Alocação de recursos em redes sem fio utilizando algoritmos baseados em Aprendizado de Máquina

Jean Lucas B. Silva[1], **Daniel Porto Q. Carneiro**[2], **Flávio Henrique T. Vieira** [3]

[1]Escola de Engenharia Elétrica, Mecânica e de Computação (UFG)
74605-010 - Goiânia, Goiás - Brasil

[2]Programa de Pós-Graduação em Engenharia Elétrica
e de Computação (PPGEEC-UFG)
Goiânia, Goiás - Brasil

[3]Escola de Engenharia Elétrica, Mecânica e de Computação (UFG)
74605-010 - Goiânia, Goiás - Brasil

jean.silva@discente.ufg.br, rynu.tjf@gmail.com, flavio_vieira@ufg.br

Abstract. *The present work introduces a study on resource allocation in a multi-user communication system based on CP-OFDM using reinforcement learning methods, specifically the DQN network associated with the Cross-Entropy method. The aim is to develop a resource scheduling algorithm that maximizes data transmission efficiency in wireless network environments. The analysis encompasses various parameters, including bandwidth, average packet size, modulation, and number of users. It details resource elements, subcarrier distribution, and transmission capacity in different modulation modes within the LTE system frame structure. It is concluded that the proposed approach, named Reinforcement Learning with Cross-Entropy, generally enhances QoS parameters such as average throughput, packet loss, buffer queue occupancy, and the energy efficiency of the communication system compared to the traditional approach based on Deep Q-Network.*

Resumo. *O presente trabalho apresenta um estudo sobre alocação de recursos em um sistema de comunicação multiusuários baseado em CP-OFDM utilizando métodos de aprendizado por reforço, especificamente a rede DQN associada ao método de Entropia Cruzada. O objetivo é desenvolver um algoritmo de escalonamento de recursos que maximize a eficiência da transmissão de dados em ambientes de rede sem fio. A análise abrange vários parâmetros, incluindo largura de banda, tamanho médio do pacote, modulação e número de usuários. Detalha elementos de recursos, a distribuição de subportadoras e a capacidade de transmissão nos diferentes modos de modulação presentes na estrutura do frame do sistema LTE. Conclui-se que a abordagem proposta, denominada Aprendizado por Reforço com Entropia Cruzada, em geral, provê melhoras nos parâmetros de QoS, como vazão média, perda de pacotes, ocupação na fila do buffer e eficiência energética do sistema de comunicação, em comparação com a abordagem tradicional baseada em Deep Q-Network.*

1. Apresentação

A expansão das redes de quinta geração de comunicação móvel (5G) em operação e a expectativa da chegada da tecnologia 6G até 2030 [Henrique and Prasad 2021], tem despertado um interesse crescente em sistemas de comunicação ultra rápidos e altamente confiáveis (URLLC) [Popovski et al. 2019]. Graças aos benefícios práticos do URLLC, que se concentra na transmissão com blocos de dados de duração limitada, tornou-se possível atender às necessidades de aplicações em crescimento, tais como sistemas de transporte inteligente [Xiang et al. 2020], automação residencial e industrial (Indústria 4.0) e saúde (cirurgia remota). Somado a isso, a dinâmica em constante evolução do tráfego real, combinada com a complexidade dos modelos de canal que incorporam múltiplos caminhos de propagação e a utilização de ondas milimétricas, requer abordagens avançadas e adaptativas para o gerenciamento de recursos.

Recentemente, houve um aumento significativo no uso de técnicas de Aprendizado de Máquina [Mitchell 1997] para lidar com desafios complexos e tomar decisões mais autônomas, abordando eficazmente uma série de desafios em comunicações e redes sem fio. O Aprendizado por Reforço (*RL*), uma das ferramentas da Aprendizagem de Máquina, demonstrou ser capaz de capacitar os dispositivos *IoT* a tomar decisões autônomas para a alocação de recursos em sistemas de comunicação.

