Algoritmo Adaptativo Multiagente Baseado em Entropia Cruzada e Arquitetura DQN Aplicado à Alocação de Recursos em Redes 5G

Daniel Porto Queiroz Carneiro, Alisson Assis Cardoso, Robson Domingos Vieira e Flávio Henrique Teles Vieira

***Resumo*— Este artigo propõe soluções de alocação de recursos para redes sem fio 5G baseadas em CP-OFDM (*Cyclic Prefix Orthogonal Frequency Division Multiplexing*) usando abordagens de agente único e de múltiplos agentes. Algoritmos de aprendizagem por reforço de rede Q-profunda (DQN) podem ser utilizados para treinar agentes com o objetivo de alocar recursos aos usuários de forma adaptativa. Nesse sentido, apresentamos um algoritmo adaptativo de aprendizado por reforço baseado no método de entropia cruzada que aprimora o treinamento destas redes DQN. Na abordagem proposta de agente único, um agente que denominamos de CEQN (*Cross-Entropy Q-Network*) aloca os recursos em sua célula correspondente. Para a abordagem de vários agentes, propomos uma função de recompensa que utiliza informações das células vizinhas nos estados dos agentes. As simulações realizadas neste trabalho consideram dados de tráfego real, estados de canal e informações de agentes vizinhos como entradas para o algoritmo CEQN adaptativo. Os resultados comprovam a superioridade do algoritmo CEQN adaptativo proposto em termos de métricas de qualidade de serviço em comparação a outras abordagens de alocação de recursos da literatura.**

***Palavras-Chave*— CP-OFDM, Multiagente, Entropia Cruzada, DQN**

***Abstract*— This paper proposes resource allocation solutions for 5G wireless networks based on CP-OFDM (*Cyclic Prefix Orthogonal Frequency Division Multiplexing*) using both single-agent and multi-agent approaches. Deep Q-Network (DQN) reinforcement learning algorithms can be used to train agents to adaptively allocate resources to users. In this context, we present an adaptive reinforcement learning algorithm based on the cross-entropy method that enhances DQN training. In the proposed single-agent approach, an agent—referred to as CEQN (*Cross-Entropy Q-Network*)—allocates resources within its corresponding cell. For the multi-agent approach, we propose a reward function that incorporates neighboring cell information into the agent states. The simulations conducted in this work consider real traffic data, channel states, and neighboring agent information as inputs to the adaptive CEQN algorithm. The results demonstrate the superiority of the proposed adaptive CEQN algorithm in terms of quality of service metrics compared to other resource allocation approaches in the literature.**

***Keywords*— CP-OFDM, Multi-Agent, Cross-Entropy, DQN**

I. Introdução

Os sistemas de comunicação sem fio vêm evoluindo para atender demandas por baixa latência e altas taxas de transmissão, diante de múltiplas solicitações simultâneas. Assim, soluções flexíveis de alocação de recursos são essenciais.

Daniel Porto, e-mail: dcarneiro@discente.ufg.br; Alisson Assis, e-mail: alsnac@ufg.br; Robson Domingues, e-mail: robson.vieira@unb.br; Flávio Henrique, e-mail: flavio_vieira@ufg.br; Centro de Excelência em Redes Inteligentes Sem Fio e Serviços Avançados (CERISE), Escola de Engenharia Elétrica, Mecânica e de Computação (EMC), UFG-Goiânia-GO.

O aprendizado por reforço é promissor nessa tarefa, mas a exploração do espaço estado-ação pode ser computacionalmente custosa, especialmente com estados contínuos. O *Cross Entropy Method* (CEM) surge como alternativa evolutiva ao gerar e refinar amostras aleatórias de soluções candidatas, além de mitigar a convergência prematura [1], [2].

Entre os trabalhos que aplicam aprendizado por reforço à alocação de recursos, destaca-se Zhu et al. [3], que usa tal abordagem em sistemas IoT multiusuário, atendendo um usuário por TTI. De forma semelhante, propomos uma função de recompensa adaptada ao atendimento simultâneo de múltiplos usuários por TTI e consideramos um cenário com múltiplas células.

Quanto ao uso do método de Entropia Cruzada, De Boer et al. [4] destacam sua eficácia em problemas de alocação, e em [5] ele é aplicado com MAB (*Multi-Armed Bandit*) em redes com ondas milimétricas. No entanto, esses trabalhos não consideram múltiplas células, modos de transmissão variáveis ou cooperação entre agentes.

