

PDO-URA-RL: Otimização de Potência e Retardo para D2D no Uplink 5G com Aprendizado por Reforço

Marcus V. G. Ferreira, Leonardo A. Melo e Flávio H. T. Vieira

Resumo— Este artigo aborda o problema de alocação de recursos no enlace uplink de redes 5G com ondas milimétricas (mmWaves) e comunicação dispositivo a dispositivo (D2D). É proposto o algoritmo PDO-URA-RL (Power and Delay Optimization based Uplink Resource Allocation with Reinforcement Learning), que integra otimização conjunta de potência e retardo com aprendizagem por reforço. O método aloca potência e taxa de transmissão aos CUEs (cellular user equipments) visando maximizar a vazão efetiva, enquanto os recursos ociosos são distribuídos aos pares D2D com foco na minimização do retardo. O desempenho do algoritmo proposto é avaliado em cenários compatíveis com redes 5G, considerando métricas de qualidade de serviço (QoS) como vazão total e retardo médio, e comparado com algoritmos representativos da literatura especializada.

Palavras-Chave— 5G, D2D, alocação de recursos.

Abstract— This paper addresses the resource allocation problem in the uplink of 5G mmWave networks with device-to-device (D2D) communication. The PDO-URA-RL (Power and Delay Optimization based Uplink Resource Allocation with Reinforcement Learning) algorithm is proposed, which integrates joint power-delay optimization with reinforcement learning. The method allocates power and transmission rate to CUEs (cellular user equipments) aiming to maximize effective throughput, while idle resources are distributed to D2D pairs focused on delay minimization. The performance of the proposed algorithm is evaluated in 5G-compatible scenarios considering quality-of-service (QoS) metrics such as total throughput and average delay, and compared with representative algorithms from the literature.

Keywords— 5G, D2D, resources allocation.

I. INTRODUÇÃO

O crescimento exponencial da demanda por comunicações sem fio tem impulsionado o desenvolvimento de soluções capazes de ampliar a vazão e reduzir a latência das redes celulares, especialmente para atender aplicações emergentes como internet das coisas, comunicação máquina a máquina e realidade aumentada. Nesse cenário, a comunicação dispositivo a dispositivo (D2D) destaca-se como uma técnica promissora para aumentar a eficiência espectral, permitindo a comunicação direta entre terminais com menor dependência da estação rádio base (eNodeB) e reduzindo a sobrecarga do tráfego centralizado. Adicionalmente, devido à assimetria característica do tráfego de dados nas redes móveis, observa-se maior disponibilidade de recursos ociosos no enlace *uplink*,

motivando pesquisas voltadas ao compartilhamento múltiplo de recursos nesse enlace.

Diversos trabalhos da literatura investigam a alocação de recursos para comunicações D2D no *uplink*. [Sun et al. 2013] propuseram o algoritmo GRA (*greedy resources allocation*) para maximizar o número de pares D2D simultaneamente ativos com mínima interferência. [Ciou et al. 2015] desenvolveram o GTM+ (*greedy throughput maximization plus*), estendendo a abordagem anterior ao cenário de compartilhamento múltiplo mediante grafos de conflito e conjuntos independentes de peso máximo, reduzindo a complexidade computacional de um problema reconhecidamente NP-difícil (não-determinístico polinomial de tempo difícil). [Sun et al. 2025] propuseram o MTMF-Q (*Multi-Type Mean-Field double deep Q-network*), combinando aprendizado por reforço multiagente (MARL) com teoria de jogos de campo médio para alocação de recursos em redes D2D heterogêneas. No contexto de redes 5G, [Mangal et al. 2025] apresentaram o RLD2D, que utiliza Q-learning para estabelecimento dinâmico de links D2D com foco em baixa latência e eficiência energética. [Attar et al. 2025] propuseram um esquema semi-distribuído que combina algoritmos de *matching* com SARSA (*State-Action-Reward-State-Action*) para alocação de potência em redes D2D com compartilhamento de sub-canais.

