Alocação de Potência em Redes Sem Fio Baseadas em Multiplexação por Divisão de Frequências Ortogonais Utilizando Aprendizagem por Reforço

Hudson H. S. Lopes[1], **Anderson da S. Soares**[2] **e Flávio H. T. Vieira**[1]

[1]Escola de Engenharia Elétrica, Mecânica e de Computação (EMC),

[2]Instituto de Informática (INF),
Universidade Federal de Goiás (UFG)

hudson_lopes@ufg.br, anderson@inf.ufg.br, flavio_vieira@ufg.br

Abstract. *In this paper, we approach the challenging problem of allocation of signal transmission power based on the Orthogonal Frequency Division Multiplexing (OFDM) technique. We propose to use Reinforcement Learning (RL) based algorithms to find the optimal policy for allocating power to wireless network devices using a reward function. More specifically, we propose to use the Double Deep Q-Network (DDQN) agent due to its higher learning capacity compared to Q-Learning and Deep Q-Network (DQN). Simulation results show that DDQN agent present promising solutions in power allocation in wireless networks.*

Resumo. *Neste artigo, abordamos o desafiante problema de alocação de potência de transmissão do sinal baseadas na técnica de Multiplexação por Divisão de Frequências Ortogonais (OFDM - Orthogonal Frequency Division Multiplexing). Propomos utilizar algoritmos baseados em Aprendizagem por Reforço (RL - Reinforcement Learning) para encontrar a política ótima para alocação de potência aos dispositivos da rede sem fio usando uma função de recompensa. Mais especificamente, propomos utilizar o agente Rede Q Profunda Dupla (DDQN - Double DQN) devido a sua maior capacidade de aprendizagem em comparação a Aprendizagem Q (Q-Learning) e a Rede Q Profunda (DQN - Deep Q-Network). Os resultados das simulações mostram que o agente DDQN apresenta soluções promissoras na alocação de potência em redes sem fio.*

1. Introdução

Em sistemas de comunicações móveis sem fio, a alocação de recurso de rádio é uma abordagem essencial para melhorar a eficiência na transmissão de dados. A Estação Base (BS - *Base Station*) pode aperfeiçoar a alocação de potência de transmissão através da seleção de Equipamento de Usuário (UE - *User Equipment*) em cada Intervalo de Tempo de Transmissão (TTI - *Transmission Time Interval*) de acordo com a Informação de Estado do Canal (CSI - *Channel State Information*), de modo a maximizar a vazão total de transmissão do sistema. A alocação de recursos em redes sem fio é normalmente representada por problemas de otimização com o objetivo de otimizar algum parâmetro (vazão ou atraso de pacotes) sujeito a algumas restrições sobre os recursos limitados (por exemplo, tempo de transmissão, largura de banda ou potência). A modelagem matemática da alocação de

recursos em um curto intervalo de tempo assume que o canal de comunicação é quase estático. Estes problemas são geralmente definidos como problemas de otimização determinísticos [Ding 2019].

De maneira similar em [Ding 2019], formulamos o problema de alocação de potência da BS aos UEs considerando transições de estado, onde a chegada de pacotes e a CSI dos UEs são variáveis aletórias. Estes problemas são geralmente referidos como problemas estocásticos de otimização. Neste artigo, utilizamos os agentes de Aprendizagem por Reforço (RL - *Reinforcement Learning*) e Aprendizagem por Reforço Profundo (DRL - *Deep Reinforcement Learning*), um ramo da Inteligência Artificial (IA), que utilizam redes neurais profundas (DNNs - *Deep Neural Networks*) na otimização do problema de alocação de potência.

As principais contribuições deste trabalho são resumidas abaixo:

- Proposta de utilização de algoritmos de RL e DRL aplicados ao problema de alocação de potência da BS aos UEs sem o conhecimento prévio da CSI e da chegada de pacotes;
- Modelagem do problema utilizando transições de estado e uma função de recompensa sem exigir uma expressão matemática de forma fechada, onde o processo de aprendizagem dos agentes ocorre através da interação com o ambiente estocástico;
- Validação do desempenho dos algoritmos de RL e DRL através de extensas simulações de rede. Os resultados mostram que a DRL supera significativamente a Aprendizagem Q (Q-*Learning*) que não é baseado em uma DNN.