Redes 5G e pós-5G baseadas em CP-OFDM permitem alocação simultânea em frequências ortogonais, otimizando o uso para usuários heterogêneos [6]. Em trabalho anterior [7], adotamos aprendizado por reforço com um modelo Markoviano e matriz de transição. Neste artigo, propomos uma abordagem mais geral que dispensa essa suposição, utilizando redes DQN combinadas com CEM, formando o algoritmo CEQN adaptativo. Apresentamos uma nova função de recompensa aplicável a cenários com um ou múltiplos agentes. A discretização dos estados (ganhos de canal, ocupação de *buffer*) e ações permite aplicar a CEQN adaptativa com eficiência. Como mostramos neste artigo, a abordagem apresenta desempenho superior às existentes na literatura.

Este artigo está organizado da seguinte forma: a Seção II descreve a rede de comunicação e o agente de aprendizado; a Seção III apresenta o modelo do sistema; a Seção IV introduz a rede DQN e a função de recompensa proposta; a Seção V detalha a rede CEQN adaptativa; a Seção VI apresenta os resultados de simulação; e a Seção VII traz as conclusões.

II. Descrição do Sistema de Comunicação e do Agente de Aprendizado por Reforço

A rede de comunicação analisada é composta por K usuários com demandas reais da série de tráfego MAWI [8]. As simulações consideram apenas transmissão *downlink*, com estações-base e usuários equipados com antenas simples (SISO), simplificando o cenário sem perda de generalidade.

Assume-se que $M < K$ usuários são atendidos por um sistema CP-OFDM com tecnologia LTE a cada *TTI*. Cada um dos M usuários recebe um canal independente. Os modos de transmissão seguem a modulação 2^j-QAM, com $j = [2, J]$, sendo J o número máximo de bits/símbolo. A mobilidade dos usuários é restrita à própria célula. No cenário multiagente, considera-se a interferência entre sinais OFDM de diferentes células.

A. Estados do Sistema

O estado do sistema é composto pelo estado do *buffer* de cada usuário, a taxa média de chegada de pacotes e o desvio padrão da taxa de chegada de pacotes dos dados recebidos de cada usuário, além do estado do canal. O estado do sistema é utilizado como entrada para a rede DQN, que tem como objetivo fornecer a melhor ação para o estado de entrada. O estado do sistema S é dado por:

$$S = (l_1 \dots l_K, \lambda_1 \dots \lambda_K, \sigma_1 \dots \sigma_K, \bar{g_1} \dots \bar{g_K}) \quad (1)$$

onde l_k é um valor inteiro de tamanho máximo L, medido em pacotes, que representa o estado do *buffer* do usuário k, λ_k é a taxa média de chegada de pacotes, σ_k é o desvio padrão da taxa de chegada de pacotes e $\bar{g_k}$ é o estado do canal.

Neste artigo, o ganho de canal é considerado uma variável aleatória. A amplitude do ganho de canal $|h|$ é assumida como seguindo uma distribuição de Rayleigh devido ao efeito de multipercurso (multipath), e os parâmetros da distribuição são extraídos do modelo de canal TDL-B de [9]. O ganho de potência $g = |h|^2$ segue uma distribuição exponencial. Para obter a perda de percurso (path loss) para o ganho de canal, com base em [9], é assumida uma distância máxima de D metros entre o usuário e a estação-base.

Nas simulações do cenário de comunicação, para considerar os efeitos de múltiplos agentes, são incluídas as ações de outros agentes (ações atrasadas) como parte do estado do agente atual, conforme segue:

$$S_M = (l_1 \dots l_K, \lambda_1 \dots \lambda_K, \sigma_1 \dots \sigma_K, \bar{g_1} \dots \bar{g_K}, a_{d+1} \dots a_K) \quad (2)$$

onde S_M representa o estado multiagente do sistema e $a_{d+1} \dots a_K$ são as ações atrasadas dos outros agentes. Além disso, consideramos nas simulações que, cada agente é designado a sua própria rede, ou seja, este fica responsável por gerenciar os primeiros d usuários. Uma ação consiste em d escolhas categóricas, uma para cada usuário dentro da célula.

