

Otimização da Eficiência Espectral em Arquitetura Cell-Free com Deep Reinforcement Learning

André Santos, Albert dos Santos, Reyso Teixeira, Matheus Pantoja, André Cavalcante, João Weyl, Diego Cardoso

Resumo— Com o crescimento exponencial de dispositivos conectados e a complexidade crescente das redes sem fio, sistemas *Cell-Free Massive Multiple-Input Multiple-Output* surgem como solução promissora para oferecer cobertura uniforme e eficiente. No entanto, a alocação de potência em ambientes densos impõe desafios de escalabilidade e justiça entre usuários. Este trabalho propõe uma abordagem baseada em aprendizado por reforço profundo, utilizando o algoritmo *Proximal Policy Optimization* para alocação de potência. O método visa otimizar o problema do max-min em relação à SE dos usuários, garantindo justiça max-min e adaptabilidade dinâmica, mesmo em cenários com elevado número de usuários e pontos de acesso.

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

Abstract— With the exponential growth of connected devices and the increasing complexity of wireless networks, Cell-Free Massive Multiple-Input Multiple-Output systems have emerged as a promising solution to provide uniform and efficient coverage. However, power allocation in dense environments poses challenges in terms of scalability and fairness among users. This work proposes a deep reinforcement learning approach using the Proximal Policy Optimization algorithm for power allocation. The method aims to optimize the max-min fairness problem with respect to users's spectral efficiency, ensuring both fairness and dynamic adaptability, even in scenarios with a high number of users and access points.

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

I. INTRODUÇÃO

A arquitetura *Cell-Free* (CF) *Massive Multiple-Input Multiple-Output* (MIMO) tem se destacado como uma solução promissora para as redes sem fio de próxima geração, proporcionando comunicação com alta confiabilidade e desempenho uniforme entre os usuários [1]. Nessa abordagem, um grande número de pontos de acesso (APs) cooperam de forma distribuída para atender simultaneamente os usuários (UEs), compartilhando os mesmos recursos tempo-frequência por meio de duplexação por divisão no tempo (TDD) [2]. A ausência de células fixas e fronteiras rígidas oferece vantagens significativas, como redução da interferência entre usuários e maior eficiência espectral (*Spectral Efficiency* – SE).

A escalabilidade de sistemas CF *Massive MIMO* enfrenta desafios significativos, especialmente devido à crescente complexidade computacional exigida pelo processamento de sinais, alocação de recursos e demanda por *fronthaul* conforme

aumenta o número de APs e UEs. Essa complexidade tende a crescer de forma polinomial, tornando a implementação prática mais desafiadora em redes de grande porte. Além disso, com o número de dispositivos conectados aumentando rapidamente, a estratégia de elevar a potência de transmissão para melhorar a SE mostra-se inviável a longo prazo, considerando os impactos econômicos e operacionais [3].

Nos últimos anos, métodos baseados em aprendizado profundo (*Deep Learning* – DL) têm sido propostos como alternativas promissoras para acelerar a alocação de recursos em sistemas sem fio. Contudo, a maioria dessas abordagens fundamenta-se em aprendizado supervisionado, o que requer conjuntos de dados extensos previamente gerados a partir de soluções ótimas obtidas por algoritmos clássicos de otimização [4]. Essa exigência não apenas impõe um custo computacional elevado na fase inicial, como também compromete a adaptabilidade dos modelos a ambientes dinâmicos, sujeitos a variações no canal ou nas demandas dos usuários.

Diante dessas limitações, métodos baseados em *Reinforcement Learning* (RL), especialmente *Deep Reinforcement Learning* (DRL), surgem como alternativas promissoras para alocação de recursos em sistemas *Massive MIMO*. O DRL combina a rápida convergência do DL com a capacidade adaptativa do RL, o que é crucial para decisões dentro do tempo de coerência do canal, geralmente na ordem de milissegundos. Além disso, em cenários de alta mobilidade, o DRL mostra-se robusto frente ao *channel aging* e ao conhecimento imperfeito do canal (*imperfect CSIT*), dispensando modelagens detalhadas da topologia da rede [5].

