

Aprendizado por Reforço Profundo para Otimização da Eficiência Espectral em Arquitetura Cell-Free com RSMA

André Santos, Albert dos Santos, Reyso Teixeira, Carlos Vitelli, André Cavalcante, João Weyl, Diego Cardoso

Resumo—A crescente densidade de usuários em redes 5G impõe desafios significativos para a alocação eficiente de recursos em sistemas sem fio. Nesse contexto, arquiteturas *Cell-Free Massive MIMO* (CF-mMIMO) combinadas com *Rate-Splitting Multiple Access* (RSMA) surgem como uma solução promissora para mitigar interferências e melhorar a eficiência espectral. No entanto, a alocação de potência nesses sistemas resulta em um problema não convexo e de alta dimensionalidade. Este trabalho propõe uma abordagem baseada em aprendizado por reforço profundo para otimizar a alocação de potência no *downlink*. Utiliza-se o algoritmo *Proximal Policy Optimization* (PPO), com uma função de recompensa projetada para maximizar a eficiência espectral privada, ao mesmo tempo em que garante estabilidade no treinamento e respeito às restrições de potência. Os resultados demonstram que o método proposto é capaz de aprender políticas eficientes, alcançando ganhos simultâneos em eficiência espectral e redução do consumo de potência.

Palavras-Chave—Cell-Free Massive MIMO, RSMA, aprendizado por reforço profundo, PPO, alocação de potência.

Abstract—The increasing density of users in 5G networks poses significant challenges for efficient resource allocation in wireless systems. In this context, *Cell-Free Massive MIMO* (CF-mMIMO) architectures combined with *Rate-Splitting Multiple Access* (RSMA) emerge as a promising solution to mitigate interference and improve spectral efficiency. However, power allocation in such systems leads to a non-convex and high-dimensional optimization problem. This work proposes a deep reinforcement learning-based approach to optimize downlink power allocation. The *Proximal Policy Optimization* (PPO) algorithm is employed, along with a reward function designed to maximize private spectral efficiency while ensuring training stability and compliance with power constraints. The results demonstrate that the proposed method is capable of learning efficient policies, achieving simultaneous gains in spectral efficiency and reduction in power consumption.

Keywords—Cell-Free Massive MIMO, RSMA, deep reinforcement learning, PPO, power allocation.

I. INTRODUÇÃO

A evolução das redes móveis de quinta geração (5G) tem imposto requisitos rigorosos em termos de eficiência espectral, conectividade e robustez à interferência. Em cenários com alta densidade de usuários e forte reutilização espectral, a interferência co-canal emerge como um dos principais fatores

limitantes do desempenho do sistema, degradando significativamente a qualidade dos enlaces, especialmente para usuários em condições de canal adversas [1].

Nesse contexto, a arquitetura CF-mMIMO tem se destacado como uma alternativa promissora ao paradigma celular convencional. Ao permitir que múltiplos pontos de acesso distribuídos atendam cooperativamente todos os usuários, essa abordagem reduz a interferência intercelular e melhora a uniformidade da taxa na área de cobertura [2], [3].

Apesar desses avanços, a interferência residual entre usuários ainda limita o desempenho em cenários práticos, especialmente sob *Channel State Information* (CSI) imperfeito. Para mitigar esse efeito, o RSMA tem sido proposto como uma estratégia eficiente de gerenciamento de interferência, ao dividir as mensagens em componentes comum e privada, proporcionando maior flexibilidade no tratamento da interferência [5], [6]. Trabalhos recentes indicam que a integração entre RSMA e CF-mMIMO pode trazer ganhos adicionais de eficiência espectral [7].

Entretanto, a alocação de potência em sistemas CF-mMIMO com RSMA resulta em um problema não convexo e de alta dimensionalidade, dificultando a obtenção de soluções ótimas por métodos tradicionais.

Nesse cenário, técnicas de *Deep Reinforcement Learning* (DRL) surgem como alternativas promissoras para o gerenciamento adaptativo de recursos, permitindo a aprendizagem de políticas diretamente da interação com o ambiente [8], [9]. No entanto, sua aplicação em sistemas CF-mMIMO com RSMA ainda é limitada.