B. Ações do Algoritmo de Aprendizagem por Reforço

A ação de cada agente determina o modo de transmissão da comunicação no *downlink*. Uma vez selecionado o modo de transmissão, a alocação de potência necessária para atingir uma Taxa de Erro de Bits (BER) mínima pode ser calculada analiticamente por uma equação que relaciona BER e potência (ver equação (6)). Este método permite um espaço de ações discreto, adequado para um Deep Q-Network (DQN), apesar da natureza contínua da alocação de potência. O número de possíveis ações N_A por agente será determinado pelo número de modos de transmissão para cada usuário e pelo número de usuários, conforme dado por:

$$N_a = n_a^K \quad (3)$$

onde n_a é o número de possíveis ações para cada usuário e K o número de usuários.

Redes DQN requerem um número finito de ações, que são representadas na saída finita da rede. Neste artigo, diferentes conjuntos de ações de usuários, que seão treinados por diferentes redes, são considerados para comparação. Eles são denominados A_1 e A_2, sendo que:

$$A_1 = [2, 6, 10] \cdot cdRate \quad (4)$$

corresponde aos modos 4QAM (2 bits/símbolo), 64QAM (6 bits/símbolo) e 1024QAM (10 bits/símbolo) e, combinados, fornecem 3^K possíveis ações do agente. O cdRate é um parâmetro utilizado para converter os bits/símbolo do modo de transmissão selecionado em pacotes/TTI (intervalo de tempo de transmissão). Mais informações sobre o cdRate são disponibilizadas na Seção III. Em relação ao outro conjunto de ações de usuários, temos que:

$$A_2 = [-1, 1] \quad (5)$$

que corresponde aos modos capazes de transmitir λ - 1 pacotes e λ +1 pacotes e, combinados, fornecem 2^K possíveis ações do agente, sendo λ é a média móvel dos pacotes recebidos para cada usuário.

C. Potência do Canal

A potência do canal $P(c, j)$ depende do modo de transmissão j e da potência de ruído WN_0. A potência mínima necessária para assegurar uma BER máxima (neste trabalho de 10^{-4}) é expressa por:

$$P(c, j) \geq \frac{(2^j - 1)\ln(5p_{BER}(c, j))}{-1.6\rho / WN_0^c} \quad (6)$$

onde $pBER$ é a probabilidade de erro de bit e ρ o ganho de potência do canal.

No cenário de comunicação deste artigo, são consideradas mobilidade dos usuários e perda de percurso associada à distância entre usuários e estações-base, fazendo com que WN_0^c varie individualmente conforme o índice c.

III. Modelo do Sistema de Comunicação

Nesta seção, apresentamos o modelo de canal e os parâmetros considerados na simulação do sistema 5G baseado em CP-OFDM.

O sistema consiste em uma estação-base central em uma célula fixa, atendendo K usuários com M sub-bandas. Cada estação-base, atuando como agente, decide a alocação de recursos a cada *TTI* (1 ms). As ações do agente correspondem aos J modos de transmissão definidos pelas subportadoras OFDM. Cada usuário possui um *buffer* individual de comprimento L.

No LTE, tempo e frequência são divididos em blocos de recursos. Na numerologia 0, as subportadoras têm espaçamento de 15 kHz. Um bloco de recurso possui 180 kHz (12 subportadoras), 14 símbolos e dura 1 ms. A conversão de modo de transmissão (bits/símbolo) para pacotes/TTI é dada por:

$$cdRate = 12 \cdot 14 \cdot 0.9 \cdot \frac{64000}{180 \cdot 8 \cdot 3360} \tag{7}$$

Esta equação converte os 14 símbolos do bloco de recurso de 12 subportadoras de cada TTI em pacotes de 3360 bytes e uma largura de banda de 64 MHz alocada. O fator 0.9 cobre a banda de guarda de 10 porcento.

A. Modelo de Múltiplos Percursos do Canal

Em sistemas de comunicação, o sinal transmitido se propaga em múltiplas direções, sofrendo reflexões que geram cópias com diferentes amplitudes. Para representar esse comportamento, utiliza-se um modelo de canal de multipercurso [9].