O restante deste trabalho está organizado da seguinte forma: na seção 2, relatam-se os principais trabalhos relacionados com este artigo; na seção 3, apresenta-se o cenário do sistema de comunicação sem fio considerado; na seção 4, descreve-se o modelo de alocação de potência considerado; na seção 5, formula-se o problema de alocação de de potência através das transições de estados; na seção 6, descrevem-se algumas características da RL e sobre a motivação de evoluir para a DRL; na seção 7 discutem-se os resultados obtidos pelos agentes que aprendem através da função de recompensa; por fim, na seção 8, sintetizam-se as conclusões obtidas.

2. Trabalhos Relacionados

Levando em conta os desafios relatados para a alocação de recurso em redes sem fio moderna, tais como as redes da 5ª geração (5G) e 6ª geração (6G) de comunicação móvel, na última década, muitos projetos de pesquisas foram desenvolvidos alcançando resultados promissores. Em [Mauricio et al. 2019], os autores estudam o problema de alocação de recurso de rádio envolvendo a otimização de eficiência energética com requisitos específicos de Qualidade de Serviço (QoS - *Quality of Service*) em cenários multisserviços. Neste trabalho, vai-se além ao utilizar a aprendizagem por reforço para realizar alocação de potência. Em [Koo et al. 2019], os autores tratam do problema na alocação de recursos em um cenário de fatiamento de rede utilizando uma abordagem baseada no Processo de Decisão de Markov (MDP - *Markov Decision Process*) e na DRL. O modelo do sistema utiliza tráfego de dados 4G reais e sintéticos e os recursos que se pretendem alocar são a largura de banda e Máquinas Virtuais (MV). Porém, os autores não consideram a alocação de potência aos UEs, um fator que influência na qualidade do canal de comunicação.

Em [VASCONCELOS et al. 2020], os autores propõem a utilização do algoritmo Q - *Learning* para controlar a transmissão de pacotes de múltiplos dispositivos em um sistema de comunicação sem fio baseado no conceito de Internet das Coisas (IdC) Cognitivo. A abordagem proposta consiste em adotar uma cadeia de Markov para modelar os estados do sistema de comunicação e suas transições, fornecendo os parâmetros necessários para determinar ações para o sistema. Utilizando também o modelo do sistema de comunicação Markoviano os autores em [Carneiro et al. 2021] apresentam um algoritmo para alocação de recursos baseado em aprendizagem por reforço para um sistema de comunicação multiportadora considerando múltiplos usuários sobre os efeitos de desvanecimento e multipercurso em uma transmissão assumindo ondas milimétricas.

Em [Liu et al. 2021], os autores formulam o problema de alocação de recursos como um MDP Restritiva e o resolvem utilizando a aprendizagem por reforço restritiva e assumem que os padrões de tráfego e mobilidade dos usuários são desconhecidos para os algoritmos, os algoritmos exploraram e aprendem com a rede sem esses conhecimentos prévios. Uma das razões pelas quais as abordagens baseadas na aprendizagem, que incorporam a exploração, funcionam melhor do que os métodos tradicionais que não consideram a aleatoriedade do sistema e são baseados apenas na observação de estados.

A coexistência dos diversos serviços na rede sem fio utilizando o mesmo recurso de rádio leva a um problema de alocação de recursos desafiante que não é fácil modelar matematicamente devido ao compromisso com os vários requisitos de QoS (latência, potência, eficiência espectral e etc). O principal objetivo deste trabalho é resolver de forma flexível e eficiente o problema de alocação de potência para melhorar o atendimento aos requisitos de QoS desejados com a menor quantidade de potência possível. Para atingir este objetivo, propõe-se utilizar algoritmos de DRL que explorem e aprendam com o ambiente sem assumir o conhecimento de modelos matemáticos precisos.

3. Modelo do Sistema

O modelo do sistema de comunicação sem fio considerado neste trabalho está no sentido *downlink* e consiste de uma pequena célula (*Small Cell*) com uma BS no seu centro conforme mostra a Fig.1. A alocação de recursos é realizada a cada TTI e as posições dos UE na área de cobertura variam ao longo do tempo, resultando em CSIs variáveis através da distribuição do desvanecimento do canal de Rayleigh. Além disso, a chegada de pacotes é modelada utilizando a distribuição de Poisson.

Figura 1. Modelo do sistema de comunicação sem fio

Neste artigo é considerada a Multiplexação por Divisão de Frequências Ortogonais (OFDM - *Orthogonal Frequency Division Multiplexing*) como a técnica na transmissão LTE *downlink*, pois permite a transmissão simultânea de diferentes pacotes de dados, atribuindo diferentes subportadoras ao usuário. No domínio do tempo, a duração do *frame downlink* é de 10 ms. Este *frames* são divididos em 10 *sub-frames*, onde cada *sub-frame* representa um TTI de 1 ms.