Apesar dos avanços, tais abordagens não contemplam de forma integrada a minimização do retardo e a otimização conjunta da potência de transmissão e da alocação de recursos. Diferentemente dos trabalhos existentes, este artigo propõe o algoritmo PDO-URA-RL (*power and delay optimization based uplink resource allocation with reinforcement learning*), que integra otimização de potência, minimização de retardo e aprendizagem por reforço na formação dos links D2D. O algoritmo é estruturado em duas etapas principais. Na primeira, o eNodeB aloca recursos para os CUEs (*cellular user equipments*) considerando simultaneamente a otimização do retardo e da potência transmitida, maximizando o *goodput* em função da relação sinal-interferência-ruído (SINR). Na segunda etapa, os recursos ociosos são reutilizados pelos DUEs (*D2D user equipments*) por meio de uma estratégia baseada em Q-learning combinada ao algoritmo DMCG (*delay minimization conflict graph*), que minimiza o retardo mediante grafos de conflito e estimativas fundamentadas em processos envelope multifractais e curvas de serviço.

O cenário de simulação considera comunicações em ondas milimétricas na frequência de 26 GHz (faixa FR2 do 5G), modelo de canal CDL-D e métricas de qualidade de serviço (QoS)

como vazão total e retardo médio. Os resultados demonstram que o PDO-URA-RL proporciona ganhos significativos em vazão e redução do retardo quando comparado aos algoritmos GTM+ e DMCG, especialmente em cenários com maior número de CUEs.

II. MODELO DO SISTEMA

O modelo de sistema para comunicações D2D considerado neste trabalho é ilustrado na Figura 1. Nesse cenário, DUEs (D2D user equipments) reutilizam os recursos de rádio previamente alocados aos CUEs, estabelecendo comunicação direta entre si.

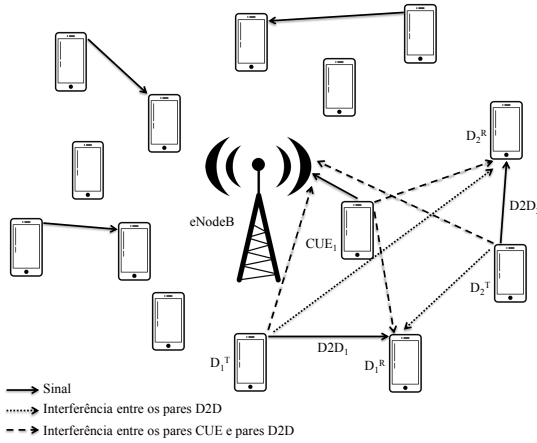


Fig. 1

MODELO DE SISTEMA PARA COMUNICAÇÕES D2D COM DUEs
REUTILIZANDO RECURSOS *uplink* DE CUEs.

Assume-se que os CUEs reais não utilizam RBs (*resource blocks*) ociosos. Mais especificamente, cada RB ocioso, isto é, não utilizado por nenhum CUE real, é associado a um CUE virtual com potência de transmissão nula e sem exigência de um valor mínimo de SINR. Assim, o sistema D2D é composto por N DUEs, M CUEs — incluindo CUEs reais e virtuais —, K RBs ociosos agendados para transmissão durante um TTI (*transmission time interval*) e $M - K$ CUEs reais. Um RB representa a menor unidade de alocação no sistema de comunicação sem fio. Os CUEs são denotados por $C_1, C_2, \dots, C_M$ e os DUEs por $D_1, D_2, \dots, D_N$. Para cada par de DUE n , o transmissor é denotado por $D_{n,Tx}$ e o receptor por $D_{n,Rx}$.

