem características peculiares que podem ser encontradas em determinados meios de propagação e em diferentes transceptores [2].

Recentemente, foi proposta a utilização de técnicas de aprendizado de máquina para a implementação de AMC. Em [2], [3], a adaptação de enlace é modelada como um problema de classificação, resolvido por meio do algoritmo kNN (*k-nearest neighbors*). Em [4], uma máquina de vetor de suporte (SVM, *support vector machine*) é sugerida como forma alternativa de resolver o mesmo problema. Apesar de exibirem melhor capacidade de adaptação, estas técnicas dependem fortemente do conjunto de dados que é utilizado para seu treinamento, pois constituem formas de aprendizado supervisionado. Frequentemente, é impraticável obter exemplos e dados que sejam acurados e significativos para o treinamento e aprendizado (sob o ponto de vista do número de diferentes situações com as quais o transmissor possa se deparar) [5]. Neste sentido, as soluções descritas são pouco viáveis para aprendizado em tempo real, o que sugere outras formas de abordar o problema.

Neste contexto, técnicas de aprendizado por reforço têm atraído alguma atenção para o campo dos sistemas de comunicação, sobretudo no contexto de rádios cognitivos, em que a capacidade de adaptação é exigida tanto para os transmissores quanto para os receptores [6], [7]. Entre as vantagens do aprendizado por reforço, estão a capacidade de aprender diretamente com o ambiente, sem a necessidade de um supervisor ou um especialista para seu treinamento, e sua capacidade de operação em tempo real. Pode-se citar ainda que são métodos independentes de modelo (*model free*), o que os torna atrativos quando a parametrização e a modelagem analítica do problema se tornam muito complexas, como é o caso do AMC em sistemas OFDM (*orthogonal frequency-division multiplexing*) [2], [6].

Este trabalho apresenta como contribuição a proposta de uma solução baseada em aprendizado por reforço para a solução do problema de AMC. Utilizando os resultados de interações passadas com o canal rádio móvel, um agente é capaz de aprender quais são as melhores combinações de modulação e codificação para um dado estado do canal, fazendo um número reduzido de conjecturas sobre o ambiente. O problema da escolha da técnica de modulação e codificação é tratado como um processo de decisão de Markov cujo objetivo é maximizar a eficiência espectral do sistema, e para

João P. Leite, Paulo H. P. de Carvalho, Departamento de Engenharia Elétrica (ENE), Universidade de Brasília (UnB), Brasília, Brasil. Robson D. Vieira, Instituto Nokia de Tecnologia (INdT), Brasília, Brasil.

sua solução é utilizado o algoritmo *Q-learning*. Não é objetivo do trabalho investigar métricas que possam ser utilizados para a caracterização do canal de rádio [3], ainda que o aprendizado possa ser facilitado pela utilização de novas métricas.

O artigo está organizado da seguinte forma: na Seção II é apresentado o modelo do sistema OFDM utilizado. A Seção III apresenta a teoria sobre aprendizado por reforço e a solução proposta para o problema de AMC. Na Seção IV são apresentados os resultados de simulação do desempenho da abordagem proposta, e as conclusões do trabalho se encontram na Seção V.

II. Modelo do Sistema

Será considerado o enlace direto de um sistema OFDM cujos procedimentos de transmissão são similares aos encontrados nos padrões 3GPP LTE (*Third-Generation Partnership Project Long Term Evolution*) e WiMAX (*Worldwide Interoperability for Microwave Access*). A transmissão é feita em pacotes, que são alocados em um conjunto de subportadoras ao longo de um ou mais símbolos OFDM dentro de um mesmo quadro de transmissão. Nesta alocação, todas as subportadoras que transmitem um determinado pacote utilizam a mesma técnica de modulação e de codificação. Cada pacote possui um campo CRC (*cyclic redundancy check*) que permite verificar a taxa de erro de pacote. A informação transmitida é também protegida por meio de um código corretor de erros do tipo convolucional.

A cada símbolo OFDM transmitido, é inserido um prefixo cíclico, de forma a eliminar a interferência intersimbólica causa pela transmissão da informação em um canal com múltiplos percursos. Supõe-se que a resposta do canal possa variar consideravelmente entre símbolos OFDM diferentes, mas não dentro do mesmo símbolo (desvanecimento por blocos). No receptor, o sinal recebido é equalizado utilizando-se o algoritmo *zero-forcing*, e a informação é decodificada utilizando-se o algoritmo de Viterbi. Detalhes mais especificos serão considerados na Seção IV.