A amplitude do sinal (v), em um ambiente com ruído gaussiano branco aditivo (AWGN), segue uma distribuição de Rayleigh, enquanto sua potência, correspondente ao quadrado da amplitude, é descrita por uma distribuição exponencial [10]. O ganho de canal ρ, com ganho médio ρ_m, tem sua distribuição probabilística dada por:

$$p_\rho(\rho) = \frac{1}{\rho_m} \exp\left(\frac{-\rho}{\rho_m}\right) \tag{8}$$

Um modelo bastante adotado para descrever canais com características temporais variáveis é o modelo TDL-B (*Tapped Delay Line*) [9], cuja resposta ao impulso é expressa como:

$$h(t,\tau) = \sum_{i=1}^{N} v_i(t)\delta(\tau - \tau_i) \tag{9}$$

onde $v_i(t)$ representa a amplitude ao longo do tempo, δ é a função impulso, e τ_i são os 23 atrasos do modelo, conforme [9]. Esse modelo é amostrado em uma simulação de Monte Carlo. Posteriormente, uma distribuição exponencial (que modela o efeito de múltiplos percursos no ganho de potência) é amostrada durante o treinamento dos agentes.

Ao relacionar ganhos de sub-banda e usuários, obtém-se uma matriz de canal. Para reduzir o espaço de decisão e agilizar o modelo, adota-se uma estratégia adaptativa com o Algoritmo Húngaro [11], associando canais aos usuários a cada TTI. Isso beneficia o DQN ao reduzir o número de ações e permitir escolhas baseadas em ganhos mais confiáveis.

B. Modelo de Perda de Sinal por Distância

A atenuação do sinal aumenta proporcionalmente à distância entre o transmissor e o receptor. O modelo de perda de percurso média considerado neste artigo segue a equação de Friis [9], dada por:

$$PL_{[dB]} = 32.4 + 20\log(f_c) + 30\log(D) \tag{10}$$

onde f_c (GHz) é a frequência da portadora definida em 6 GHz e D é a distância em metros entre o transmissor e o receptor.

IV. Rede DQN

A arquitetura DQN faz uso neste trabalho, conforme mostrada na Figura 1, de um Perceptron Multicamadas (MLP) com 2 camadas ocultas, de forma semelhante ao que é feito em [12]. O agente age com base nos valores Q gerados pela rede. A entrada $s = (l, \lambda, \sigma, \bar{g}, a)$ possui $5K$ neurônios para múltiplos agentes e $4K$ para um único agente. As camadas ocultas possuem 150 e 300 neurônios, e a saída tem N_a neurônios, um por ação. O total de parâmetros é de $45451 \cdot N_a \cdot 5K$ para múltiplos agentes.

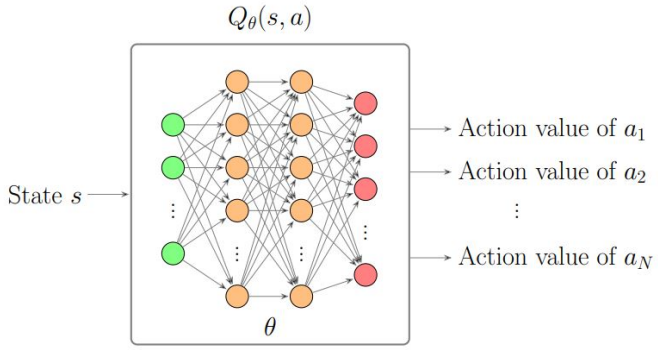

Fig. 1: Rede Multicamadas Perceptron da Deep Q-Network

A. Função de Recompensa

Neste trabalho, tanto as Redes Deep-Q (DQNs) quanto o Método de Entropia Cruzada (CEM) são treinados de forma adaptativa, utilizando a mesma função de recompensa, aplicada a dois conjuntos distintos de ações. Propomos que a função de recompensa seja dada por:

$$x(s,a) = \frac{1}{K}\left(\sum_{k=1}^{K} \frac{f_k(s,a)}{P_k(s,a)}\right) - \frac{1}{K}\left(\sum_{k=1}^{K} \lambda_k \text{Lost}_k + \exp\left[press \cdot (B_k + \text{Lost}_k)\right]\right) \tag{11}$$

$$R(s,a) = sign(x) \cdot log(|x| + 1) \tag{12}$$

onde $f_k(s,a)$ representa a vazão do usuário k, $P_k(s,a)$ a potência alocada ao dispositivo do usuário necessária para atingir a BER desejada, $Lost_k$ e B_k o número de pacotes perdidos e o estado do *buffer* após a ação a, respectivamente. A taxa média de chegada de pacotes é dada por λ_k, e a pressão do *buffer*, denotada por $press$, assume o valor de 0,5. Sign(x) é a função sinal de x.