Assim, este trabalho tem como objetivo otimizar o problema do max-min da SE, abordando um dos principais desafios na alocação de potência em transmissões *downlink*. Para isso, propõe-se a utilização de algoritmos de DRL, em especial o *Proximal Policy Optimization* (PPO), para realizar a alocação de potência de forma distribuída e eficiente em sistemas CF *Massive MIMO*. A abordagem emprega ações contínuas, que representam o percentual de potência de cada AP alocado para determinado UE, permitindo um controle mais preciso da alocação de potência e contribuindo para a otimização do problema de max-min da SE, respeitando as restrições de potência de cada AP e garantindo a escalabilidade do sistema mesmo em cenários com elevado número de usuários e pontos de acesso.

II. TRABALHOS CORRELATOS

Nos últimos anos, a alocação de potência em arquiteturas CF *Massive MIMO* tem sido amplamente investigada por

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

meio de diferentes abordagens, especialmente aquelas baseadas em otimização convexa e algoritmos heurísticos. No entanto, essas técnicas geralmente sofrem com limitações de escalabilidade e adaptabilidade em cenários dinâmicos. Diante disso, abordagens baseadas em RL vêm ganhando destaque como alternativas promissoras.

Em [6], os autores investigam três algoritmos baseados em aprendizado por reforço — *Q-learning*, SARSA e eSARSA — aplicados ao problema de alocação de potência com base na métrica *max-product SINR*. Notavelmente, os algoritmos foram capazes de aprender políticas eficazes considerando apenas as distâncias entre os usuários e os pontos de acesso, sem a necessidade de informações detalhadas do canal. [11] investigaram o uso de dois métodos de alocação de potência baseados em DRL, o *Deep Q-Network* (DQN) e o *Deep Deterministic Policy Gradient* (DDPG), em sistemas CF Massive MIMO operando em ondas milimétricas com UEs móveis. O objetivo foi maximizar a sum-SE no *downlink*, considerando imperfeições na estimativa do canal (CSI) e contaminação dos pilotos no *uplink*. Os resultados mostraram que ambos os algoritmos alcançaram desempenho competitivo em relação ao tradicional algoritmo WMMSE, superando-o em até 38% na soma da SE em configurações específicas e mantendo uma média de 33% superior em diferentes cenários.

Em [7], é investigada a alocação de potência em sistemas CF Massive MIMO utilizando aprendizado por reforço profundo com o algoritmo DDPG. Inicialmente, os autores aplicam o DDPG ao problema de controle de potência no *downlink* com critério de equidade max-min. Em seguida, o estudo é estendido para duas outras métricas de desempenho amplamente utilizadas: *sum-rate* e *max-product SINR*. Os resultados mostram que o DDPG é eficaz nas três abordagens, superando tanto técnicas baseadas em aprendizado profundo supervisionado quanto métodos de otimização convexa. Para o cenário max-min, o DDPG alcança uma solução próxima da ótima obtida por *solvers* de otimização convexa. Nos casos de *sum-rate* e *max-product SINR*, o algoritmo apresentou ganhos de desempenho de 10,3% e 30,2%, respectivamente, em comparação com os métodos supervisionados.

Desta forma, estratégias de alocação de potência baseadas em RL e DRL têm se destacado como alternativas promissoras aos métodos tradicionais de otimização, especialmente em cenários dinâmicos e de grande escala, razão pela qual foram adotadas como foco deste trabalho.

III. MODELO DO SISTEMA

Para o sistema considera-se uma rede CF Massive MIMO com M APs e K UEs equipados com apenas uma antena, onde todas as APs estão ligadas a uma unidade central de processamento (CPU) através de um *fronthaul*. Todos os K UEs são atendidos simultaneamente por todos os M APs dentro da cobertura da rede. O canal é modelado como [1]:

$$g_{mk} = \beta_{mk}^{\frac{1}{2}} h_{mk}, \quad (1)$$

onde h_{mk} representa o ganho de pequena escala e β_{mk} representa o ganho de larga escala. Considera-se que h_{mk} são variáveis aleatórias independentes e identicamente distribuídos

com distribuição $\mathcal{CN}(0, 1)$. [8] modela o *path loss* e o *shadow fading* de β_{mk} como

$$\beta_{mk} = PL_{mk} 10^{\frac{\sigma_{sh} z_{mk}}{10}}, \quad (2)$$

onde PL_{mk} é o *path loss*, e $10^{\frac{\sigma_{sh} z_{mk}}{10}}$ representa o *shadow fading* com desvio padrão σ_{sh} e z_{mk} segue uma distribuição $\mathcal{N}(0, 1)$.