III. Aprendizado por Reforço

O conceito de um sistema que opera utilizando aprendizado por reforço é mostrado de modo simplificado na Fig. 1. Há um agente, que interage com o ambiente por meio de ações, e percebe os efeitos destas ações por meio de um sinal de recompensa e pela transição do ambiente para um novo estado. O objetivo do agente é maximizar, para todos os estados, uma soma ponderada de todos os sinais de recompensa recebidos ao longo do tempo de operação [8]. Diferentemente de técnicas de aprendizado supervisionado, o agente aprende diretamente de suas experiências de interação com o ambiente.

Como será detalhado posteriormente, o interesse em aplicar aprendizado por reforço para o problema de AMC é para maximizar a vazão do sistema para um dado estado do ambiente, representado pelo canal de rádio e descrito por meio da razão sinal-ruído. O processo de otimização é dirigido por meio da seleção de um conjunto de combinações entre modulação e codificação.

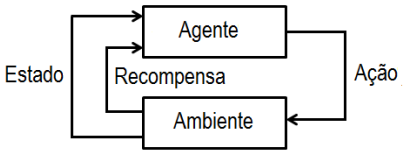

Fig. 1: Diagrama de blocos para um sistema de aprendizado por reforço.

A. Processo de Decisão de Markov

Problemas de aprendizado por reforço podem ser formalizados utilizando a teoria de processos de decisão de Markov [9], descritos por meio de 4 elementos:

- $\mathcal{S} = \{s_1, s_2, \ldots s_n\}$ é um conjunto de n possíveis estados, que descrevem o ambiente;
- $\mathcal{A} = \{a_1, a_2, \ldots a_m\}$ é um conjunto de m possíveis ações que o agente pode tomar;
- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0; 1]$ é um modelo de transição markoviano que descreve a dinâmica do ambiente, sendo $P(s, a, s')$ a probabilidade de transição para o estado $s' \in \mathcal{S}$ ao tomar a ação $a \in \mathcal{A}$ no estado $s \in \mathcal{S}$;
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ é uma função recompensa, em que $R(s, a, s')$ é uma recompensa imediata recebida pelo agente e originada da transição do estado s para o estado s' ao tomar a ação a.

Uma política π representa o comportamento do agente, e é definida como um mapeamento entre ações e estados. Matematicamente, $\pi : \mathcal{S} \rightarrow \mathcal{A}$. A notação $\pi(s)$ indica a ação que deve ser tomada no estado s. O valor do estado sob a política π, ou função-V, representado por $V^{\pi}(s)$, define a recompensa cumulativa esperada ao seguir a política π partir do estado s [10]:

$$V^{\pi}(s) = \mathbb{E}\left\{\sum_{t=0}^{\infty} \gamma^t r_t\right\}, \tag{1}$$

em que r_t é a recompensa recebida no instante t, e γ, com $0 \leq \gamma \leq 1$ é um fator de desconto para as recompensas futuras, e determina a importância relativa destas com relação à recompensa recebida no instante t . Valores próximos de zero geram um agente imediatista, ou míope, enquanto que valores próximos de 1 valorizam igualmente as recompesas presentes e futuras.

O objetivo do agente em um problema de aprendizado por reforço é encontrar a política ótima $\pi^*(s)$ que maximiza (1), isto é, deve-se obter $V^*(s) = \max_{\pi} V^{\pi}(s), \forall s \in \mathcal{S}$.

A forma mais usual de caracterizar uma política é por meio da função-Q, que representa o quão adequado (sob a ótica de maximizar a soma cumulativa das recompensas) é tomar a ação a quando o ambiente se encontra no estado s e, a partir deste ponto, seguir a política π [10].

Sua definição é dada por:

$$Q^{\pi}(s, a) = \mathbb{E}\left\{\sum_{t=0}^{\infty} \gamma^t r_t | s_{t=0} = s, a_{t=0} = a\right\}. \tag{2}$$

Definindo-se $Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$, pode-se mostrar que, sob a política ótima, é válida a igualdade [8]

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a). \quad (3)$$

Dessa forma, a política ótima pode ser obtida selecionando a ação que maximiza $Q(s,a)$ para um dado estado s, isto é,

$$\pi^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a). \quad (4)$$

Para a maioria dos problemas de aprendizado por reforço, não existe o conhecimento prévio das quantidades $\mathcal{R}(s,a)$ e $P(s,a,s')$, de tal forma que estas devem ser estimadas indiretamente para a obtenção da função-Q e, consequentemente, para a obtenção da política ótima. Esta tarefa pode ser realizada por meio de algoritmos como SARSA (*State-Action-Reward-State-Action*) e *Q-learning* [11].