Essa formulação proposta considera as ações de todos os agentes. Ou seja, ainda que cada agente decida apenas pelas ações de seus próprios usuários, todos compartilham o mesmo objetivo global, sendo, portanto, influenciados pelas boas ou más decisões dos demais.

V. Rede CEQN Adaptativa

O método de Entropia Cruzada (CEM) busca otimizar os valores dos pesos de uma rede neural sem retropropagação [1], [2]. Assim, avaliamos múltiplos MLPs que mapeiam $Q(s,a)$, aplicando a melhor ação e classificando com base em recompensas acumuladas em 10 passos. A população é inicializada com pesos aleatórios e usamos uma média móvel em uma janela de 10 passos.

Os passos do CEM são:

1) Avaliar a aptidão da população;
2) Selecionar o mais apto;
3) Repopular com cópias dos mais aptos;

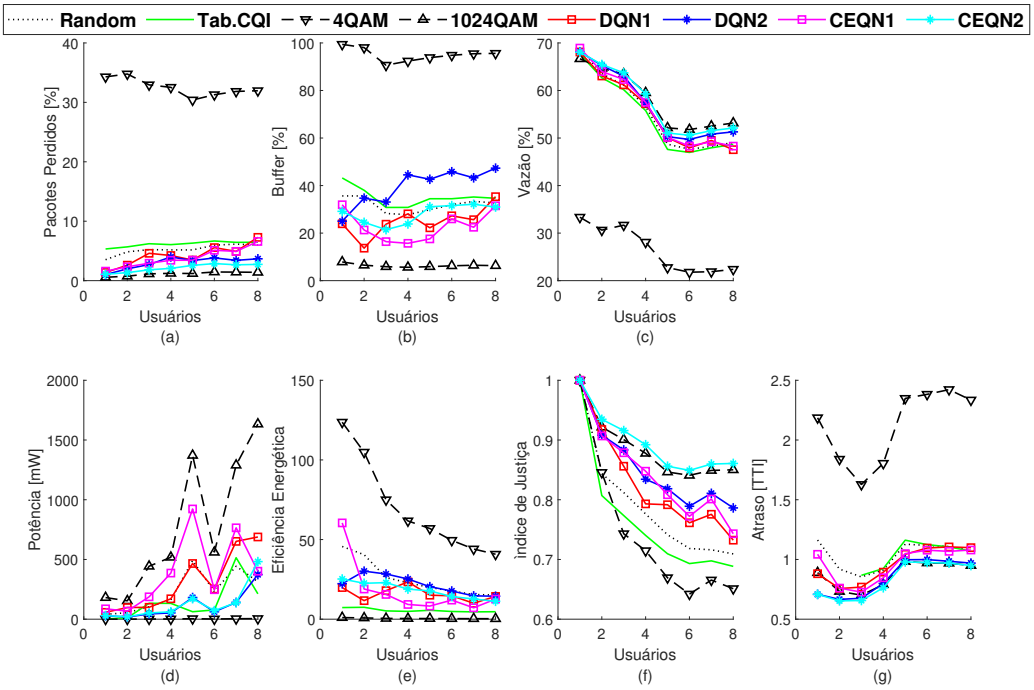

Fig. 2: QoS de um único agente com número variável de usuários

4) Introduzir mutações gaussianas nos pesos.

A aptidão é calculada por:

$$Q(s_0, a_0) = \sum_{i=0}^{9} \gamma^i R(s_i, a_i) \tag{13}$$

com $\gamma = 0.9$ e R da Eq.(12).

Os 40 melhores indivíduos definem as estatísticas dos pesos. A nova população é formada por esses indivíduos e amostras gaussianas, até atingir 100 redes.

O CEM se destaca quando a função de otimização possui gradiente ruidoso e muitos mínimos locais [13], o que o torna adequado para os desafios do aprendizado por reforço.

Em resumo, a Rede CEQN adaptativa proposta para alocação de recursos em redes sem fio 5G e pós-5G consiste na utilização da função de recompensa Eq. 11 em uma rede DQN sendo que o método de entropia cruzada (CEM) descrito nesta seção é aplicado para melhorar o desempenho de treinamento desta rede DQN.

VI. Resultados

Nesta seção, compararamos os desempenhos de diferentes abordagens: Aleatório (modo de transmissão aleatório), Tab. CQI (limiares fixos de ganho de canal), 4QAM e 1024QAM (modos fixos), DQN e CEQN. As abordagens DQN1 e CEQN1 usam o conjunto de ações da Eq. 4, DQN2 e CEQN2 usam o da Eq. 5.