O *path loss* PL_{mk} é modelado utilizando o modelo de *three-slope* descrito em [9]. O expoente de *path loss* assume o valor 3,5 quando a distância entre a m -ésima AP e a k -ésima UE, denotada por d_{mk} , é maior que d_1 ; considera-se o valor 2 quando $d_1 \geq d_{mk} > d_0$; e é igual a 0 quando $d_{mk} \leq d_0$, para certos valores de d_0 e d_1 . Quando $d_{mk} > d_1$, utiliza-se o modelo de propagação Hata-COST231. O *path loss* é modelado como, em dB

$$PL_{mk} = \begin{cases} -L - 35\log_{10}(d_{mk}), & \text{se } d_{mk} > d_1 \\ -L - 15\log_{10}(d_1) - 20\log_{10}(d_{mk}), & \text{se } d_0 < d_{mk} \leq d_1 \\ -L - 15\log_{10}(d_1) - 20\log_{10}(d_0), & \text{se } d_{mk} \leq d_0 \end{cases} \quad (3)$$

onde

$$L \triangleq 46,3 + 33,9\log_{10}(f) - 13,82\log_{10}(h_{AP}) - (1,1\log_{10}(f) - 0,7)h_u + (1,56\log_{10}(f) - 0,8), \quad (4)$$

onde f é a frequência da onda portadora (em MHz), h_{AP} a altura da antena da AP (em metros) e h_u sendo a altura da antena da UE (em metros).

Os coeficientes do *shadow fading* são usados com o modelo com dois componentes [10]

$$z_{mk} = \sqrt{\delta a_m} + \sqrt{1 - \delta} b_k, \quad (5)$$

onde a_m e b_k seguem uma distribuição $\mathcal{N}(0, 1)$ com variáveis aleatórias independentes. A variável a_m representa os efeitos de *shadow fading* causados por obstáculos nas proximidades da m -ésima AP, afetando de forma semelhante todos os usuários atendidos por essa AP. Já a variável b_k representa os efeitos de *shadow fading* causados por objetos próximos ao k -ésimo usuário, impactando de maneira igual os canais entre esse usuário e todas as APs. O parâmetro $\delta \in [0, 1]$ controla o balanço entre essas duas fontes de *shadow fading*: Quando $\delta = 0$, o *shadow fading* é específico de cada usuário (mesmo para diferentes APs). Quando $\delta = 1$, o *shadow fading* é específico de cada AP (mesmo para diferentes usuários). Valores intermediários de δ representam uma combinação entre essas duas situações.

A. Treinamento do Uplink

Para estimar o canal, durante a fase de treinamento as UEs transmitem sequências de pilotos simultaneamente, onde cada sequência de piloto φ_k possui comprimento τ_p (em amostras) para cada k -ésima UE. Sendo $\sqrt{\tau_p} \varphi_k \in \mathbb{C}^{\tau_p \times 1}$, onde $\|\varphi_k\|^2 = 1$. Por isso, o $\tau_p \times 1$ vetor do sinal piloto recebido para a m -ésima AP é dado por [1]

$$y_{p,m} = \sqrt{\tau_p} \rho_p \sum_{k=1}^K g_{mk} \varphi_k + \mathbf{w}_{p,m}, \quad (6)$$

$$SINR_k = \frac{\rho_d (\sum_{m=1}^M \eta_{mk}^{1/2} \gamma_{mk})^2}{\rho_d \sum_{k' \neq k}^K (\sum_{m=1}^M \eta_{mk'}^{1/2} \gamma_{mk'} \frac{\beta_{mk}}{\beta_{mk'}})^2 |\varphi_k^H \varphi_{k'}|^2 + \rho_d \sum_{k'=1}^K \sum_{m=1}^M \eta_{mk'} \gamma_{mk'} \beta_{mk'} + 1} \quad (10)$$