4. Alocação de Potência

O método de Enchimento de Água (*Water Filling*) tem como objetivo alocar a potência do sinal da BS aos UEs de tal forma a maximizar a soma da taxa de dados da rede e atender a restrição de alocar potência com valores positivos aos UEs mantendo-se dentro da disponibilidade de potência da BS [Ding 2019]. Neste trabalho foi escolhido o método *Water Filling* por ele resolver o problema de otimização descrito pelas Equações (1), (2) e (3) com complexidade computacional linear através da solução de um sistema com (Z + 1) equações e (Z + 1) incógnitas. Esta complexidade computacional é inferior a de outros métodos da literatura que poderiam ser utilizados para resolver o problema como os algoritmos genéticos.

$$\max_{P} \quad \sum_{i=1}^{Z} log\left(1 + p_i.\frac{||h_i(t)||^2}{\sigma_z^2}\right), \tag{1}$$

$$S.a \quad \sum_{i=1}^{Z} p_i \leq P_{bs}, \tag{2}$$

$$p_i \geq 0, \quad i = 1, 2, ..., Z. \quad P_{bs} \geq 0, \tag{3}$$

onde Z é o número de UEs, P_{bs} é a potência da BS e p = $p_1, p_2, ..., p_Z$ é a potência alocada aos UEs. A expressão $\frac{||h_i(t)||^2}{\sigma_z^2}$ representa o CSI e é dada pela distribuição do desvanecimento de Rayleigh. Aplicando o método Dual de Lagrange ($\mathcal{L}$) temos que a função objetivo (1) pode ser reduzida para:

$$\mathcal{L}(p, \beta) = \sum_{i=1}^{Z} log\left(1 + p_i.\frac{||h_i(t)||^2}{\sigma_z^2}\right) - \beta(\sum_{i=1}^{Z} p_i - P_{bs}), \tag{4}$$

onde β é o multiplicador de Lagrange. Uma solução para o problema dual da Equação (4) é obtida igualando a zero o gradiente, isto é $\nabla\mathcal{L}(p, \beta)$ = 0, conforme a seguir:

$$\frac{\partial\mathcal{L}(p, \beta)}{\partial p} = \frac{\frac{||h_i(t)||^2}{\sigma_z^2}}{1 + p_i.\frac{||h_i(t)||^2}{\sigma_z^2}} - \beta = 0. \tag{5}$$

$$p_i = \frac{1}{\beta} - \frac{1}{\frac{||h_i(t)||^2}{\sigma_z^2}}. \tag{6}$$

A escolha do valor de β determina o "nível da água" para a restrição da soma de potência alocada aos UEs, ou seja, a potência total consumida da BS. Sendo assim, a alocação de potência para todos os UE pode ser reescrita como:

$$p_i(t) = max\left\{0, \frac{1}{\beta_t} - \frac{\sigma_z^2}{||h_i(t)||^2}\right\}, \forall i = 1, 2, ..., Z. \tag{7}$$

Como a quantidade de potência alocada aos UEs com melhor qualidade de canal é maior, o algoritmo interpreta que não é necessário que alguns UEs tenham potência alguma alocada, já que o nível de ruído está demasiadamente alto.

5. Formulação do Problema

Neste artigo, propõe-se a aplicação dos algoritmos de RL e DRL para resolver o problema de alocação de potência em redes sem fio baseadas em OFDM. Nesta abordagem de aprendizado por reforço, considera-se que os estados do sistema sejam representados pelo número de pacotes que estão esperando para transmissão no *buffer* $q(t)$, e a ação é dada pela variável β que relaciona-se com a potência alocada aos UEs. A função de recompensa consiste em minimizar o retardo médio $\frac{q(t)}{\lambda}$ e a soma da potência que é alocada aos UEs ($\sum_{i=1}^{Z} p_i$), onde λ é o parâmetro da distribuição de Poisson para a chegada de pacotes.

Os principais elementos para a aplicação dos agentes ao problema de alocação de potência são descritos a seguir:

Estado do sistema: considera-se o número de pacotes que estão esperando para ser transmitido no t-ésimo *subframe* como o estado do sistema $s(t)$, ou seja, os pacotes que estão no *buffer*:

$$s(t) = \{q(t)\}. \tag{8}$$

Um *buffer* de tamanho R permite armazenar no máximo R pacotes para todos os UEs. Os estados do sistema são obtidos levando em conta os (R + 1) estados possíveis com inclusão do zero. Dessa forma, há mudanças no estado do *buffer* com a chegada ou saída de pacotes a cada *subframe*.