Neste trabalho, é utilizado o algoritmo *Q-learning* para a solução do problema de aprendizado por reforço. A quantidade $Q^*(s,a)$ é estimada de forma recursiva e iterativa por meio da relação de atualização

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_a Q(s',a) - Q(s,a) \right], \quad (5)$$

em que α é a taxa de aprendizado.

É necessário lidar com um último aspecto de problemas de aprendizado por reforço, conhecido como dilema *exploration vs. exploitation* [8]. Como considerado anteriormente, a dinâmica do ambiente não é conhecida. Dessa forma, deve-se buscar um equilíbrio entre coletar informações sobre o ambiente para que sua dinâmica seja aprendida (*exploring*), selecionando ações diferentes para um mesmo estado, e tirar proveito das melhores ações determinadas até o momento (*exploiting*). Para lidar com este aspecto do problema, foi utilizado o método *softmax* para seleção de ações [8], no qual a próxima ação a é selecionada com probabilidade $Pr(a)$ dada por

$$Pr(a) = \frac{e^{Q(s,a)/\tau}}{\sum_{i=1}^{m} e^{Q(s,a_i)/\tau}}, \quad (6)$$

em que τ controla o grau de exploração para a seleção das ações. Valores próximos de zero geram um comportamento guloso (*greedy*) para o algoritmo.

B. Solução Proposta

Considera-se que o problema de AMC pode ser tratado como uma versão do problema *one-armed bandit* [12], no qual existe um conjunto de estados diversas ações $a \in \mathcal{A}$ com diferentes recompensas esperadas, que podem possuir características variantes no tempo (o que justifica a utilização de técnicas como o aprendizado por reforço). O principal objetivo é maximizar a soma das recompensas a longo prazo por meio de uma determinada sequência de ações. Como consequência, não há relação imediata entre s_{t+1} e a_t.

O estado do sistema é determinado por meio da razão sinal-ruído média do conjunto de subportadoras que compõem um bloco de recursos (ver seção IV). Por razões de ordem prática, os valores de SNR foram restritos ao intervalo $[0; 40]$ dB. Para evitar lidar com um conjunto infinito de estados, a SNR foi discretizada em passos de 1 dB, totalizando 41 estados. Dessa forma, cabe ao agente descobrir, para cada estado do sistema, o melhor esquema de modulação e codificação para um dado valor de SNR, no sentido de maximizar a eficiência espectral. Consequentemente, o conjunto de ações $\mathcal{A}$ é formado por todas as combinações disponíveis de modulação e codificação, como citado no início da Seção III.

Por fim, define-se a função recompensa R como a vazão alcançada pelo sistema ao tomar a ação a quando o ambiente encontra-se no estado s, que é dada por:

$$R(s,a,s') = \log_2(M_a)\, \rho_a \left[1 - PER(s,a)\right] \quad (7)$$

em que M_a é a ordem da modulação da ação a, ρ_a é a taxa de codificação da ação a e $PER(s,a)$ é a taxa de erro de pacote (PER, *packet error rate*) resultante da ação a sobre o estado de canal s. Como um campo CRC está presente em cada pacote, é possível ao receptor identificar pacotes recebidos com erro, e a PER pode estão ser estimada [13]. Esta estimativa é enviada ao transmissor para que (7) possa ser calculada e, por meio do algoritmo *Q-learning*, o melhor esquema de modulação e codificação possa ser encontrado (em termos de minimizar a PER, aumentando a vazão efetiva do sistema).

IV. Simulação e Resultados

O desempenho do algoritmo de aprendizado por reforço para a adaptação de enlace foi avaliado por meio de simulações computacionais, e foi comparado com a solução por tabelas de consulta.