onde ρ_p é a razão sinal-ruído (SNR) normalizado de cada símbolo piloto e $\mathbf{w}_{p,m}$ é o vetor do ruído aditivo na m-ésima AP, sendo $\mathbf{w}_{p,m}$ variáveis aleatórias independentes e identicamente distribuídos com distribuição $\mathcal{CN}(0, 1)$. O canal estimado g_{mk} da m-ésima AP é baseado a partir do sinal do piloto recebido $y_{p,m}$, assim denotando a projeção de $y_{p,m}$ em φ_k^H , como demonstrado a seguir [1]:

$$\check{y}_{p,mk} = \sqrt{\tau_p \rho_p} g_{mk} + \sqrt{\tau_p \rho_p} \sum_{k' \neq k}^K g_{mk'} \varphi_k^H \varphi_{k'} + \varphi_k^H \mathbf{w}_{p,m}. \quad (7)$$

Em questão, para sequências de pilotos, $\check{y}_{p,mk}$ não constitui uma estatística suficiente para a estimativa do canal g_{mk} , sendo possível apenas obter estimativas subótimas. Por outro lado, quando duas sequências de pilotos são idênticas ou ortogonais, logo $\check{y}_{p,mk}$ torna-se uma estatística suficiente, permitindo estimativas ótimas por meio do estimador MMSE [1]

$$\hat{g}_{mk} = \frac{\mathbb{E} \{ \check{y}_{p,mk} g_{mk} \}}{\mathbb{E} \{ |\check{y}_{p,mk}|^2 \}} \check{y}_{p,mk} = c_{mk} \check{y}_{p,mk}, \quad (8)$$

onde

$$c_{mk} \triangleq \frac{\sqrt{\tau_p \rho_p} \beta_{mk}}{\tau_p \rho_p \sum_{k'=1}^K \beta_{mk'} |\varphi_k^H \varphi_{k'}|^2 + 1}. \quad (9)$$

B. Transmissão dos Dados em Downlink

Cada AP considera o canal estimado como se fosse o canal verdadeiro e aplica *beamforming* conjugado para transmitir sinais à k-ésima UE. Assim, o sinal transmitido pela m-ésima AP é dado por [1]:

$$x_m = \sqrt{\rho_d} \sum_{k=1}^K \eta_{mk}^{1/2} \hat{g}_{mk}^* q_k, \quad (11)$$

onde q_k é o símbolo associado para a k-ésima UE, que satisfaz $\mathbb{E} \{ |q_k|^2 \} = 1$, e η_{mk} são os coeficiente de controle de potência que satisfaz a restrição de potência para cada AP, demonstrado por.

$$\mathbb{E} \{ |x_m|^2 \} \leq \rho_d. \quad (12)$$

Logo, a restrição de potência pode ser descrita como

$$\sum_{k=1}^K \eta_{mk} \gamma_{mk} \leq 1, \quad \text{para todas as APs}, \quad (13)$$

onde,

$$\gamma_{mk} \triangleq \mathbb{E} \{ |\hat{g}_{mk}|^2 \} = \frac{\tau_p \rho_p (\beta_{mk})^2}{\tau_p \rho_p \sum_{k'=1}^K \beta_{mk'} |\varphi_k^H \varphi_{k'}|^2 + 1}. \quad (14)$$

Logo, o sinal correspondente recebido da k-ésima UE corresponde a

$$\begin{aligned} r_{d,k} &= \sum_{m=1}^M g_{mk} x_{mk} + w_{d,k} \\ &= \sqrt{\rho_d} \sum_{m=1}^M \sum_{k'=1}^K \eta_{mk'}^{1/2} g_{mk} \hat{g}_{mk'}^* q_{mk'} + w_{d,k}, \end{aligned} \quad (15)$$

onde $w_{d,k}$ é o ruído aditivo com distribuição $\mathcal{CN}(0, 1)$ para a k-ésima UE