Portanto, este trabalho objetiva-se apresentar os resultados e validar um esquema adaptativo de alocação de recursos, voltado para sistemas de comunicação multiusuários e multiportadora. Nas seções seguintes estão descritas uma revisão bibliográfica, o método proposto, os resultados e discussões e, por fim, a conclusão do trabalho.

2. Revisão Bibliográfica

Esta seção apresenta os principais trabalhos relacionadas ao uso de técnicas de Aprendizado de Máquina para melhorar a transmissão multiusuários e eficiência energética em redes sem fio.

Em [Liang et al. 2019], fornece uma visão abrangente do potencial do aprendizado profundo na resolução de problemas de alocação de recursos em redes sem fio, com foco particular em redes veiculares. Os autores utilizam de uma arquitetura de treinamento centralizada multiagente para permitir que veículos aprendam como alocar recursos (como largura de banda e potência) de maneira distribuída e coordenada. O método proposto incentiva a cooperação entre veículos sem qualquer coordenação online, e mostrou-se significativamente mais eficaz em termos de desempenho de rede do que uma abordagem baseada em um único agente.

Em [Zhang M 2022], é proposto uma abordagem DQN para otimizar a alocação de recursos em uma rede de acesso múltipla não ortogonal (NOMA). A abordagem utilizada de técnicas de aprendizado por reforço junto com o método de Entropia Cruzada para maximizar a taxa de sigilo em uma rede NOMA, com o objetivo de alcançar baixas taxas de erro. A abordagem proposta superou outros métodos em termos de desempenho e eficiência computacional.

Em [Carneiro 2022], foi proposto um algoritmo de alocação e recursos em redes sem fio multiportadoras com ondas milimétricas utilizando aprendizado por reforço baseado em conjunto com uma rede neural DQN. O algoritmo proposto visa maximizar a

eficiência energética e a taxa de transmissão dos usuários, levando em consideração as restrições do sistema, como largura de banda disponível, a potência máxima de transmissão e a qualidade do canal, entre outros. Os resultados mostram um impacto positivo ao se escolher uma determinada função de recompensa e algoritmo.

Semelhante a [Carneiro 2022], é utilizado a abordagem de alocação de recursos em redes sem fio multiportadora e multiusuários utilizando aprendizado por reforço. Para fins de comparação, é aplicado o método de Entropia Cruzada para avaliar o desempenho do agente.

3. Proposta de alocação de recursos utilizando RL

A tecnologia LTE-OFDM emprega um espaçamento de subportadoras de 15 kHz com 12 subportadoras por bloco de recurso. Baseada em um esquema de modulação OFDM, permite a transmissão simultânea de dados em diferentes frequências. Devido à natureza ortogonal das frequências, pode ocorrer sobreposição entre as faixas de frequências da subportadora. Assim, resulta na utilização mais eficiente do espectro, uma vez que a largura de faixa é menor em comparação com o FDM (*Frequency Division Multiplexing*). Ademais, para uma mesma largura de faixa disponível, a comunicação CP-OFDM acomoda mais sub-portadoras, levando vantagem em quantidade de símbolos. Além disso, permite altas taxas de transmissão com símbolos de duração mais longa no domínio do tempo, reduzindo a interferência dos efeitos de múltiplos percursos do sinal.

A Figura 1 ilustra a configuração de um *frame* no sistema de comunicação LTE CP-OFDM, revelando o tamanho do bloco de recursos para um *slot* do *subframe*. No modo normal há 7 símbolos por bloco de recurso. Consequentemente, um bloco de 12 subportadoras terá 12 X 7 = 84 elementos de recurso. Cada elemento de recurso representa uma combinação de frequência e símbolo, variando de acordo com o modo de transmissão. A tecnologia LTE opera com os modos QPSK/4QAM, 16QAM, 64QAM e 256QAM que correspondem a transmissão de 2, 4, 6 e 8 bits por símbolo (elemento de recurso).