Política de controle: dado a distribuição aleatória da CSI dos UEs, a ação de controle é a alocação de potência utilizando a Eq. (7) através do valor de β que neste trabalho é discretizado em duas casas decimais $\in$ (0,01, 1):

$$a(t) = \{\beta\}. \tag{9}$$

Transição de estados: dado o estado do sistema, CSI dos UEs e a ação de controle do t-ésimo *subframe*, a transição de estado é representada pela equação de fila dinâmica dada por:

$$q(t+1) = max\{0, q(t) - d(t)\} + c(t), \tag{10}$$

onde $q(t+1)$ é o novo estado do sistema $s(t+1)$, $c(t)$ é o número de pacotes que chegam no t-ésimo *subframe*. Assume-se que $c(t)$ siga a distribuição de Poisson com parâmetro λ. Sendo assim, há em média λ pacotes em cada *subframe*. Note também que $d(t)$ é a capacidade do canal e representa a quantidade de pacotes que são transmitidos no t-ésimo *subframe* e é dada por:

$$d(t) = \left(\frac{\sum_{i=1}^{Z} L_i \times log_2(1 + \frac{P_i . ||h_i||^2}{\sigma_z^2})}{B} \right), \tag{11}$$

onde L_i é a largura de banda, neste artigo, assume-se uma alocação de largura de banda igualitária para todos os UEs, B é a quantidade de *bits* em um pacote, foi considerado um tamanho de pacote de 1024 *bytes*, o tamanho do pacote é baseado no padrão IEC-61850 para distribuição de energia de média e alta voltagem [Wong and Das 2014].

Função Recompensa: Segundo a lei de *Little's* o retardo médio de pacotes é dado por [Kleinrock 1975]:

$$W = \frac{Q}{\lambda} = \lim_{T \to \infty} E\left(\frac{1}{T} \sum_{i=1}^{T} \frac{q(t)}{\lambda} \right), \tag{12}$$

onde Q é o número médio de pacotes à espera de ser transmitido. Sendo assim, a função de recompensa é dada por:

$$r(t) = -\sum_{i=1}^{Z} p_i - \alpha * W, \tag{13}$$

onde α é o peso sobre o retardo médio na transmissão de pacotes. Geralmente os agentes de RL aprendem com o objetivo principal de maximizar a função de recompensa. Neste trabalho, o objetivo principal é a redução do retardo médio com a menor quantidade de potência de transmissão do sinal possível, a função de recompensa possui valor negativo na equação. Portanto, o seu valor é minimizado. A fim de obter maior prioridade, o retardo médio é ponderado pelo peso α.

6. Aprendizagem por Reforço

Os algoritmos baseados em valor são utilizados para estimar a função valor do agente. Esta função valor é então utilizada para se obter de forma implícita e gananciosa uma política ótima. Existem dois tipos de funções baseada em valor: a função valor $V^{\pi}(s)$ e a função estado-ação $Q(s(t), a(t))$. Ambas representam a recompensa descontínua acumulada esperada recebida quando tomando uma ação $a(t)$ no estado $s(t)$ para a função valor

ou para a função estado-ação. Essas funções são bastante importantes, uma vez que representam a ligação entre a formulação matemática do MDP e a RL. A função valor $V^*(s)$ e a função estado-ação $Q^*(s(t), a(t))$ ótima são obtidas pelas equações de otimilidade de Bellman [Alwarafy et al. 2021]:

$$V^*(s) = \max_{a(t)}[r_t(s(t), a(t)) + \gamma E_\pi V^*(s(t+1))]. \tag{14}$$

$$Q^*(s(t), a(t)) = r_t(s(t), a(t)) + \gamma E_\pi[\max_{a(t+1)} Q^*(s(t+1), a(t+1))], \tag{15}$$

onde γ é o fator de desconto $\in$ (0,1) e $r_t(s(t), a(t))$ é a recompensa de se tomar a ação $a(t)$ no estado $s(t)$. O principal objetivo do MDP é obter a política ótima π^*, ou seja, mapear os estados para otimizar as ações. Para a função valor a política é dada por:

$$\pi^* = \arg\max_{a(t)} = E[\sum_{t=1}^{T} \gamma r_t(s(t), \pi(s(t)))] \tag{16}$$

Para a função Q, a política ótima se torna:

$$\pi^*(s) = \arg\max_{a} Q^{\pi *}(s(t), a(t)) \tag{17}$$

Em RL, o Q-*Learning* é o algoritmo mais usado para abordar MDPs. Obtém ótimos valores da função $Q(s(t), a(t))$ utilizando iterativamente a regra de atualização da equação de Bellman [Alwarafy et al. 2021]:

$$Q(s(t), a(t)) = Q(s(t), a(t)) + \omega[r_t(s(t), a(t)) + \gamma \max_{a(t+1)} Q^*(s(t+1), a(t+1)) \\ -Q(s(t), a(t))], \tag{18}$$

onde ω é a taxa de aprendizagem.