Dessa forma, este trabalho propõe uma abordagem baseada em DRL para a alocação de potência em sistemas RSMA CF-mMIMO. O problema é formulado com o objetivo de maximizar a eficiência espectral total associada às mensagens privadas, sujeito a restrições de potência por ponto de acesso. Para viabilizar o treinamento, é adotada uma função de recompensa projetada para estabilizar o aprendizado.

II. MODELO DO SISTEMA

Considera-se uma rede CF composta por M *access points* (APs), cada um equipado com uma única antena, conectados via *fronthaul* a uma unidade central de processamento (CPU), que atendem cooperativamente K *user equipment* (UE), também com uma única antena. O canal entre a m -ésima AP e a k -ésima UE é modelado por

$$g_{mk} = \sqrt{\beta_{mk}} h_{mk}, \quad (1)$$

André Santos, Albert dos Santos, Reyso Teixeira, Carlos Vitelli, João Weyl, Diego Cardoso, Instituto de Tecnologia (ITEC), Universidade Federal do Pará (UFPA), Belém, PA, Brasil, email:[andre.santos,albert.santos,reyso.teixeira,carlos.vitelli]@itec.ufpa.br; [jweyl,diego]@ufpa.br. André Cavalcante, Ericsson Telecomunicações Ltda, Indaiatuba, SP, Brasil, email: andre.mendes.cavalcante@ericsson.com

em que β_{mk} representa o ganho de larga escala, incluindo *path loss* e *shadow fading* [2], enquanto $h_{mk} \sim \mathcal{CN}(0, 1)$ corresponde ao ganho de pequena escala.

Durante a fase de treinamento *uplink*, considera-se que todos os UEs ativos utilizam a mesma sequência piloto com potência total ρ_{tr} . Assim, a estimativa MMSE do canal é dada por [7]

$$\hat{g}_{mk} = \beta_{mk} Q_m^{-1} \left(\sum_{i=1}^K g_{mi} + \frac{1}{\sqrt{\rho_{tr}}} z_m \right), \quad (2)$$

onde $z_m \sim \mathcal{CN}(0, \sigma_{ul}^2)$ representa o ruído térmico na m-ésima AP, e

$$Q_m = \sum_{k=1}^K \beta_{mk} + \frac{\sigma_{ul}^2}{\rho_{tr}}. \quad (3)$$

A estimativa $\hat{g}_{mk}$ possui distribuição $\hat{g}_{mk} \sim \mathcal{CN}(0, \gamma_{mk})$ com

$$\gamma_{mk} = \frac{\beta_{mk}^2}{Q_m}. \quad (4)$$

A. Transmissão Downlink com RSMA

No enlace *downlink*, emprega-se o RSMA, em que a mensagem destinada a k-ésima UE, denotada por W_k , é dividida em mensagem comum $W_{c,k}$ e em mensagem privada $W_{p,k}$. A mensagem comum de todos os UEs são combinadas em uma única mensagem comum W_c , codificada no símbolo s_c , enquanto cada mensagem privada é codificada em s_k . O sinal transmitido pela m-ésima AP é dado por [7]

$$x_m = \sqrt{\rho_T \eta_{c,m}} w_{c,m} s_c + \sum_{i=1}^K \sqrt{\rho_T \eta_{mi}} w_{mi} s_i, \quad (5)$$

onde $w_{c,m} \in \mathbb{C}$ é o precoder da mensagem comum e $w_{mk} \in \mathbb{C}$ é o precoder da mensagem privada para a k-ésima UE, ρ_T é a potência total transmitida em *downlink* na m-ésima AP com $\eta_{c,m}$ e η_{mk} sendo os coeficientes de controle de potência das mensagens comum e privada, respectivamente. O sinal recebido pela k-ésima UE é expresso por [7]

$$y_k = \sum_{m=1}^M g_{mk}^* x_m + n_k, \quad (6)$$

em que $n_k \sim \mathcal{CN}(0, \sigma^2)$ representa o ruído aditivo gaussiano.

De acordo com o princípio do RSMA, cada UE inicialmente realiza a decodificação da mensagem comum x_c , tratando as mensagens privadas como interferência. Em seguida, após a remoção da mensagem comum por meio de SIC, a k-ésima UE procede à decodificação do seu respectiva mensagem privada x_k .