A. Parâmetros do Sistema

Para propósitos de simulação, foi considerado o enlace direto do padrão 3GPP LTE, com largura de banda de 10 MHz. Um quadro de transmissão possui duração de 10 ms e contém 10 subquadros, cada um de duração 1 ms. Cada subquadro é dividido em 2 *slots*, cada um transportando 7 símbolos OFDM. O espaçamento entre as subportadoras é de 15 kHz, e o prefixo cíclico possui duração de 4,6 μs [14]. Transmissor e receptor possuem apenas uma antena. As combinações possíveis entre as formas de modulação e codificação são mostradas na Tabela I. O código corretor de erros é do tipo convolucional, com taxas de código 1/2, 2/3 e 3/4. O codificador base é de taxa 1/2 com geradores $[133, 171]$ (em octal), e é utilzado puncionamento para a obtenção das demais taxas.

Para considerar um cenário mais realista, o modelo de canal variante no tempo utilizado para as simulações foi o *Spatial Channel Model Extended (SCME)*, modelo de referência padronizado pelo 3GPP [15] e implementado pelo conjunto de funções em [16]. Os valores utilizados para os diferentes parâmetros de canal estão especificados na Tabela II.

Quanto ao algoritmo de aprendizado por reforço, tem-se um conjunto de $m = 6$ ações, correspondentes a cada uma das combinações possíveis de modulação e codificação. Exceto quando indicado o contrário, utilizou-se $\gamma = 0,65$ para o fator de desconto, $\alpha = 0,5$ para a taxa de aprendizado do algoritmo

TABELA I: Esquemas de Modulação e Codificação.

Número do esquema (Ação m)	Modulação	Taxa de Código
1	QPSK	1/2
2	QPSK	3/4
3	16QAM	1/2
4	16QAM	3/4
5	64QAM	2/3
6	64QAM	3/4

TABELA II: Parâmetros do modelo SCME.

Parâmetros	Valor
Frequência central da portadora	2 GHz
Velocidade do móvel	11 m/s
Número da antenas na estação rádio-base	4
Número de antenas na estação móvel	1
Cenário	*Suburban Macro*
Número de múltiplos percursos	19

Q-learning e $\tau = 0,3$ no critério softmax para escolha de ações. Estas escolhas serão justificadas na próxima subseção.

Um total de 500 simulações foram consideradas, e, em média, 50 pacotes foram utilizados para a estimativa da probabilidade de erro de pacote em cada estado.

B. Tabelas de Consulta

As tabelas de consulta foram geradas utilizando-se o mapeamento do tipo *RawBER* [1], e os limiares de SNR foram definidos utilizando-se um limite de 10% para a probabilidade de erro de pacote, em um canal com ruído branco aditivo gaussiano.

Uma das desvantagem do método, além do tempo de simulação computacional requerido, reside no fato do desempenho da estratégia depender das condições de simulação para as quais as tabelas foram obtidas [1]. Quando há presença de interferência, muitas vezes a hipótese de que esta e ruído branco podem ser modelados por meio de apenas uma distribuição gaussiana não é válida [5]. Claramente, é impraticável simular e gerar dados para todas as situações possíveis e, consequentemente, tabelas de consulta podem conduzir a soluções subótimas (a mesma observaçao é válida para grande parte das técnicas baseadas em aprendizado supervisionado).

C. Resultados de Simulação

As Fig. 2 and Fig. 3 mostram o efeito do fator de esquecimento (γ) e da taxa de aprendizado (α) na convergência do algoritmo. O erro quadrático médio (MSE) foi calculado considerando-se a diferença entre a vazão do sistema sob o melhor esquema de modulação e codificação disponível para uma dada SNR e a política aprendida pelo algoritmo *Q-learning* ao longo de suas iterações.

Como mostrado na Fig. 2, quanto maior o valor da taxa de aprendizado, mais rápida é a convergência do algoritmo. A escolha $\alpha = 0,5$ resulta em um tempo de convergência de aproximadamente 250 quadros (inferior a três segundos), consistindo em uma escolha razoável, de acordo com as necessidades de adaptação do transmissor e do receptor.

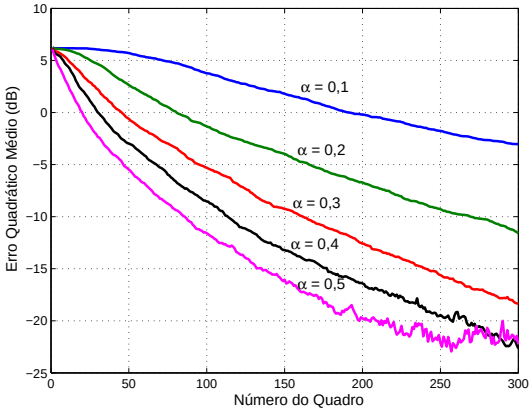

Fig. 2: Influência da taxa de aprendizado α na convergência do algoritmo *Q-learning* para $\gamma = 0,65$.