Antes do compartilhamento dos RBs com os DUEs, um conjunto de RBs é pré-alocado a cada CUE. A largura de banda W_m atribuída ao CUE m é proporcional ao número de RBs alocados. No modelo de sistema adotado, considera-se a possibilidade de um grande número de conexões simultâneas entre pares de DUEs, uma vez que cada CUE pode compartilhar seus RBs com múltiplos pares de DUEs, conforme o conceito de alocação de recursos com compartilhamento múltiplo.

O conjunto Θ_n representa a relação entre CUEs e o par de DUE n , sendo definido como o conjunto de CUEs que

compartilham seus RBs com esse par. De forma complementar, o conjunto Δ_m representa o conjunto de pares de DUEs que reutilizam os RBs alocados ao CUE m . Os conjuntos $\Delta_1, \Delta_2, \dots, \Delta_M$ são subconjuntos de $\{1, 2, \dots, N\}$.

Considera-se que a interferência impõe restrições à alocação de RBs para os DUEs no *uplink*. Assim, um par de DUEs pode reutilizar os RBs de um CUE m somente se a interferência gerada na transmissão de m atender ao requisito de SINR. Desse modo, o SINR recebido por qualquer CUE $m \in \{1, 2, \dots, M\}$ deve satisfazer:

$$\frac{P_m G_{mB}}{\sigma_m^2 + \sum_{n \in \Delta_m} P_n G_{nB}} \geq \gamma_m, \quad (1)$$

onde P_m é a potência de transmissão do CUE m , G_{mB} é o ganho de canal entre o CUE m e o eNodeB, σ_m^2 é a potência de ruído associada ao CUE m , P_n é a potência de transmissão do par de DUE n e G_{nB} é o ganho de canal entre o transmissor do par de DUE n e o eNodeB para SINR maior que zero e γ_m é o limiar de SINR exigido pelo CUE m . Os ganhos de canal G_{mB} e G_{nB} dependem do modelo de canal e das condições de propagação do sinal. Neste trabalho, tais valores são obtidos a partir do modelo de canal CDL-D (*clustered delay line*) [3gp 2017].

De forma análoga, também existem requisitos de SINR para os pares de DUEs. Os RBs podem ser reutilizados por um par de DUE n apenas se o SINR recebido satisfizer:

$$\frac{P_n G_{nn}}{\sigma_n^2 + P_m G_{mn} + \sum_{n' \in \Delta_m - \{n\}} P_{n'} G_{n'n}} \geq \gamma_n, \forall m \in \Theta_n, \quad (2)$$

onde G_{nn} é o ganho de canal entre os dois terminais do par de DUE n , σ_n^2 é a potência de ruído associada ao par de DUE n , G_{mn} é o ganho de canal do CUE m para o par de DUE n , $G_{n'n}$ é o ganho de canal entre $D_{n',Tx}$ e $D_{n,Rx}$ para SINR maior que zero e γ_n é o limiar de SINR exigido pelo par de DUE n . Os ganhos de canal G_{nn} , G_{mn} e $G_{n'n}$ também dependem da modelagem de canal e das condições de propagação do sinal, sendo igualmente obtidos por meio do modelo CDL-D [3gp 2017].

Assume-se ainda que o eNodeB dispõe de informações relevantes do estado do canal, tais como potência de transmissão, ganho de canal e potência de ruído dos enlaces de comunicação associados aos CUEs m e aos pares de DUEs n , bem como os respectivos limiares de SINR. Essa hipótese é amplamente adotada em trabalhos relacionados à alocação de recursos em redes sem fio, como [Sun et al. 2013], [Ciou et al. 2015] e [Ferreira and Vieira 2021], com o objetivo de simplificar os cenários de simulação. Neste artigo, foca-se especificamente no problema de alocação de recursos, sem perda de generalidade, assumindo-se que os parâmetros de ganho de canal utilizados nas simulações do algoritmo proposto poderiam ser estimados pelo eNodeB por meio de técnicas de estimativa de canal amplamente discutidas na literatura [Belgiovine et al. 2021].