Considerando a ausência de transmissão de pilotos em *downlink*, considera-se que os UEs não dispõem de conhecimento instantâneo dos termos $g_{mk}^* w_{c,m}$ e $g_{mk}^* w_{k,m}$. Dessa forma, cada UE baseia sua detecção nas médias estatísticas desses termos, isto é, $\mathbb{E}\{g_{mk}^* w_{c,m}\}$ e $\mathbb{E}\{g_{mk}^* w_{k,m}\}$. O sinal recebido para a mensagem comum pode ser expresso como [7]

$$y_{c,k} = T_{c,k} s_c + B I_{c,k} s_c + \sum_{i=1}^K I_{ki} s_i + \eta_k, \quad (7)$$

onde

$$T_{c,k} = \sqrt{\rho_T} \sum_{m=1}^M \sqrt{\eta_{c,m}} \mathbb{E}\{g_{mk}^* w_{c,m}\}, \quad (8)$$

$$B I_{c,k} = \sqrt{\rho_T} \sum_{m=1}^M \sqrt{\eta_{c,m}} g_{mk}^* w_{c,m} - T_{c,k}, \quad (9)$$

$$I_{ki} = \sqrt{\rho_T} \sum_{m=1}^M \sqrt{\eta_{mi}} g_{mk}^* w_{mi}. \quad (10)$$

Após o cancelamento da mensagem comum, o sinal recebido da mensagem privada da k-ésima UE é dada por

$$y_{p,k} = T_{p,k} s_k + B I_{p,k} s_k + B I_{c,k} s_c + \sum_{i \neq k}^K I_{ki} s_i + \eta_k, \quad (11)$$

onde

$$T_{p,k} = \sqrt{\rho_T} \sum_{m=1}^M \sqrt{\eta_{mk}} \mathbb{E}\{g_{mk}^* w_{mk}\}, \quad (12)$$

$$B I_{p,k} = \sqrt{\rho_T} \sum_{m=1}^M \sqrt{\eta_{mk}} g_{mk}^* w_{mk} - T_{p,k}. \quad (13)$$

Após a decodificação sucessiva, as eficiências espectrais das mensagens comum e privada são dadas por

$$SE_{c,k} = \frac{\tau_d}{\tau} \log_2 (1 + \text{SINR}_{c,k}), \quad (14)$$

$$SE_{p,k} = \frac{\tau_d}{\tau} \log_2 (1 + \text{SINR}_{p,k}), \quad (15)$$

onde $\text{SINR}_{c,k}$ e $\text{SINR}_{p,k}$ representam a relação sinal interferência-ruído (SINR) da mensagem comum e privada, respectivamente, que são dadas pelas Equações 16 e 17.

III. PROJETO DO PRECODER

As expressões de eficiência espectral obtidas podem ser utilizadas com diferentes estratégias de precodificação para as mensagens comum e privada. Para as mensagens privadas, adota-se o *conjugate beamforming*, definido por $w_{mk} = \hat{g}_{mk}$. O projeto do precodificador da mensagem comum deve maximizar a menor mensagem comum entre todos os UEs. Assim, o problema pode ser formulado como [7]

$$\max_{\{w_{c,m}\}} \min_k \pi_k^{-1} \left| \sum_{m=1}^M \mathbb{E}\{g_{mk}^* w_{c,m}\} \right|^2 \quad (18)$$

sujeito à

$$\mathbb{E}\{|w_{c,m}|^2\} = 1, \quad \forall m, \quad (19)$$

$$\frac{\rho_T \left(\sum_{m=1}^M (\sqrt{\eta_{c,m}} \gamma_{c,mk}) \right)^2}{\rho_T \sum_{i=1}^K \sum_{m=1}^M (\sqrt{\eta_{mi}}^2 \gamma_{mi} \beta_{mk}) + \rho_T \sum_{i=1}^K \left(\sum_{m=1}^M (\sqrt{\eta_{mi}} \gamma_{mi} \frac{\beta_{mk}}{\beta_{mi}}) \right)^2 + \rho_T \sum_{m=1}^M (\sqrt{\eta_{c,m}}^2 I_{c,k}^m) + \sigma^2} \quad (16)$$