A modelagem do sistema de comunicação CP-OFDM realizada nas simulações deste trabalho utiliza dados reais de tráfego de [MAWI 2019].Isso ocorre, porque, na prática, não se tem controle sobre os valores quando se trata de dados sintéticos nas simulações. Nesta alocação adaptativa de recursos avalia-se os intervalos de λ que fazem com que o sistema de comunicação apresenta melhor desempenho em termos de pacotes perdidos, ocupação média do *buffer*, vazão, potência média consumida, eficiência energética e índice de justiça. Dado que os valores da taxa média de chegada apresentam flutuações, é adotado limiar de 0.15λ para determinar quando é necessário implementar uma política de ação. Neste trabalho, a frequência de portadora foi mantida fixa em 6 GHz, além de fixar o tamanho dos pacotes em 3360 bytes e um intervalo de TTI (*Time Transmission Interval*) igual a 0,5 ms. A taxa de codificação para o downlink do sistema LTE é dada pela seguinte equação:

$$v = \frac{12 \cdot 7 \cdot 0.9 \cdot BW}{15000 \cdot M \cdot 8 \cdot Pacote} \tag{1}$$

onde BW é a largura de banda em Hz (descontada a banda de guarda), M é o número de canais (sub-bandas) e Pacote é o tamanho do pacote em bytes. Essa consideração implica

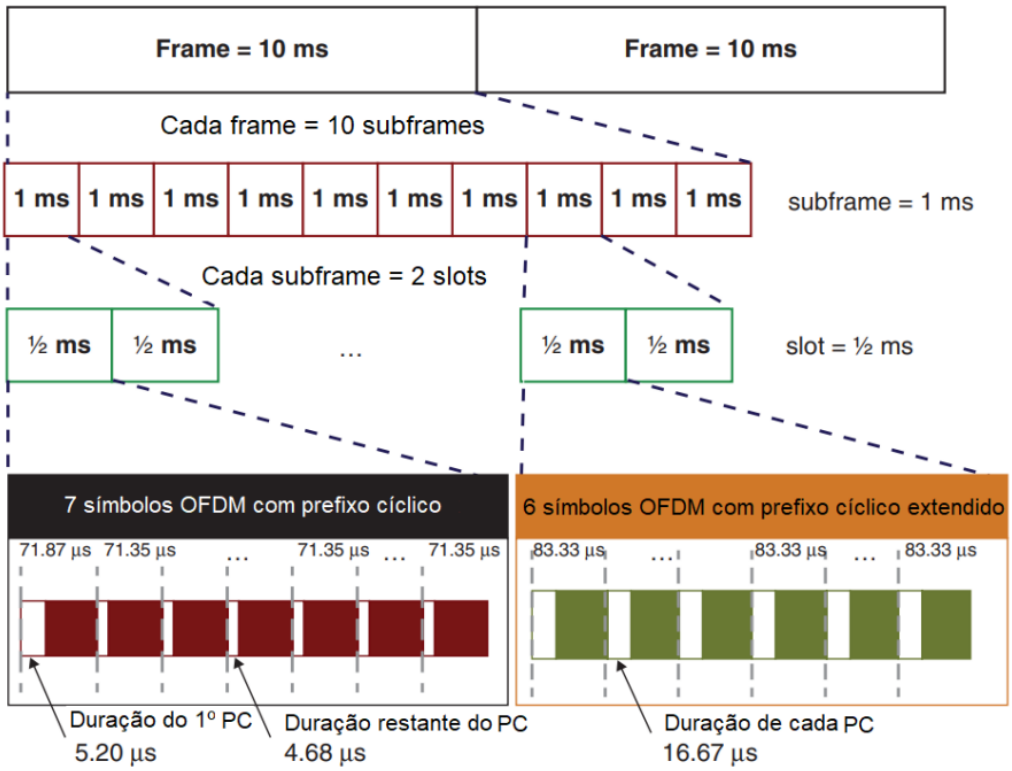

Figura 1. Estrutura de um *frame* para o sistema de comunicação LTE-OFDM. Fonte: Zarrinkoub (2014)

que a comunicação transcorre de maneira contínua, ou seja, a sequência de pacotes é transmitida de maneira gradual e é perceptível em nível de bits. Essa abordagem também viabiliza a aplicação das equações de BER (*Bit Error Rate*) ao problema em questão. A adoção do modelo contínuo minimiza o número de eventos simulados, resultando em maior eficiência no processo de simulação [Carneiro 2022].