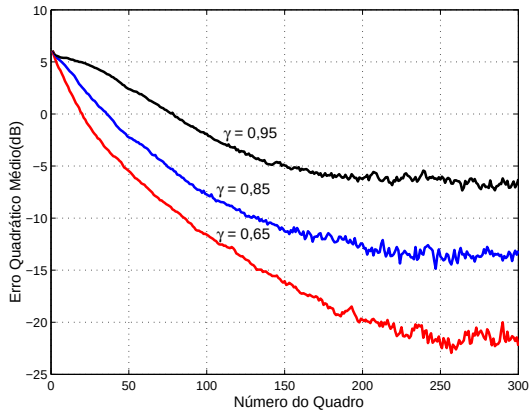

Fig. 3: Influência do fator de desconto γ na convergência do algoritmo *Q-learning* para $\alpha = 0,5$.

A Fig. 3 mostra a influência do fator de desconto na convergência do algoritmo. Um valor baixo para o fator de desconto implica em um comportamento míope do agente, uma vez que recompensas imediatas serão preferidas em detrimento do comportamento a longo prazo.

A Fig. 4 mostra a eficiência espectral do sistema em função da SNR em um canal com ruído do tipo aditivo gaussiano. As técnicas de aprendizado por reforço e tabelas de consulta apresentam desempenho similar, uma vez que a métrica de qualidade de enlace utilizada é a mesma para ambas. A perda de vazão observada no algoritmo de aprendizado por reforço deve-se à quantização dos valores de SNR. Determinar a melhor partição para o espaço de estados pode ser um problema complexo: como os limiares de AMC não são conhecidos *a priori*, uma discretização muito grosseira pode causar grande perda de capacidade, ao passo que uma discretização mais pormenorizada aumenta consideravelmente o número de estados do sistemas, aumentando de forma proibitiva o tempo de convergência do algoritmo. O passo de discretização escolhido, de 1 dB, parece uma solução razoável em termos de dimensionalidade e perda de desempenho. Entretanto, cabe observar que a técnica de aprendizado por reforço opera

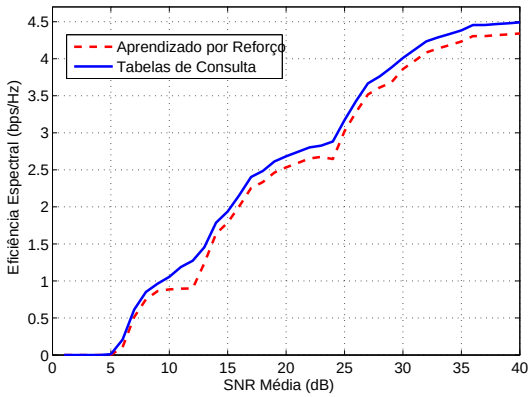

Fig. 4: Eficiência espectral média das técnicas de aprendizado por reforço e tabelas de consulta em um canal com ruído branco gaussiano.

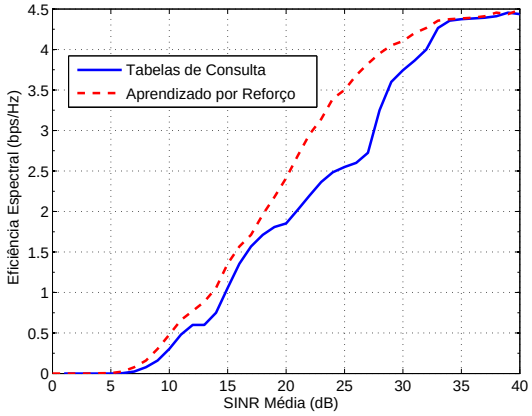

Fig. 5: Eficiência espectral média das técnicas de aprendizado por reforço e tabelas de consulta em um canal com ruído branco gaussiano e interferência colorida.

em tempo real e não há necessidade de um especialista ou um conjunto grande de simulações computacionais *off-line* para a obtenção dos limiares de SNR que são utilizados nas abordagens tradicionais de AMC.