$$\frac{\rho_T \left(\sum_{m=1}^M (\sqrt{\eta_{mk}} \gamma_{mk}) \right)^2}{\rho_T \sum_{i=1}^K \sum_{m=1}^M (\sqrt{\eta_{mi}}^2 \gamma_{mi} \beta_{mk}) + \rho_T \sum_{i \neq k} \left(\sum_{m=1}^M (\sqrt{\eta_{mi}} \gamma_{mi} \frac{\beta_{mk}}{\beta_{mi}}) \right)^2 + \rho_T \sum_{m=1}^M (\sqrt{\eta_{c,m}}^2 I_{c,k}^m) + \sigma^2} \quad (17)$$

onde $\pi_k = \sum_{i=k}^K \mathbb{E}\{|I_{ki}|^2\} + \sigma^2$. Considera-se que o precodificador comum em cada AP seja uma combinação linear das estimativas de canal dos UEs [7]:

$$w_{c,m} = \sum_{k=1}^K a_{mk} \hat{g}_{mk}, \quad (20)$$

em que a_{mk} é o peso associado a k -ésima UE na m -ésima AP.

Substituindo essa expressão no problema anterior, obtém-se uma formulação equivalente para o projeto do precodificador comum:

$$\max_{\mathbf{A}} \min_k \left(\sum_{m=1}^M \sum_{i=1}^K a_{mi} \beta_{mi} Q_m^{-1} \beta_{mk} \right)^2 \quad (21)$$

sujeita à

$$\mathbb{E} \left\{ \left| \sum_{i=1}^K a_{mi} \hat{g}_{mi} \right|^2 \right\} = 1, \quad \forall m. \quad (22)$$

A matriz $\mathbf{A} \in \mathbb{R}^{M \times K}$ contém todos os coeficientes do precodificador comum. Como o problema é convexo após transformação apropriada, sua solução pode ser obtida por métodos de otimização convexa, como *interior-point methods*. Uma vez obtidos os coeficientes ótimos, o precodificador comum final é reconstruído por [7]

$$w_{c,m} = \sum_{k=1}^K a_{mk}^* \hat{g}_{mk}. \quad (23)$$

Como essa solução depende apenas dos coeficientes de larga escala β_{mk} , o precodificador pode ser reutilizado durante vários intervalos de coerência, reduzindo significativamente a complexidade computacional do sistema.

IV. FORMULAÇÃO DO PROBLEMA

Neste trabalho, a alocação de potência é formulada com base na maximização da eficiência espectral total associada às mensagens privadas em sistemas RSMA CF-mMIMO. A escolha pela componente privada se deve à sua relação direta com os coeficientes de potência e à menor complexidade de modelagem no contexto de aprendizado por reforço.

A eficiência espectral privada da k -ésima UE é definida como

$$SE_{p,k} = \log_2(1 + \text{SINR}_{p,k}), \quad (24)$$

e o problema de otimização é dado por

$$\max_{\boldsymbol{\eta}} \sum_{k=1}^K SE_{p,k} \quad (25)$$

sujeito à restrição de potência em cada ponto de acesso,

$$\sum_{k=1}^K \eta_{mk} \gamma_{mk} \leq 1, \quad \forall m. \quad (26)$$

V. APRENDIZADO POR REFORÇO

O problema é modelado como um processo de decisão de Markov (MDP), sendo implementado em *Python* com a biblioteca *Gymnasium*, enquanto o treinamento do agente é realizado com a *Stable-Baselines3*.

O ambiente segue a estrutura padrão com as funções *reset* e *step*. A função *reset* reinicializa o episódio mantendo fixos o posicionamento das APs, dos UEs e os coeficientes de canal. Essa escolha foi motivada por testes preliminares, nos quais a variação da topologia dificultou a convergência do agente devido à elevada não estacionariedade do ambiente.

A função *step* executa a ação do agente, atualizando o estado do sistema, calculando a recompensa e verificando o término do episódio.

O estado, a ação e a recompensa são definidos de modo a capturar as condições da rede e permitir que o agente aprenda uma política eficiente de alocação de potência no cenário considerado.

A. Espaço de Estados

O estado do ambiente é representado por um vetor contendo os valores instantâneos do SINR de cada UE ativo no sistema. Essa escolha permite que o agente tenha acesso direto à qualidade dos enlaces de comunicação, refletindo as condições atuais da rede e possibilitando decisões mais informadas.