O aprendizado por reforço é uma técnica de aprendizado de máquina que permite que um agente aprenda a tomar decisões em um ambiente incerto, através da interação com esse ambiente [Sutton and BARTO 2018]. O objetivo do agente é maximizar uma recompensa numérica, que é recebida após cada ação tomada. Além disso, o modelo é utilizado para estimar recompensa esperada de cada ação em cada estado, o que permite que o agente tome decisões mais informadas. A função $Q(s, a)$ é uma extensão do algoritmo de aprendizado temporal proposto por Bellman e é expressa da seguinte forma:

$$Q^{\pi}(s, a) = r + \gamma max Q^{\pi}(s', a') \tag{2}$$

onde *r* é a recompensa imediata obtida e γ é o fator de desconto para recompensas futuras, *s, s', a e a'* os estados e ações presentes e futuros respectivamente.

Para o DQN, não é necessário especificar o tamanho do espaço de estados, apenas a sua forma. Todavia, é necessário listar o espaço de ações (discreta), uma vez que cada ação é representada na camada de saída da rede neural. Para essa aplicação é escolhido um vetor coluna com os estados do sistema, sendo considerado os estados do *buffer* de cada usuário e do canal, seguido pela quantidade de pacotes demandado para cada usuário, as taxas médias de chegada λ.

Na estrutura do DQN, uma rede classificadora com múltiplas saídas é empregada juntamente com o algoritmo *Q-Learning*, o que implica a associação do cálculo de *Q* com a determinação da melhor ação através da rede neural. Com esse enfoque, a rede aproxima a função *Q* e oferece a ação ideal. A Equação de Bellman pode ser representada em sua forma alternativa, onde a função $Q(s, a)$ é atualizada com base na taxa de aprendizagem α:

$$Q(s_i, a_i) \leftarrow (1 - \alpha) \cdot Q(s_i, a_i) + \alpha \cdot (R_i + \gamma max a' Q(s', a')) \tag{3}$$

Na equação 3, a cada novo passo gera-se uma tupla (s_i, a_i, R_i, s).

O método de Entropia Cruzada, descrito em [Boer et al. 2005] é um Algoritmo Evolutivo aplicado no treinamento de agentes *RL*, ajudando a ajustar as ações do agente para maximizar as recompensar esperadas, destacando as ações ruins tomadas e treinar o agente com a melhores ações. O princípio fundamental do método envolve a seleção de várias entradas do algoritmo, a avaliação de suas saídas e, subsequentemente, o foco nas entradas que produzem resultados mais favoráveis. Esta abordagem iterativa visa ajustar o agente até que o desempenho desejado seja alcançado. Neste algoritmo, uma distribuição gaussiana $N(\mu, \sigma)$ é empregada para representar os pesos da rede neural θ. Outro aspecto importante envolve a geração de conjuntos de θ, amostrados a partir da distribuição gaussiana com o tamanho *N* especificado. Esses conjuntos são então avaliados usando a função de custo da rede neural. Essa função de custo é uma medida do erro entre a saída da rede neural e o valor esperado da recompensa de longo prazo. Posteriormente, são escolhidas as amostras θ mais promissoras, levando ao cálculo de novos parâmetros μ e σ para a distribuição gaussiana. Esta sequência de passos é repetida iterativamente até que a convergência seja alcançada. A política inicial é estabelecida de forma aleatória e melhorada de forma incremental ajustando os parâmetros da rede neural.