6.1. Double Deep Q-Network (DDQN)

O algoritmo Q-*Learning* se baseia na construção de uma tabela para os valores da função Q. Devido a esta razão, quando o espaço de estado e o espaço de ação tornam-se grandes como nos casos normalmente encontrados nos problemas de gerenciamento de recursos de rádio em sistemas sem fios modernos, obter a política ótima pode ser extremamente demorado.

Para solucionar este problema surgiu a Rede Q Profunda (DQN - *Deep Q Network*) que herda as vantagens das técnicas comumente utilizadas no Q- *Learning* e na Aprendizagem Profunda (DL - *Deep Learning*). A ideia principal é substituir a tabela do algoritmo Q-*Learning* por uma Rede Neural Profunda (DNN - *Deep Neural Network*) que aproxima os valores de Q. A saída da DNN do DQN é o valor de Q para todas as possíveis ações do problema.

A DNN também é chamada de função de aproximação universal e é designada por $Q(s(t), a(t)|\Theta)$, onde Θ representa os parâmetros ou os pesos da DNN. Para aumentar a estabilidade do DQN, é utilizada outra rede neural, chamada de rede Q objetivo, cujos pesos $\Theta^{'}$ serão periodicamente atualizados para seguir os da rede neural Q principal [Alwarafy et al. 2021]. O algoritmo DQN é otimizado por iteratividade atualizando os pesos Θ da sua DNN para minimizar a seguinte função de perda de Bellman:

$$L(\Theta_t) = E[r_t(s(t), a(t)) + \gamma \max_{a(t+1)} Q(s(t+1), a(t+1)|\Theta^{'}) - Q(s(t), a(t)|\Theta)]^2, \quad (19)$$

onde $\Theta^{'}$ são os pesos da rede Q objetivo.

O algoritmo DQN tende a sobrestimar os valores Q, o que pode degradar o processo de treinamento e conduzir a políticas sub-ótimas. A sobrestimação resulta do enviesamento positivo causado pela operação máxima empregada na equação de Bellman. Especificamente, a causa raiz é que as mesmas transições de treinamento são utilizadas na seleção e avaliação de uma ação [Alwarafy et al. 2021].

Como solução para este problema, neste artigo, propõe-se utilizar a técnica Duplo DQN (DDQN - *Double* DQN), onde emprega-se duas funções para o valor Q, uma para selecionar a melhor ação e a outra para avaliar a melhor ação. A seleção da ação ainda baseia-se nos pesos Θ, enquanto que os segundos pesos $\Theta^{'}$ são utilizados para avaliar o valor desta política conforme mostra a Fig. 2.

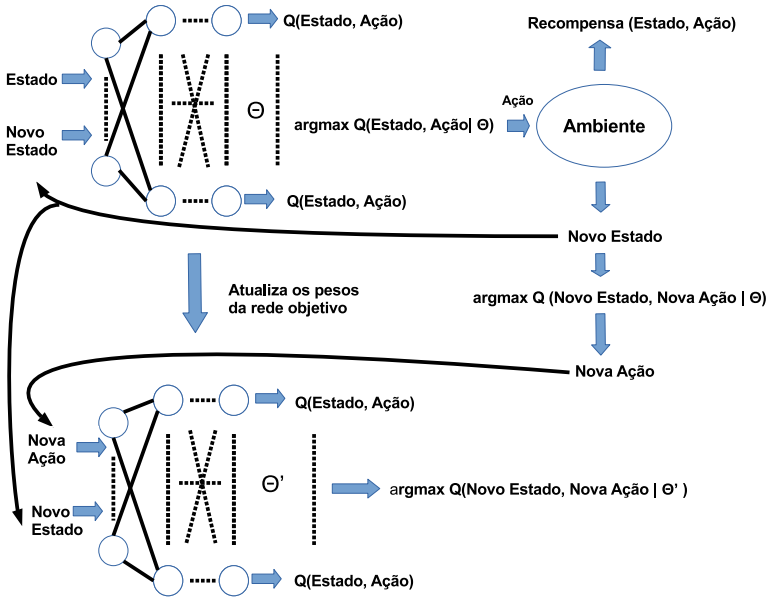


Figura 2. *Double* DQN

Portanto, assim como no DQN, o valor da política ainda é estimado com base no valor de Q atual. Os pesos $\Theta^{'}$ são atualizados através de Θ [Alwarafy et al. 2021]. O