B. Espaço de Ações

As ações são definidas como um vetor contínuo de dimensão $K \times M$, onde K corresponde ao número de UE e M ao número de AP. Cada elemento da ação representa o nível relativo de potência associado à transmissão entre um determinado par AP-UE.

Devido à parametrização adotada pelos algoritmos de aprendizado profundo, os valores das ações são inicialmente gerados no intervalo contínuo $[-1, 1]$. Esses valores são posteriormente transformados em coeficientes de alocação de potência,

levando em consideração as condições de canal. Assim, a potência efetivamente alocada é dada por

$$\eta_{mk} = \frac{a_{mk}}{\gamma_{mk}}, \quad (27)$$

em que a_{mk} representa a ação produzida pelo agente e γ_{mk} corresponde ao ganho de canal estimado entre a m -ésima AP e a k -ésima UE.

C. Função de Recompensa

Embora o objetivo do problema seja a maximização da eficiência espectral total, a função de recompensa utilizada no treinamento do agente é uma versão modificada desse objetivo, projetada para melhorar a estabilidade e a convergência do processo de aprendizado.

Quando as restrições de potência são satisfeitas, isto é,

$$\sum_{k=1}^K \eta_{mk} \gamma_{mk} \leq 1, \quad \forall m, \quad (28)$$

a recompensa no instante t é definida como

$$r_t = \begin{cases} 1 + \overline{SE}_t, & \text{se } Q_{0.05}^{(t)} \geq Q_{0.05}^{(t-1)}, \\ -1, & \text{caso contrário,} \end{cases} \quad (29)$$

onde $\overline{SE}_t$ representa a eficiência espectral média e $Q_{0.05}^{(t)}$ o quinto percentil da distribuição de eficiência espectral.

O uso da média $\overline{SE}_t$ em vez da soma da $SE_{p,k}$ tem como objetivo normalizar a escala da recompensa em relação ao número de UEs, evitando valores excessivamente elevados que dificultam a convergência do agente.

O termo constante $1+$ é introduzido para garantir recompensas positivas durante o treinamento, favorecendo a estabilidade dos gradientes no algoritmo de aprendizado.

Já o critério baseado no percentil atua como um mecanismo de regularização temporal, impedindo que o agente degrade significativamente o desempenho dos UEs em piores condições entre passos consecutivos. Ressalta-se que esse termo não altera o objetivo de otimização, mas apenas guia o processo de aprendizado.

Quando as restrições de potência são violadas, a recompensa é definida como

$$r_t = -\mathcal{P}_t = \sum_{m=1}^M \max \left(0, \sum_{k=1}^K \eta_{mk} \gamma_{mk} - 1 \right), \quad (30)$$

quantificando o excesso de potência transmitida.

D. Algoritmo de Aprendizado

Para a resolução do problema, foi empregado o algoritmo *Proximal Policy Optimization* (PPO), amplamente utilizado em cenários com espaços de ação contínuos. O PPO pertence à classe de métodos *actor-critic*, nos quais duas redes neurais são treinadas de forma conjunta.

A rede *actor* é responsável por parametrizar a política estocástica, mapeando estados em distribuições de probabilidade sobre ações. Já a rede do crítico estima a função de valor associada aos estados, fornecendo uma medida de qualidade das decisões tomadas.

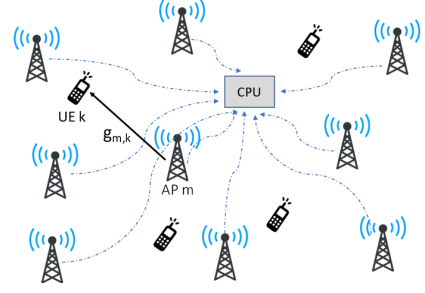


Fig. 1. Sistema Cell-Free Massive MIMO (retirado de [11])

Uma característica importante do PPO é a utilização de um mecanismo de atualização com restrição de proximidade, que evita mudanças abruptas na política durante o treinamento. Isso contribui para maior estabilidade e eficiência no processo de aprendizado. No contexto considerado, o agente atua como um controlador centralizado, responsável por determinar simultaneamente a alocação de potência em todos os enlaces AP-UE do sistema.