Em um sistema de comunicação D2D, o objetivo do problema de alocação de recursos com compartilhamento múltiplo consiste em determinar quais pares de DUEs reutilizam os

RBs dos CUEs de modo a maximizar o rendimento total do sistema. Esse problema é NP-difícil, conforme demonstrado em [Sun et al. 2013] e [Ciou et al. 2015]. Tal complexidade reforça a necessidade de abordagens como a proposta neste artigo, capazes de fornecer soluções rápidas e eficientes para o problema de alocação de recursos.

III. ALGORITMO DE ALOCAÇÃO DE RECURSOS NO UPLINK BASEADO NA OTIMIZAÇÃO DE POTÊNCIA E RETARDO E NA APRENDIZAGEM POR REFORÇO (PDO-URA-RL)

Nesta seção, propõe-se um algoritmo de alocação de recursos para dispositivos em redes sem fio com compartilhamento múltiplo no *uplink*, utilizando CP-OFDM (*cyclic prefix-orthogonal frequency division multiplexing*) e comunicações D2D. O algoritmo proposto é estruturado em duas etapas. Na primeira etapa, os recursos da rede são alocados aos CUEs em termos de potência e taxa de transmissão, por meio de uma abordagem que visa maximizar a vazão. Na segunda etapa, os recursos ociosos são alocados com foco na minimização do retardo, empregando um método de aprendizagem por reforço baseado em Q-learning para a formação dos links D2D. Para essa finalidade, utiliza-se adicionalmente o algoritmo DMCG, no qual os recursos ociosos são compartilhados com os DUEs considerando os grafos de conflito formados e as informações de retardo estimadas por meio de conceitos de Cálculo de Rede, tais como o processo envelope e a curva de serviço.

A. Determinação de Pares D2D Utilizando Q-Learning

A abordagem de aprendizagem por reforço adotada tem como objetivo a construção de uma rede de comunicação D2D com alta conectividade e retardo mínimo. O método baseia-se no algoritmo Q-learning, amplamente utilizado em aprendizagem por máquina para tomada de decisões em ambientes dinâmicos e com informações incompletas [Nagarajan and Mohamad 1992]. Uma de suas principais vantagens é a simplicidade de implementação, uma vez que incorpora os principais fatores que afetam o desempenho da rede e não requer conhecimento prévio detalhado da topologia ou das condições da rede de comunicação D2D.

Os principais elementos do modelo de aprendizagem por reforço são: agente, estado, ação, recompensa e ambiente. Na rede de comunicação D2D, os dispositivos responsáveis pela criação dos links de comunicação atuam como agentes. O estado do ambiente S de cada dispositivo d_i é definido pelo conjunto de seus dispositivos vizinhos e pelo número de pacotes presentes em sua fila de *buffer*, sendo esse estado utilizado como critério de decisão pelo algoritmo. Assim, tem-se $S = \{(n_1, p_1), (n_2, p_2), \dots, (n_M, p_M)\}$, onde $n_1, n_2, \dots, n_M$ representam o número de dispositivos vizinhos e $p_1, p_2, \dots, p_M$ correspondem ao número de pacotes na fila de *buffer* dos dispositivos $d_1, d_2, \dots, d_M$, respectivamente. Para cada estado $s \in S$, o dispositivo executa uma ação a , que determina o link de comunicação D2D a ser estabelecido. A ação escolhida influencia tanto o estado subsequente quanto a recompensa associada, uma vez que o agente interage com o ambiente e atualiza o valor da recompensa RW .