algoritmo DDQN utiliza a seguinte função de perda de Bellman modificada para atualizar os seus pesos:

$$L(\Theta_t) = E[r_t(s(t), a(t)) + \gamma Q(s(t+1), \arg\max_{a(t+1)} Q(s(t+1), a(t+1)|\Theta), \Theta^{'}) \quad (20)$$
$$-Q(s(t), a(t)|\Theta)]^2.$$

7. Resultados e Discussões

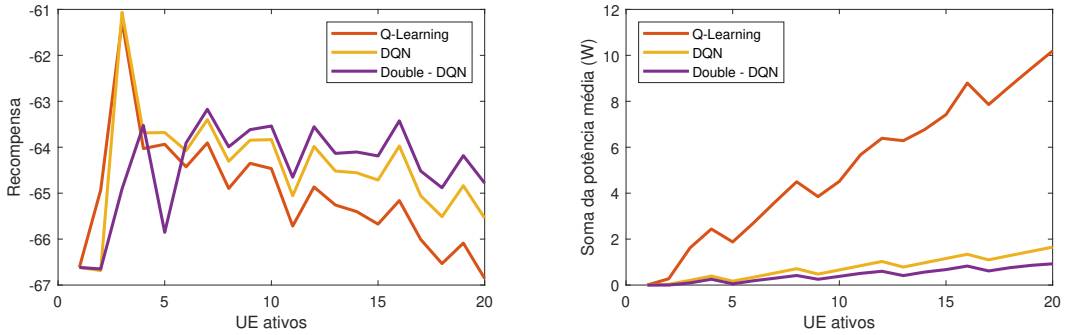
Neste trabalho, foram realizadas simulações utilizando as seguintes configurações de *software* e *hardware*: *software* Matlab versão R2021, processador Intel Core i5-1035G1 de 1,00 GHz; 8 GB de RAM sem placa de vídeo dedicada. O conjunto completo de parâmetros das simulações é fornecido na Tabela 1.

Tabela 1. Parâmetros da Simulação

Parâmetros	Valores
Potência de transmissão da BS (P_{bs})	10 W
Tamanho máximo do buffer (R)	4 pacotes
Intervalo de tempo de transmissão	1 ms
Número de símbolos OFDM por TTI	7
Largura de banda	20 MHz
Modelagem do canal	Distribuição de Rayleigh
Chance de escolha aleatória (ϵ)	0,1
Taxa de aprendizagem (ω)	0,01
Peso sobre o retardo médio (α)	50
Fator de desconto (γ)	0,90
Taxa de chegada de pacotes (λ)	Distribuição de Possoin
Camadas oculta da rede neural densa	3
Quantidade de neurônios em cada camada	500
Função de ativação de cada camada oculta	Leaky ReLU
Quantidade de *bits* em um pacote (B)	8192
Iterações de treinamento	10000
Tempo de simulação	1000 TTI

As Figs. 3-5 mostram os desempenhos dos algoritmos de RL e DRL (Q-*Learning*, DQN e DDQN) na alocação de potência da BS a medida que aumenta o número UEs ativos na rede. A Fig. 3(a) apresenta a recompensa dos agentes e o algoritmo DDQN é o que apresenta o maior valor de recompensa entre os três algoritmos. Os maiores valores de recompensa ao longo do tempo obtidos pelo algoritmo DDQN em relação aos outros algoritmos considerados indicam que este apresenta uma maior capacidade de aprendizagem para o problema proposto. A Fig. 3(b) mostra a média da potência do sinal de transmissão que é alocada aos UEs com intuito de aumentar a capacidade do canal de transmissão e enviar os pacotes que estão em espera na fila. Pela Fig. 3(b), observa-se que o Q-*Learning* consome uma maior potência que quase extrapola a potência total disponível da BS. Note que o algoritmo DDQN é o que apresenta maior economia em termos de potência comparada aos outros algoritmos considerados.