Este estudo, baseado no trabalho de [Carneiro 2022] que adotou uma abordagem de alocação de recursos por meio de uma rede DQN, examina a aplicação do método de Entropia Cruzada para melhorar a rede DQN. É explorada uma rede DQN para resolver a tarefa de alocar recursos em contexto SISO (*Single Input Single Output*) ODFM, que englobam múltiplos usuários, sujeitos a uma restrição de BER máxima. Além de ser utilizado uma rede DQN *off-policy*, ou seja, a função *Q-Learning* aprende com ações que estão fora da política atual. Avaliamos as funções de recompensa de [Carneiro 2022], juntamente com a proposta por [Zhu 2018]. A função de recompensa de [Zhu 2018] considera o tamanho do *buffer* pacotes transmitidos e a potência alocada, enquanto que [Carneiro 2022] propõe quatro funções que combinam o tamanho do *buffer* com pacotes perdidos. Adicionalmente, utiliza-se o algoritmo de "Ação Fixa"com limite superior para potência e fluxo de pacotes, adotando 256QAM para todos os canais de comunicação e selecionando usuários de forma aleatória. Ademais, também foi adotada nas simulações abordagens de seleção aleatória de uma ação a partir do conjunto de opções possíveis.

Abaixo são apresentas as funções de recompensas utilizadas para avaliar o desempenho dos algoritmos de aprendizado por reforço:

$$R_{Zhu_{QL}}(s, a) = \frac{EE(s, a)}{\sum_{k=1}^{K} e^{0.5.l_k}} \tag{4}$$

$$R_{Prop1}(s,a) = \frac{EE(s,a)}{\sum_{k=1}^{K} E[e^{0.5(B_k+Lost_k)}]} \tag{5}$$

$$R_{Prop2}(s,a) = \frac{EE(s,a)}{\sum_{k=1}^{K} E[e^{0.5Lost_k} + B_k]} \tag{6}$$

$$R_{Prop3}(s,a) = \frac{EE(s,a)}{\sum_{k=1}^{K} E[\frac{e^{0.5(l_k+Lost_k)}}{B_k}]} \tag{7}$$

$$R_{Prop4}(s,a) = \frac{EE(s,a)}{\sum_{k=1}^{K} E[Lost_k x \lambda_k + e^{0.5(B_k+Lost_k)}]} \tag{8}$$

onde *EE(s,a)* representa a eficiência energética, $P_k(s,a)$ reflete a porção da potência direcionada para assegurar a taxa de erro de bits mínima (BER) ao usuário *k*. Enquanto isso, $Lost_k$ e B_k indicam a quantidade de pacotes perdidos e o estado atual do *buffer* do usuário *k*, após a tomada da ação *a* no estado *s*, respectivamente. Adicionalmente, λ_k designa a taxa média de chegada de pacotes para o usuário *k*.

4. Resultados e discussão

Conforme discutido nas seções anteriores, foi realizada uma análise comparativa para o desempenho da DQN sem e com a aplicação do algoritmo de Entropia Cruzada. Nas simulações conduzidas nesse artigo, a seguinte configuração do sistema de comunicação foi adotada para ambas análises: foram utilizados um total de *K*=5 usuários, um tamanho de *buffer* de *L*=5, número de canais *M*=5, número de modos de transmissão *J*=4 e taxa de codificação *V*=2. Dois *clusters* foram formados, um com 3 usuários e outro com 2, totalizando 2 *clusters*. Ao todo são simulados 373700 TTI's que representam 3 minutos.

Quanto à estrutura da rede neural adotada, o número de neurônios na camada de entrada e de saída são respectivamente 9 e 125 para o *cluster* de 3 usuários. Para o *cluster* de 2 usuários, o número de neurônios foi de 6 na camada de entrada e 25 na camada de saída. As três camadas ocultas foram configuradas com um número de neurônios igual à média do número de neurônios na camada de entrada e na camada de saída, resultando em 67 neurônios para cada camada oculta no *cluster* de 3 usuários e 16 neurônios o *cluster* de 2 usuários. Foi utilizada a função de ativação ReLU *(Rectified Linear Unit)* que é uma função que introduz não linearidade ao permitir que valores positivos fluam sem alterações, enquanto zera os valores negativos.