O desempenho da rede ao executar uma ação em um determinado estado é avaliado por meio de uma função de recompensa. O conjunto de vizinhos de um dispositivo depende diretamente de sua cobertura, de modo que a maximização da recompensa tende a melhorar o desempenho global da rede. Nesta abordagem, a função de recompensa de cada dispositivo considera o número de dispositivos vizinhos n_i , denominado grau de vizinhança, e o número de pacotes na fila do *buffer*. Esses parâmetros, em conjunto, definem o estado s . O alcance de comunicação r caracteriza a ação a . A função de recompensa do dispositivo d_i , denotada por $RW_{t,i}$ no instante t , é composta por três termos independentes [Sreedevi and Rama Rao 2019]:

$$RW_{t,i} = 1 - \omega_n \frac{n_{t,i}}{N_{max}} - \omega_r \frac{r_{t,i}}{R_{max}} - \omega_p \frac{p_{t,i}}{P_{max}}, \quad (3)$$

onde N_{max} é o grau máximo de vizinhança dos dispositivos, R_{max} é o alcance máximo de comunicação do dispositivo i , P_{max} é o número máximo de pacotes possível no *buffer* do dispositivo i , e ω_n , ω_r e ω_p são os coeficientes de ponderação do grau de vizinhança do dispositivo, do alcance de comunicação e do número de pacotes no *buffer*.

O algoritmo proposto emprega a técnica de Q-learning com política de seleção ϵ -greedy. Nessa abordagem, cada dispositivo seleciona seu alcance de comunicação considerando tanto o alcance máximo quanto o tamanho da fila do *buffer*. O número de pacotes no *buffer* é utilizado como uma métrica de latência, de modo que filas menores estão associadas a menores valores de retardo ao longo do enlace. Com probabilidade elevada ($1 - \epsilon_t$), o dispositivo seleciona a ação associada ao maior valor Q , enquanto as demais ações são escolhidas com probabilidade reduzida. O mecanismo ϵ -greedy é definido para $\epsilon \in [0, 1]$. Assim, a probabilidade $\pi(s, a)$ de selecionar uma ação a em um estado s é dada por [Sreedevi and Rama Rao 2019]:

$$\pi(s, a) = \begin{cases} 1 - \frac{\epsilon}{|A(s)|}, & \text{se } a = a^*, \text{ onde } a^* = \max_{A(s)} Q(s, a), \\ \frac{\epsilon}{|A(s)|}, & \text{em outro caso,} \end{cases} \quad (4)$$

onde o valor de Q é definido como $Q(s, a) = R(s, a) + \gamma \max_{A(s)} Q(s', a)$, sendo $\gamma = 0,5$ o fator de desconto adotado para proporcionar melhor adaptabilidade a ambientes não estacionários. O valor de Q corresponde à soma da recompensa imediata associada à ação a no estado s e do maior valor esperado de Q no estado subsequente s' , considerando todas as ações possíveis em $A(s)$, o conjunto de ações possíveis no estado s [Sreedevi and Rama Rao 2019].

A rede de comunicação D2D considerada é composta por M dispositivos comunicantes. Cada dispositivo transmite informações relacionadas à potência de transmissão, ao alcance máximo de comunicação e ao número de pacotes em sua fila de *buffer*, além de coletar informações de seus dispositivos vizinhos. Com base nesses dados, cada dispositivo executa o algoritmo de aprendizagem por reforço e decide o link de comunicação D2D que maximize a cobertura, levando em consideração o alcance de comunicação e o estado do *buffer*.

B. Algoritmo Geral para Alocar Recursos para CUEs e DUEs

Com o objetivo de sintetizar o funcionamento do algoritmo PDO-URA-RL, que considera simultaneamente as comunicações de CUEs e DUEs em uma rede CP-OFDM, apresenta-se o pseudo-código descrito no **Algoritmo 1**. Esse algoritmo fornece soluções que visam maximizar a taxa de transferência e minimizar o retardo dos usuários, decompondo o problema em etapas de otimização distintas. Inicialmente, a potência é alocada aos usuários com o objetivo de minimizar a taxa de erro de bits (*bit error rate* - BER) do enlace de transmissão. Em seguida, o número de blocos de recursos, correspondente à taxa de transmissão dos CUEs, é determinado considerando a minimização do retardo. Por fim, os blocos de recursos não alocados aos CUEs são distribuídos de forma otimizada aos DUEs, empregando a abordagem de aprendizagem por reforço descrita anteriormente em conjunto com o algoritmo de grafos de conflito (**Algoritmo DMCG** [Ferreira et al. 2022]) para minimização do retardo.