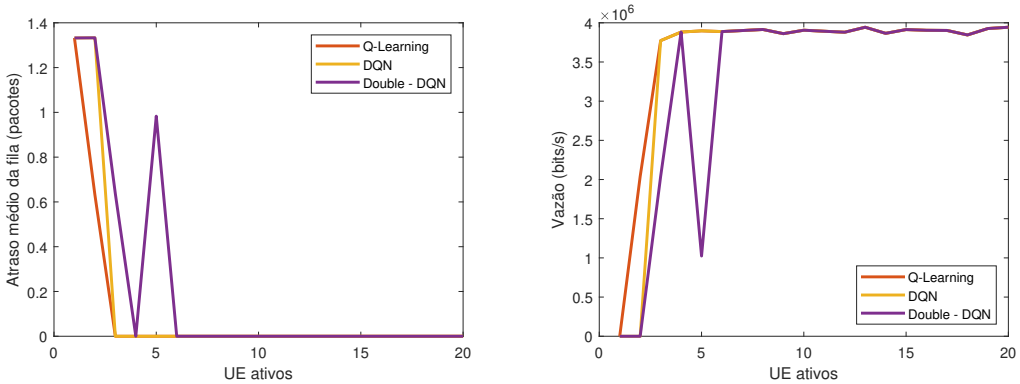
Figura 3. Recompensa média e a potência média total alocada aos UEs pelos agentes Q-Learning, DQN e DDQN.



(a) Recompensa média (b) Soma da potência média alocada aos UEs

Neste trabalho, foi atribuído um alto peso sobre o atraso médio na transmissão de pacotes, para que a prioridade na função de recompensa seja diminuir o atraso com a menor quantidade de potência possível. Os resultados da Fig. 4 revelam que o maior consumo de potência obtido com o algoritmo Q-*Learning* conforme mostra a Fig. 3(b) não trouxe benefícios em termos do atraso médio e vazão de pacotes. À medida que o número de UEs ativos no sistema aumenta, temos um aumento da soma de potência média alocada (Fig. 3(b)) que faz com que a vazão média (Fig. 4(b)) também aumente e se mantenha aproximadamente constante a partir de 7 UEs no sistema. Esse aumento de vazão faz com que a partir de 7 UEs todos os algoritmos em média esvaziem o *buffer* gerando atraso médio nulo (Fig. 4(a)). Note também que o algoritmo DDQN obteve em média os mesmos resultados de atraso de pacotes e vazão a partir de 7 UEs ativos na rede entre os algoritmos de alocação de potência considerados mas com o menor consumo de potência.

Figura 4. Atraso médio e vazão média na transmissão de pacotes para cada agente.

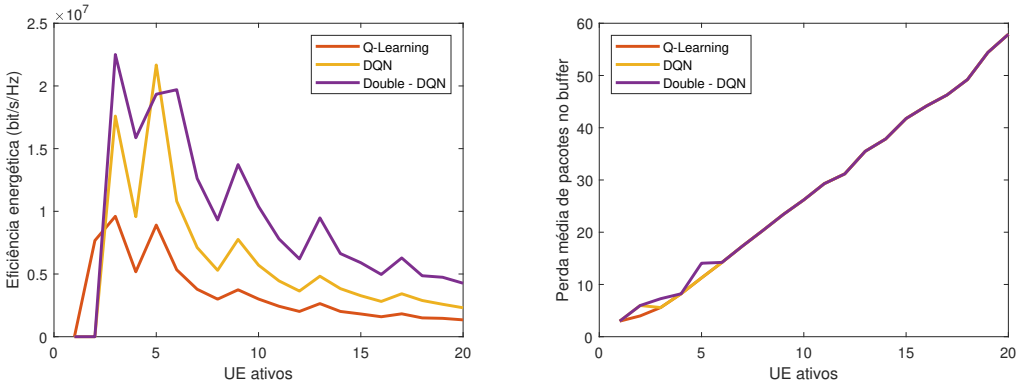


(a) Atraso médio (b) Vazão média

A Eficiência Energética (EE) emergiu como um indicador chave de desempenho para as futuras redes, uma mudança para comunicações otimizadas. A EE pode ser melhorada utilizando diferentes estratégias, tais como planejamento e desenvolvimento da infra-

estrutura de rede, captação de energia e alocação de recurso de rádio [Buzzi et al. 2016]. A métrica de EE considerada neste trabalho é a razão entre a vazão média pela potência média total de transmissão consumida. Conforme mostra a Fig. 5(a), o agente DDQN apresenta o melhor resultado em termos de EE. Outro indicador analisado é a perda de pacotes no *buffer*. Conforme mostra a Fig. 5(b), com o aumento do número de UEs ativos no sistema, aumenta-se a chegada de pacotes além da capacidade do *buffer* gerando perda de pacotes. Os algoritmos de alocação de potência provêm perdas de pacotes que aumentam proporcionalmente com o número de UEs ativo na rede. Entretanto, a partir de 7 UEs no sistema, as perdas de pacotes praticamente se igualam para todos os algoritmos de alocação de potência considerados.