A. Cenário de Agente Único

Adota-se nas simulações $K = [1, 2, 3, 4, 5, 6, 7, 8]$ usuários, $L = 30$ pacotes, $M = K$ canais, máximo de bits $J = 10$ (1024-QAM), TTI de $1ms$, largura de banda de 64 MHz/usuário, pacotes de 3360 KB, frequência de 6 GHz e alcance de 50 metros.

Na Figura 2, pode-se observar o resumo dos resultados para diferentes abordagens. O conjunto de ações 2 (CEQN2 e DQN2) se destaca com melhores valores para perda de pacotes (Fig. 2–a), vazão (Fig. 2–c), consumo de potência (Fig. 2–d), eficiência energética (Fig. 2–e), índice de justiça (Fig. 2–f) e atraso médio (Fig. 2–g) em relação aos algoritmos avaliados. Já o conjunto 1 de ações (DQN1 e CEQN1) provê melhor ocupação do *buffer* em relação aos algoritmos considerados. Ao comparar CQN2 e DQN2 observa-se uma vantagem do CQN2 na maioria dos resultados e expressiva vantagem no indice de justiça (Fig. 2–f).

B. Cenário de Múltiplos Agentes com Seleção de Usuários

Conforme mencionado anteriormente, propõe-se neste artigo também uma abordagem com múltiplos agentes. Esta abordagem, assim como a com agente único, é combinada a um algoritmo de seleção de usuários, onde apenas $M < K$ usuários são atendidos a cada TTI, com prioridade definida por:

$$Prioridade = \log_2(PcktIN + Buff) \cdot \frac{\rho}{\overline{\rho}} \tag{14}$$

Note que todas abordagens, exceto DQN e CEQN, selecionam usuários aleatoriamente. As simulações com três células, oito usuários e até oito subcanais por célula mostraram que a perda de pacotes diminui conforme mais usuários são atendidos, com maior vantagem ao selecionar 4 de 8 usuários. A Figura 3 evidencia que a solução CEQN2 apresenta os menores valores de perda e maiores valores de vazão (Fig. 3–a, c), superando até mesmo a abordagem 1024QAM com seleção aleatória quando até 5 dos 8 usuários da célula são antendidos por intervalo.

VII. Conclusão

Neste trabalho, abordamos a limitação da escalabilidade de redes DQN, causada pelo crescimento do espaço de ações, usando um conjunto reduzido de ações discretas, mas alcançando ampla cobertura do espaço de alocação de potência. A alocação de potência baseada em restrições de