Algoritmo 1: Algoritmo de Alocação de Recursos no Uplink Baseado na Otimização de Potência e Retardo e na Aprendizagem por Reforço (PDO-URA-RL)

- 1 Receba os parâmetros iniciais: M CUEs, N pares de DUEs, K RBs ociosos e $M - K$ CUEs reais;
 - 2 Para calcular a taxa para os CUEs a fim de minimizar o retardo e otimizar a alocação de potência:
 /* Executa **Algoritmo de Otimização de Potência e Retardo** [Ferreira et al. 2022] */
 - 3 Para alocar recursos ociosos para os DUEs considerando a minimização de retardo e a aprendizagem por reforço:
 /* Executa **Algoritmo DMCG** [Ferreira et al. 2022] para formar os grafos de conflito */
 /* Forma os links D2D com base na função de recompensa (3), conforme o método descrito na Seção III-A */
-

IV. SIMULAÇÕES E RESULTADOS

Esta seção apresenta o modelo de canal adotado, os parâmetros do sistema sem fio e os resultados numéricos obtidos por meio de simulações computacionais. O desempenho do algoritmo proposto é comparado aos algoritmos DMCG [Ferreira and Vieira 2021] e GTM+ [Ciou et al. 2015], devido à similaridade entre seus objetivos, especialmente em relação à minimização do retardo em comunicações D2D.

Os cenários simulados consideram propagação por ondas milimétricas, característica das redes 5G, utilizando níveis de potência em conformidade com os limites estabelecidos pela ICNIRP (*International Commission on Non-Ionizing Radiation Protection*) [Karipidis et al. 2021]. O modelo de canal CDL-D foi empregado para representar os efeitos de múltiplos

percursos, enquanto a frequência portadora de 26 GHz foi adotada com base em estudos recentes da ANATEL (Agência Nacional de Telecomunicações). As rotinas de simulação foram implementadas manualmente em MATLAB versão R2025b (licença número 41288152), permitindo maior controle sobre os parâmetros do sistema e da modelagem do canal.

Para representar o tráfego de dados, foram utilizadas cinco séries reais de tráfego TCP/IP (*transmission control protocol/internet protocol*) provenientes da Universidade de Waikato [wai], agregadas em intervalos de 1 ms e distribuídas aleatoriamente entre os usuários. As simulações foram realizadas em um sistema *uplink* CP-OFDM com numerologia $\mu = 3$, prefixo cíclico normal, duração de *slot* de 125 μ s e espaçamento entre subportadoras de 120 kHz, conforme as especificações da 3GPP [3gp 2017]. Os parâmetros e condições de canal adotados, apresentados na Tabela I, foram definidos para representar cenários realistas de propagação em ondas milimétricas, seguindo recomendações de estudos anteriores [3gp 2017], [Sun et al. 2013], [Ciou et al. 2015].

TABELA I
PARÂMETROS DE SIMULAÇÃO

Modelo multipercurso	CDL-D
Cenário	UMa
Potência de transmissão do CUE	23 dBm
Potência de transmissão do DUE	10 dBm
Requisito de SINR para cada CUE	7 dB
Requisito de SINR para cada par DUE	4.7 dB
Raio de cobertura do eNodeB	500 m
Distância entre cada par DUE	15 m
Densidade de potência do ruído branco	-174 dBm/Hz
Modelo de perda de percurso para CUE e DUE	$128.1 + 37.6 \log_{10}(d_{dis}[km])$
Modelo de perda de percurso para pares DUE	$148 + 40 \log_{10}(d_{dis}[km])$
Largura de banda do sistema	396 MHz
Número de RBs	275
Frequência da portadora	26 GHz
Número de CUEs reais	40
Número de CUEs	No intervalo entre 50 e 100
Número de pares DUE	160
Número de TTIs	1000
Duração do <i>slot</i>	125 μ s