VI. RESULTADOS

Para os resultados, foi considerado a rede com $M = 16$ e $K = 4$, onde cada AP e UE são distribuídos de forma aleatória em um cenário com tamanho 250×250 metros, representado na Fig. 1. Os parâmetros utilizados são descritos na Tabela I, em concordância com os valores utilizados em [2] [7]

TABELA I
PARÂMETROS DO CENÁRIO.

Parâmetros	Valor
Frequência da portadora	1,9 GHz
Largura de banda	20 MHz
Altura da AP (h_{AP})	10 m
Altura da UE (h_u)	1,65 m
Potência do ruído (σ^2)	-94 dBm
Distâncias de referências (d_1), (d_0)	50 m, 10 m
Potências (ρ_{tr}), (ρ_T)	10 dBm, 20 dBm

Para verificar o desempenho do algoritmo PPO, foi realizada uma comparação com uma abordagem do *baseline*, na qual a alocação de potência prioriza as UEs com os melhores canais (*baseline* canal-forte). O treinamento do agente PPO foi conduzido utilizando 22 milhões de passos, devido a convergência do algoritmo, por meio da biblioteca SB3. Foram mantidos os hiperparâmetros padrão do algoritmo [12], com exceção do parâmetro *batch_size*, cujo valor padrão de 64 foi aumentado para 256. Essa modificação foi adotada com o objetivo de melhorar a estabilidade do treinamento, uma vez que lotes maiores proporcionam estimativas de gradiente menos ruidosas.

A Fig. 2 apresenta a função de distribuição acumulada (CDF) da eficiência espectral por UE para o esquema RSMA em 10 cenários distintos, comparando o método *baseline* com o PPO, tanto para a SE total quanto para a mensagem privada. Observa-se que as curvas do PPO estão, de forma geral,

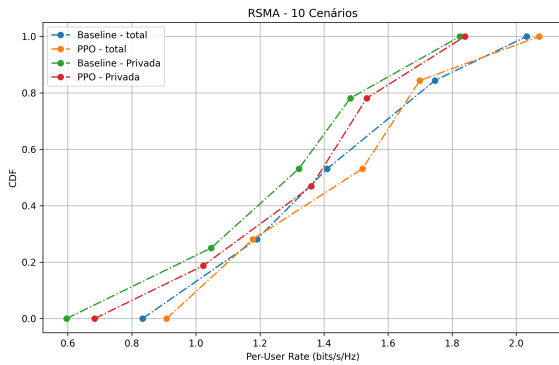


Fig. 2. CDF da eficiência espectral por UE

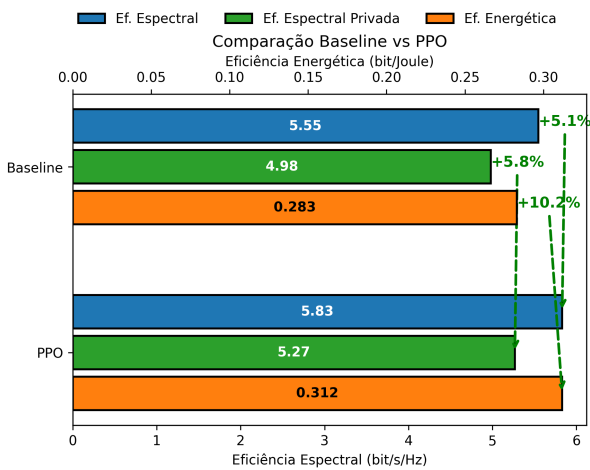


Fig. 3. Gráfico de comparação entre baseline e PPO

deslocadas para a direita, indicando melhoria consistente na distribuição da eficiência espectral entre os UEs.

Para a SE total, os ganhos são mais evidentes nas regiões intermediárias da CDF (entre 0,3 e 0,8), sugerindo aumento da eficiência espectral típica do sistema, enquanto os valores máximos permanecem semelhantes. No caso da mensagem privada, o deslocamento é ainda mais pronunciado em algumas regiões, indicando maior eficácia do PPO na alocação de potência. Além disso, a redução na probabilidade de baixas eficiências espectrais evidencia melhoria na equidade, beneficiando UEs em condições desfavoráveis.