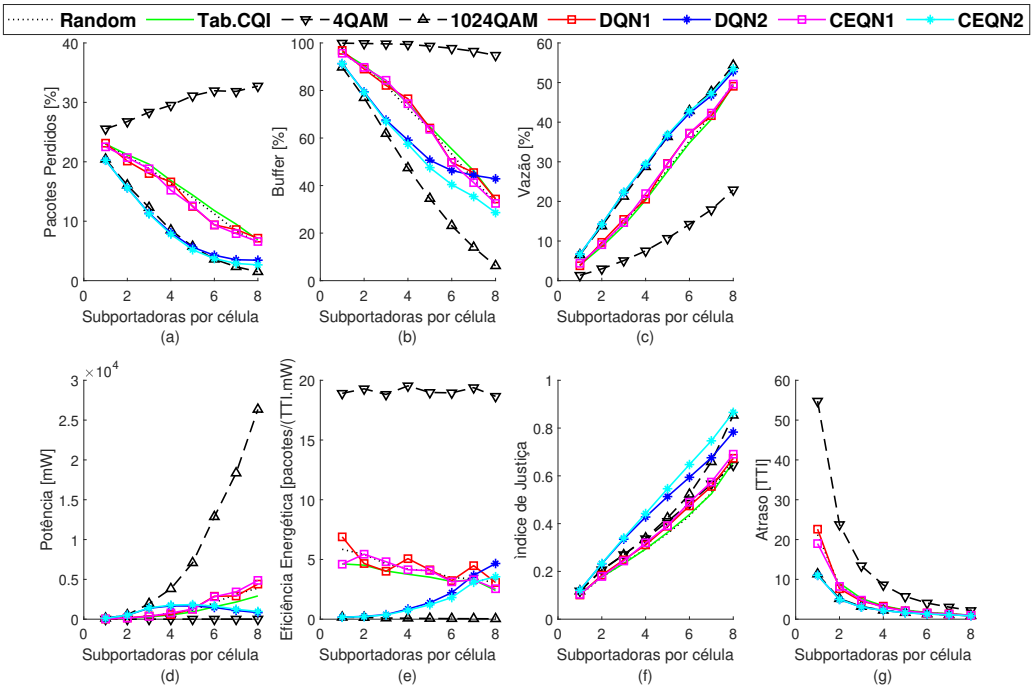

Fig. 3: QoS utilizando múltiplos agentes e variando a quantidade de usuários selecionados por célula

BER aplicada em conjunto com o algoritmo CEQN proposto se mostrou eficaz na alocação de recursos.

O método CEM foi essencial para atingir os objetivos da função de recompensa no algoritmo CEQN, favorecendo a escalabilidade da rede por exigir apenas o processo de *forwarding*. A abordagem adaptativa enfrenta desafios de tempo de processamento, tornando o uso da entropia cruzada uma alternativa eficiente. Observou-se que as estratégias baseadas em aprendizado por reforço e CEM proporcionaram à rede sem fio baixa perda de pacotes, baixa ocupação de *buffer*, alta taxa de transferência e eficiência energética em especial quando mais usuários são atendidos por célula.

A abordagem CEQN2 apresentou os melhores resultados na maioria dos parâmetros, destacando-se pelo equilíbrio entre perda de pacotes e eficiência energética. A utilização do segundo tipo de ação (Eq. (5)) influenciou positivamente o desempenho na alocação de recursos, auxiliando o agente na adaptação a fluxos de dados heterogêneos.

Além disso, observou-se que a aplicação de métodos de seleção de usuários em cenários com recursos limitados pode melhorar o desempenho dos algoritmos, sendo o CEQN2 a solução recomendada.

AGRADECIMENTOS

Os autores agradecem ao Centro de Excelência em Redes Inteligentes Sem Fio e Serviços Avançados (CERISE) e à Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) pelo apoio e financiamento à pesquisa.

REFERÊNCIAS

[1] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018.

[2] Kuan-Chou Lee, Sen-Hung Wang, Chih-Peng Li, Ho-Hsuan Chang, and Hsueh-Jyh Li. Adaptive resource allocation algorithm based on cross-entropy method for ofdma systems. *IEEE Transactions on Broadcasting*, 60(3):524–531, 2014.

[3] J. Zhu, Y. Song, D. Jiang, and H. Song. A new deep-q-learning-based transmission scheduling mechanism for the cognitive internet of things. *IEEE Internet of Things Journal*, 5(4):2375–2385, 2018.

[4] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67, 2005.

[5] Muhammad Anjum Qureshi and Cem Tekin. Fast learning for dynamic resource allocation in ai-enabled radio networks. *IEEE Transactions on Cognitive Communications and Networking*, 6(1):95–110, 2019.

[6] Patteti Krishna, Tipparti Anil Kumar, and Kalitkar Kishan Rao. M-qam ber and ser analysis of multipath fading channels in long term evolutions (lte). *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9(1):361–368, 2016.

[7] Daniel PQ Carneiro, Alisson A Cardoso, and Flávio HT Vieira. Adaptive resource allocation in 5g cp-ofdm systems using markovian model-based reinforcement learning algorithm. *Neural Computing and Applications*, 35(13):9421–9435, 2023.

[8] MAWI Working Group et al. Traffic archive. *mawi . wide . ad . jp/mawi/*.

[9] 3GPP Radio Access Network Working Group et al. Study on channel model for frequencies from 0.5 to 100 ghz (release 15). Technical report, 3GPP TR 38.901, 2018.

[10] Bernard Sklar. *Digital communications: fundamentals and applications*. Pearson, 2021.

[11] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52(1):7–21, 2005.

[12] Heunchul Lee, Maksym Girnyk, and Jaeseong Jeong. Deep reinforcement learning approach to mimo precoding problem: Optimality and robustness. *arXiv preprint arXiv:2006.16646*, 2020.

[13] Rodolfo E. Haber, Gerardo Beruvides, Ramón Quiza, and Alejandro Hernandez. A simple multi-objective optimization based on the cross-entropy method. *IEEE Access*, 5:22272–22281, 2017.