A vazão total do sistema, correspondente à soma das taxas efetivas de todos os dispositivos, é apresentada na Figura 2. O algoritmo proposto alcança os maiores valores de vazão, superando o DMCG e o GTM+, devido à combinação da estratégia de grafos de conflito com a minimização do retardo. Embora todos os algoritmos apresentem a mesma ordem de complexidade computacional, $O(n^4)$, sendo n o número de dispositivos da rede, observa-se nas simulações um pequeno aumento no tempo de processamento do algoritmo proposto em razão da execução do algoritmo de aprendizagem por reforço. Além disso, sua vazão aumenta com o número de CUEs, evidenciando a eficiência da alocação otimizada de recursos para dispositivos D2D.

A Figura 3 demonstra que o algoritmo proposto apresenta os menores valores de retardo médio entre os métodos avaliados, devido ao uso de aprendizagem por reforço na formação dos

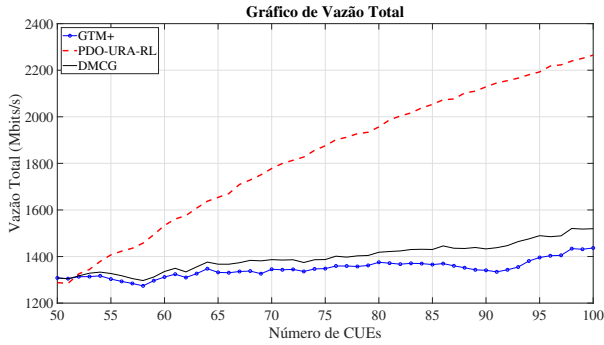


Fig. 2

VAZÃO TOTAL PARA DIFERENTES NÚMEROS DE CUEs.

links D2D e à consideração do retardo estimado como função de utilidade na alocação de recursos. O desempenho superior também decorre da eficiente alocação de potência voltada à maximização do *goodput*, resultando em maior vazão total do sistema. Além disso, observa-se redução do retardo médio com o aumento do número de CUEs, em razão da maior disponibilidade de recursos ociosos reutilizados pelos pares D2D, indicando que o sistema opera fora da condição de saturação e utiliza de forma mais eficiente os recursos de *uplink*.

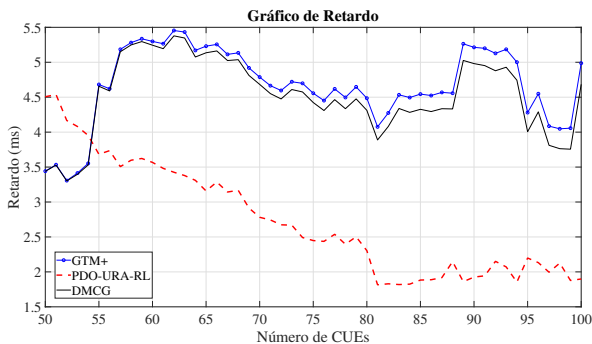


Fig. 3

RETARDO MÉDIO PARA DIFERENTES NÚMEROS DE CUEs.

V. CONCLUSÕES

Este artigo apresentou o algoritmo PDO-URA-RL para alocação de recursos no *uplink* em redes D2D, baseado na otimização conjunta de potência, retardo e aprendizagem por reforço. O algoritmo é composto por duas etapas: na primeira, realiza a alocação de recursos aos CUEs por meio da otimização de potência e retardo; na segunda, aplica o algoritmo DMCG em conjunto com aprendizagem por reforço para alocar os recursos ociosos aos dispositivos D2D. A avaliação foi realizada em um cenário CP-OFDM com propagação por ondas milimétricas, explorando a maior disponibilidade de largura de banda característica das redes 5G.