A Fig. 3 compara o baseline e o PPO em termos de eficiência espectral total, eficiência espectral privada e eficiência energética, calculada conforme a definição apresentada em [13]. O método proposto supera o baseline em todas as métricas, com ganhos de aproximadamente 5,1% na SE total, 5,8% na SE privada e 10,2% na eficiência energética. Esses resultados indicam que o PPO é capaz de aprender políticas mais eficientes, melhorando simultaneamente o desempenho espectral e o uso de potência.

VII. CONCLUSÕES

Este trabalho propôs uma abordagem baseada em DRL, utilizando o algoritmo PPO, para otimizar a alocação de potência em *downlink* de sistemas RSMA CF-mMIMO. Através do projeto de uma função de recompensa adaptada, o agente foi capaz de aprender políticas eficientes que respeitam as restrições de potência de cada AP, contornando a complexidade de otimização tradicional desses sistemas. Os resultados demonstraram que o método proposto superou de forma consistente o *baseline*. A política aprendida pelo PPO demonstrou ganhos notáveis na eficiência espectral, englobando tanto a SE total quanto as mensagens privadas. Adicionalmente, a formulação resultou em um gerenciamento de recursos mais inteligente, o que se traduziu em uma expressiva melhoria na eficiência energética global do sistema. Como trabalhos futuros, pretende-se investigar arquiteturas de Multi-Agente DRL para contornar problemas de escalabilidade em cenários com um número massivo de APs e UEs.

REFERÊNCIAS

- [1] Ericsson, "Ericsson Mobility Report," 2024. [Online]. Disponível em: <https://www.ericsson.com/en/mobility-report>
- [2] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson and T. L. Marzetta, "Cell-Free Massive MIMO Versus Small Cells," *IEEE Transactions on Wireless Communications*, v. 16, n. 3, pp. 1834–1850, Mar. 2017.
- [3] Ö. T. Demir, E. Björnson and L. Sanguinetti, *Foundations of User-Centric Cell-Free Massive MIMO*. Hanover, MA, USA: Now Publishers, 2021.
- [4] E. Björnson and L. Sanguinetti, "Scalable Cell-Free Massive MIMO Systems," *IEEE Transactions on Communications*, v. 68, n. 7, pp. 4247–4261, Jul. 2020.
- [5] Y. Mao, B. Clerckx and V. O. K. Li, "Rate-Splitting Multiple Access for Downlink Communication Systems: Bridging, Generalizing, and Outperforming SDMA and NOMA," *EURASIP Journal on Wireless Communications and Networking*, v. 2018, n. 1, pp. 1–54, 2018.
- [6] B. Clerckx, Y. Mao, R. Schober and H. V. Poor, "Rate-Splitting Multiple Access: A New Frontier for the PHY Layer of 6G," *IEEE Communications Magazine*, v. 59, n. 2, pp. 124–130, Feb. 2021.
- [7] A. Mishra, E. Björnson and L. Sanguinetti, "Rate-Splitting Assisted Massive Machine-Type Communications in Cell-Free Massive MIMO," *IEEE Transactions on Wireless Communications*, v. 21, n. 2, pp. 1350–1364, Feb. 2022.
- [8] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang and Y.-C. Liang, "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Communications Surveys & Tutorials*, v. 21, n. 4, pp. 3133–3174, 2019.
- [9] N. Zhao, Y. Chen, R. Zhang and Y. Zou, "Deep Reinforcement Learning for User Association and Resource Allocation in Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, v. 20, n. 2, pp. 1090–1105, Feb. 2021.
- [10] A. Nasir, H. Tabassum and E. Hossain, "Multi-Agent Deep Reinforcement Learning for Resource Allocation in Wireless Networks," *IEEE Transactions on Wireless Communications*, v. 20, n. 7, pp. 4459–4474, Jul. 2021.
- [11] Z. Chen and E. Björnson, "Channel Hardening and Favorable Propagation in Cell-Free Massive MIMO With Stochastic Geometry," *IEEE Transactions on Communications*, vol. 66, no. 11, pp. 5205–5219, Nov. 2018.
- [12] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. "Proximal policy optimization algorithms," *arXiv preprint arXiv*, Jul. 2017
- [13] Q. Zheng, P. Zhu, J. Li, D. Wang and X. You, "On the Spectral and Energy Efficiency of RSMA-Based Cell-Free Systems," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 11, pp. 17726–17731, Nov. 2024