Figura 5. Eficiência energética na transmissão de pacotes e perda média de pacotes no buffer para cada agente.



(a) Eficiência energética média

(b) Perda média de pacotes

8. Conclusão

Neste artigo, propõe-se a aplicação de agentes de RL e DRL para solucionar o problema de alocação de potência do sinal de transmissão da BS aos UEs. O problema foi modelado através de transições de estados sem considerar o conhecimento prévio das probabilidades das transições entre os estados, ou seja, o agente inteligente aprende adaptativamente com base nas observações dos estados aleatórios do sistema. As extensivas simulações mostraram que o agente DDQN apresenta melhores resultados em termos de eficiência energética em relação aos outros agentes devido a sua melhor capacidade de aprendizagem do ambiente através da função de recompensa proposta. Em um cenário com mais de 7 UEs ativos na rede o agente DDQN realiza um consumo eficiente de potência apresentando uma maior economia.

Os métodos de Aprendizagem por Reforço Profundo são considerados técnicas promissoras para a alocação de recursos em redes sem fio, como apontado em [Koo et al. 2019], [Liu et al. 2021] e [Zhou et al. 2021]. Devido ao seu grande potencial de aprendizagem, obtém resultados interessantes monitorando o ambiente, sem o conhecimento das estatísticas do sistema a priori. Os resultados das simulações apresentados neste artigo confirmam a eficiência da aprendizagem do algoritmo DDQN para alocação de potência aos UEs. Finalmente, em trabalhos futuros, pretende-se adaptar o cenário a fim de considerar as redes móveis sem fio da próxima geração.

9. Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa no Estado de Goiás (FAPEG) e ao Centro de Excelência em Inteligência Artificial (CEIA) pelos apoios no desenvolvimento da pesquisa.

Referências

Alwarafy, A., Abdallah, M. M., Ciftler, B. S., Al-Fuqaha, A. I., and Hamdi, M. (2021). Deep reinforcement learning for radio resource allocation and management in next generation heterogeneous wireless networks: A survey. *CoRR*, abs/2106.00574.

Buzzi, S., I, C.-L., Klein, T. E., Poor, H. V., Yang, C., and Zappone, A. (2016). A survey of energy-efficient techniques for 5g networks and challenges ahead. *IEEE Journal on Selected Areas in Communications*, 34(4):697–709.

Carneiro, D., Cardoso, A., Almeida, C., and Vieira, F. (2021). Aprendizado por reforço para escalonamento de recursos em sistema sem fio multiportadora com ondas milimétricas utilizando modelo markoviano. In *Anais da IX Escola Regional de Informática de Goiás*, pages 12–25, Porto Alegre, RS, Brasil. SBC.

Ding, Z., editor (2019). *Applications of Machine Learning in Wireless Communications*. Telecommunications. Institution of Engineering and Technology.

Kleinrock, L. (1975). *Theory, Volume 1, Queueing Systems*, volume 1. Wiley-Interscience, USA.

Koo, J., Mendiratta, V. B., Rahman, M. R., and Walid, A. (2019). Deep reinforcement learning for network slicing with heterogeneous resource requirements and time varying traffic dynamics. In *2019 15th International Conference on Network and Service Management (CNSM)*, pages 1–5.

Liu, Y., Ding, J., and Liu, X. (2021). Resource allocation method for network slicing using constrained reinforcement learning. In *2021 IFIP Networking Conference (IFIP Networking)*, pages 1–3.

Mauricio, F., Vinicius, W., Lima, M., and et. al (2019). Resource allocation for energy efficiency and qos provisioning. *Journal of Communication and Information Systems*, 34(1):224–238.

VASCONCELOS, M. M., A., C. A., and VIEIRA, F. H. T. (2020). Aprendizado por reforço e modelo markoviano para alocação de recursos em um sistema internet das coisas cognitivo. *XXXVIII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais - SBrT 2020*.

Wong, T. J. and Das, N. (2014). Modelling and analysis of iec 61850 for end-to-end delay characteristics with various packet sizes in modern power substation systems. In *5th Brunei International Conference on Engineering and Technology (BICET 2014)*, pages 1–6.

Zhou, H., Elsayed, M. H. M., and Erol-Kantarci, M. (2021). RAN resource slicing in 5g using multi-agent correlated q-learning. In *32nd IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2021, Helsinki, Finland, September 13-16, 2021*, pages 1179–1184. IEEE.