Os resultados obtidos demonstram que o algoritmo de alocação proposto alcança maiores valores de taxa de vazão e

menores valores de retardo quando comparado aos algoritmos GTM+ e DMCG, especialmente em cenários com maior número de CUEs. As simulações computacionais confirmaram que a alocação otimizada de potência e taxa para os usuários CUEs, aliada ao uso de aprendizagem por reforço e às informações de retardo estimado para a identificação de grafos de conflito e conjuntos de peso máximo independente, proporciona ganhos significativos no desempenho da alocação de recursos. Tais ganhos refletem-se simultaneamente no aumento da vazão total do sistema e na redução do retardo, em cenários de comunicação D2D baseados em ondas milimétricas.

VI. AGRADECIMENTOS

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

REFERÊNCIAS

- [wai] University of waikato - wand network research group: Series of real tcp/ip traffic. <http://wand.net.nz/wits/waikato/8/>. Jun 24, 2021.
- [3gp 2017] (2017). 3gpp tr 38.901 version 14.0.0 release 14. 5g; study on channel model for frequencies from 0.5 to 100 ghz.
- [Attar et al. 2025] Attar, I. S., Mahyuddin, N. M., and Hindia, M. H. D. N. (2025). A semi-distributed scheme for mode selection and resource allocation in device-to-device-enabled cellular networks using matching game and reinforcement learning. *Telecom*, 6(1).
- [Belgiovine et al. 2021] Belgiovine, M., Sankhe, K., Bocanegra, C., Roy, D., and Chowdhury, K. R. (2021). Deep learning at the edge for channel estimation in beyond-5g massive mimo. *IEEE Wireless Communications*, 28(2):19–25.
- [Ciou et al. 2015] Ciou, S., Kao, J., Lee, C. Y., and Chen, K. (2015). Multi-sharing resource allocation for device-to-device communication underlying 5g mobile networks. In *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1509–1514.
- [Ferreira et al. 2022] Ferreira, M. V., Vieira, F. H., and Álisson A. Cardoso (2022). Power and delay optimization based uplink resource allocation for wireless networks with device-to-device communications. *Computer Communications*, 194:226–239.
- [Ferreira and Vieira 2021] Ferreira, M. V. G. and Vieira, F. H. T. (2021). Delay minimization based uplink resource allocation for device-to-device communications considering mmwave propagation. *PeerJ Computer Science*, 7:e462.
- [Karipidis et al. 2021] Karipidis, K., Mate, R., Urban, D., Tinker, R., and Wood, A. (2021). 5g mobile networks and health - a state-of-the-science review of the research into low-level rf fields above 6 ghz. *Journal of Exposure Science and Environmental Epidemiology*, 31:585–605.
- [Mangal et al. 2025] Mangal, A., Potnis, A., Rizvi, M., Sahu, R., Singh, M., and Rathod, P. (2025). Reinforcement learning for 5g d2d communication in proximity-based networks. page 7.
- [Nagarajan and Mohamad 1992] Nagarajan, R. and Mohamad, N. M. V. (1992). Q-learning. *Machine Learning*, 8:279–292.
- [Sreedevi and Rama Rao 2019] Sreedevi, A. and Rama Rao, T. (2019). Reinforcement learning algorithm for 5g indoor device-to-device communications. *Transactions on Emerging Telecommunications Technologies*, 30(9):e3670.
- [Sun et al. 2013] Sun, H., Sheng, M., Wang, X., Zhang, Y., Liu, J., and Wang, K. (2013). Resource allocation for maximizing the device-to-device communications underlying lte-advanced networks. In *2013 IEEE/CIC International Conference on Communications in China - Workshops (CIC/ICCC)*, pages 60–64.
- [Sun et al. 2025] Sun, Q., Zhang, Y., Wu, H., Li, Y., and Petrosian, O. (2025). Multi type mean field reinforcement learning for optimal resource allocation in heterogeneous network. *Engineering Applications of Artificial Intelligence*, 158:111207.