Finalmente, a Fig. 5 considera a eficiência espectral das técnicas na presença de interferência colorida. Além do ruído branco gaussiano, considerou-se também um segundo sinal OFDM interferente, cuja potência média era oito vezes superior à variância do ruído branco. Este cenário exibe um dos problemas das tabelas de consulta: em geral é difícil obter dados representativos para sua construção. Mesmo em caso de algoritmos de aprendizado supervisionado, estes podem depender fortemente de características do sinal interferente. O algoritmo de aprendizado por reforço, por outro lado, é capaz de adaptar-se e aprender diretamente por meio de sua interação com o ambiente e mostrou melhor desempenho que a solução por tabelas de consulta. A diferença entre as abordagens pode ser superior a 1 bps/Hz, dependendo da região de operação.

V. Conclusões

Neste trabalho, foi apresentada uma proposta de para asolução do problema de modulação e codificação adaptativas com base no conceito de aprendizado por reforço. A maximização da eficiência espectral foi tratada como um processo de decisão de Markov, no qual a razão sinal-ruído média foi utilizada para caracterizar o canal móvel, e foi encontrada uma associação entre estes valores e diferentes esquemas de modulação e codificação disponíveis para transmissão. O método proposto mostrou-se apropriado para aprendizado em tempo real, pois não necessita de nenhuma etapa prévia de treinamento, diferentemente de outras soluções encontradas na literatura. Além disso, a abordagem por meio do aprendizado por reforço mostra capacidade de adaptar-se a caracterísitcas específicas do ambiente de rádio. Comparou-se também o desempenho da técnica propostra com a solução por tabelas de consulta, e estas mostraram-se subótimas quando utilizadas em cenários em que interferência colorida estava presente.

Referências

[1] S. Kant and T. L. Jensen, "Fast link adaptation for IEEE 802.11n," M. S. thesis, Aalborg University, Denmark, Feb. 2007.

[2] R. C. Daniels, C. M. Caramanis, and J. Robert W. Heath, "A Supervised Learning Approach to Adaptation in Practical MIMO-OFDM Wireless Systems," in *Proc. of IEEE Global Telecommunications Conference, 2008 - GLOBECOM 2008*, New Orleans, LO, Nov. 2008, pp. 1–5.

[3] ——, "Adaptation in Convolutionally Coded MIMO-OFDM Wireless Systems Through Supervised Learning and SNR Ordering," *IEEE Trans. Veh. Technol.*, vol. 59, no. 1, pp. 114–126, Jan. 2010.

[4] ——, "Online Adaptive Modulation and Coding with Support Vector Machines," in *Proc. of 2010 European Wireless Conference (EW)*, Lucca, Italy, Apr. 2010, pp. 718–724.

[5] M. Aljuaid and H. Yanikomeroglu, "Investigating the Gaussian Convergence of the Distribution of the Aggregate Interference Power in Large Wireless Networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 9, pp. 4418–4424, Nov. 2010.

[6] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, Mar. 2007.

[7] M. Coupechoux, J. M. Kelif, and P. Godlewski, "SMDP approach for JRRM analysis in heterogeneous networks," in *Proceedings of 14th European Wireless Conference, EW 2008.*, Prage, Czech Republic, Jun. 2008, pp. 1–7.

[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[9] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, New Jersey: Wiley-Interscience, 2005.

[10] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed. Cambridge, MA: MIT Press, 2010.

[11] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, New Jersey: Prentice Hall, 2009.

[12] A. W. M. Leslie Pack Kaelbling, Michael L. Littman, "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, May 1996.

[13] M. A. Haleem and R. Chandramouli, "Adaptive Stochastic Iterative Rate Selection for Wireless Channels," *IEEE Commun. Lett.*, vol. 8, no. 5, pp. 292–294, May 2004.

[14] 3GPP, "3GPP TR 36.211 V8.5.0: Physical Channels and Modulation (Release 8)," Tech. Rep., 2008.

[15] 3GPP, "3GPP TR 25.996 V8.5.0 - Spatial Channel Model for Multiple Input Multiple Output (MIMO) Simulations (Release 6)," Third Generation Partnership Project, Tech. Rep., Sep. 2003.

[16] J. Salo, G. Del Galdo, J. Salmi, M. Milojevic, D. Laselva, and C. Schneider. (2005, Jan.) MATLAB implementation of the 3GPP Spatial Channel Model (3GPP TR 25.996). [Online]. Available: http://www.tkk.fi/Units/Radio/scm/