Na Figura 2, é mostrado os resultados obtidos para o sistema em um cenário adaptativo com a aplicação de uma rede DQN sem o método da Entropia Cruzada, avaliando as funções de recompensa descritas em [Carneiro 2022] e [Zhu 2018]. Os resultados evidenciam que, em linhas gerais, as abordagens fundamentadas na aplicação da DQN para alocar recursos demonstram um desempenho notável nos parâmetros de Qualidade de Serviço (QoS), contudo, também revelam flutuações nos resultados.

Podemos observar que a Proposta 4 (Equação 8), converge para soluções que minimizem as perdas de pacotes (a), redução da ocupação do *buffer* (b) e se estabiliza com um comportamento de alta vazão (c). Por outro lado, a proposta por [Zhu 2018]

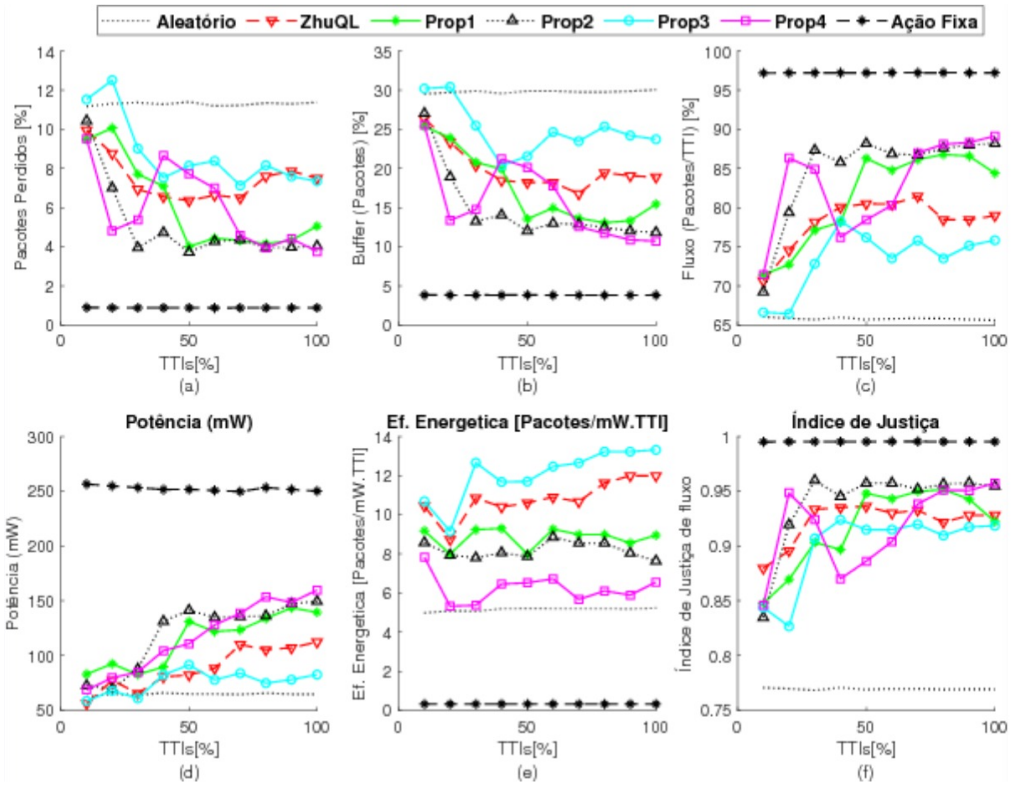

Figura 2. Parâmetros de QoS versus tempo de simulação para rede DQN simulação assíncrona: (a) Pacotes Perdidos, (b) ocupação do *buffer*, (c) Fluxo de pacotes, (d) Potência, (e) Eficiência energética e (f) Índice de justiça

(Equação 4) gasta mais potência e perde menos pacotes. Além disso, a estratégia proposta por [Zhu 2018] e a Proposta 3 (Equação 7), convergem para soluções que otimizam a eficiência energética, sendo a Proposta 3 particularmente destacada (e).

Nesta etapa, é avaliado o sistema com a aplicação de uma rede DQN com o método da Entropia Cruzada. Dentre os parâmetros de simulação, foram utilizados o mesmo número de usuários, tamanho do *buffer*, número de canais, modos de transmissão e taxa de codificação aplicados na DQN sem a utilização do método de Entropia Cruzada. Além disso, a escolha da população e da elite também é muito importante para o desempenho da DQN adaptativa. Uma população muito pequena pode levar a uma convergência prematura, enquanto uma população muito grande pode aumentar o tempo de execução. A elite, por sua vez, é responsável por selecionar os indivíduos mais bem-sucedidos para a próxima iteração. É utilizada uma população de 90 indivíduos e uma elite de 30 indivíduos foram considerados adequados para o problema em questão.

A Figura 3 apresenta os resultados obtidos com o método da Entropia Cruzada. Nota-se uma maior discrepância entre as diferentes abordagens, além de apresentar uma convergência mais rápida e com menos oscilação. Também fica evidente uma maior influência da função de recompensa. Percebe-se que a Proposta 4 (Equação 6) se sobressai em relação às demais propostas quanto a perda pacotes (a), diminuição da ocupação do *buffer* (b), alta vazão (c) e índice de justiça (f). Adicionalmente, o método de entropia cruzada realça os traços característicos da proposta de [Zhu 2018] no que tange à eficiência

energética (e). Tal destaque se origina do fato de que um tamanho de *buffer* de usuário menor tende a amplificar essa particularidade.

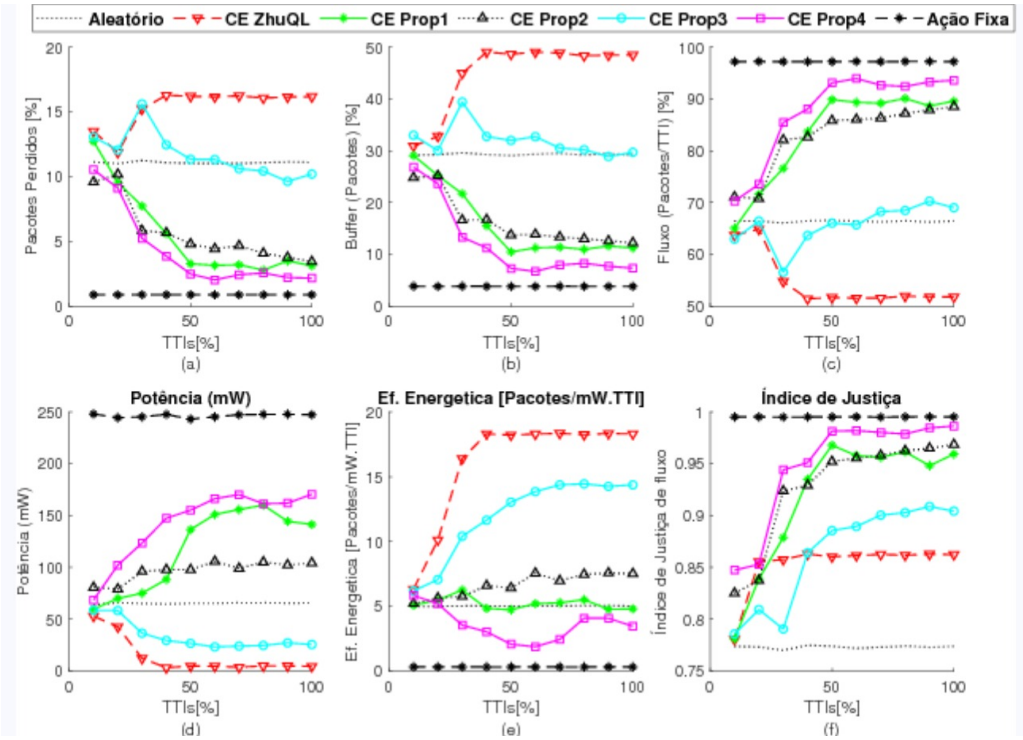

Figura 3. Parâmetros de QoS versus tempo de simulação para rede DQN com o método de Entropia Cruzada e simulação assíncrona: (a) Pacotes Perdidos, (b) ocupação do *buffer*, (c) Fluxo de pacotes, (d) Potência, (e) Eficiência energética e (f) Índice de justiça

5. Conclusão

Ao longo deste artigo, foi apresentado um esquema adaptativo de alocação de recursos em redes sem fio, utilizando algoritmos de aprendizado por reforço baseados em *Deep Q-Networks* e o método de Entropia Cruzada. O objetivo principal desse método é otimizar o desempenho de redes *IoT* de maneira adaptável e em tempo real, levando em consideração as restrições energéticas dos dispositivos alimentados por baterias e a demanda crescente por aplicações intensivas.

Os resultados oriundos da sistema treinado por redes DQN com utilização do algoritmo de Entropia Cruzada mostraram que esta abordagem proposta é capaz de melhorar significativamente a eficiência da alocação de recursos em redes sem fio de acordo com as necessidades dos usuários e as condições do ambiente, em comparação com outros métodos tradicionais. Ademais, o esquema apresentou uma capacidade de adaptação em tempo real às mudanças na demanda de tráfego e nas condições do canal de comunicação, o que é essencial para garantir a sustentabilidade e a viabilidade a longo prazo dessas redes. Com os resultados alcançados, foi possível concluir que a utilização do método de entropia cruzada apresentou melhoria nos parâmetros de qualidade do sistema, comparado ao método tradicional da DQN.

Referências

Boer, P.-T., Kroese, D., Mannor, S., and Rubinstein, R. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67.

Carneiro, D. P. Q. (2022). Alocação de recursos em redes sem fio multiportadoras com ondas milimétricas utilizando aprendizado por reforço baseado em modelo markoviano. Dissertação (Engenharia Elétrica e da Computação) - Universidade Federal de Goiás, Goiânia.

Henrique, P. S. R. and Prasad, R. (2021). *6G The Road to the Future Wireless Technologies 2030*, pages i–xxvi.

Liang, L., Ye, H., Yu, G., and Li, G. (2019). Deep learning based wireless resource allocation with application to vehicular networks.

MAWI (2019). Deep reinforcement learning approach to mimo precoding problem: Optimality and robustness. Mawi working group traffic archive.

Mitchell, T. M. (1997). *Machine learning*, volume 1. McGraw-hill New York.

Popovski, P., Stefanović, , Nielsen, J. J., de Carvalho, E., Angjelichinoski, M., Trillingsgaard, K. F., and Bana, A.-S. (2019). Wireless access in ultra-reliable low-latency communication (urllc). *IEEE Transactions on Communications*, 67:5783–5801.

Sutton, R. S. and BARTO, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Xiang, Z., Yang, W., Cai, Y., Ding, Z., Song, Y., and Zou, Y. (2020). Noma-assisted secure short-packet communications in iot. *IEEE Wireless Communications*, 27:8–15.

Zhang M, Zhang Y, C. Q. W. S. (2022). Deep learning–based resource allocation for secure transmission in a non-orthogonal multiple access network. *International Journal of Distributed Sensor Networks*.

Zhu, J. e. a. (2018). A new deep-q-learning-based transmission scheduling mechanism for the cognitive internet of thing. volume 5, pages 2375–2385. IEEE Internet of Things Journal.