IV. FORMULAÇÃO DO PROBLEMA

Neste trabalho, adota-se a técnica de *max-min fairness* para a alocação de potência em sistemas CF Massive MIMO, com o objetivo de otimizar a distribuição de potência no enlace de *downlink* entre os usuários finais (UEs). A principal motivação é aumentar a SE dos usuários em condições de canal desfavoráveis, promovendo um balanceamento equitativo no desempenho global da rede. Nesse contexto, a SE individual está diretamente associada à razão sinal-interferência-mais-ruído (SINR), conforme definido na Eq.(10), a qual representa a qualidade do enlace de comunicação de cada usuário. Assim, o problema consiste em maximizar a menor taxa de transmissão entre os usuários, dada por

$$R_k = \log_2(1 + SINR_k) \quad (16)$$

V. APRENDIZADO POR REFORÇO

O ambiente de aprendizado por reforço foi desenvolvido utilizando a linguagem de programação *Python* versão 3.9.21 em conjunto com a biblioteca *Gymnasium* (Gym)¹, que permite a criação de ambientes de RL personalizados. Para o treinamento do agente, empregou-se a biblioteca *Stable-Baselines3* (SB3)², que fornece uma variedade de algoritmos de aprendizado por reforço compatíveis com ambientes do Gym. O ambiente personalizado segue a estrutura padrão do Gym, composto por duas funções principais:

- 1) *Reset*: Esta função reinicia o ambiente ao início de cada episódio de treinamento. A cada novo episódio, o cenário de comunicação CF Massive MIMO é reconfigurado, incluindo a atualização aleatória das posições das APs e UEs.
- 2) *Step*: Esta função executa uma ação escolhida pelo agente, que consiste na alocação de potência para os usuários. A partir da ação tomada, o ambiente calcula a nova observação, recompensa, e indica se o episódio foi finalizado, completando os passos designados.

Além das funções principais do ambiente, também são necessários outros componentes essenciais para o processo de aprendizado do agente.

¹<https://gymnasium.farama.org/>

²<https://stable-baselines3.readthedocs.io/en/master/index.html>

- 1) *State*: O estado do ambiente é definido pelo vetor de valores de SINR de cada UE. Ou seja, o agente observa o SINR atual de cada UE como entrada do ambiente.
- 2) *Action*: O agente seleciona ações representadas por um vetor de dimensão $K \times M$, onde K é o número de UEs e M o número de APs. Essas ações são valores contínuos, representando o percentual de potência da m -ésima AP alocada para k -ésima UE, no intervalo $[-1, 1]$, devido à normalização exigida pelo SB3, que assume uma distribuição normal $N(0,1)$ para o espaço de ações contínuas. No entanto, esses valores ainda precisam ser transformados para representar corretamente as potências alocadas no sistema. A alocação final de potência é calculada com base nos canais de cada UE para seus respectivos APs, definidas por:

$$\eta_{mk} = \frac{a_{mk}}{\gamma_{mk}}, \quad (17)$$

onde a_{mk} são as ações contínuas geradas pelo agente.

- 3) *Reward*: A função de recompensa é projetada para otimizar o problema de max-min, buscando maximizar a menor SE entre todos os usuários, enquanto respeita a restrição de potência máxima de cada AP. Assim, a recompensa é definida com base na menor SE alcançada, desde que a soma das potências alocadas por AP não exceda seu limite máximo. Logo a função de recompensa é definida por:

$$r^t = \begin{cases} \min(SE_k), & \sum_{k=1}^K \eta_{m,k} \gamma_{m,k} \leq 1 \\ -\frac{E}{4}, & \text{caso contrário} \end{cases} \quad (18)$$

onde E é uma variável que penaliza o agente caso ele ultrapasse a restrição de potência da AP. A divisão por quatro é utilizada para que a penalidade não seja muito elevada, pois o valor de E não é constante, onde esse valor foi obtido por meio de testes.

Foi utilizado o algoritmo PPO como agente para otimizar a alocação de potência no sistemas CF Massive MIMO. No agente, todas as APs são modeladas como parte do ambiente, enquanto o agente atua como controlador centralizado, otimizando simultaneamente as ações de transmissão para APs e UEs.

O PPO pertence à classe dos métodos *actor-critic*, nos quais duas redes neurais distintas atuam de forma complementar. A rede do ator é responsável por representar a política, isto é, gerar ações com base nos estados observados e interagir diretamente com o ambiente. Já a rede do crítico estima a função de valor, avaliando o desempenho das ações tomadas pelo ator e fornecendo sinal de aprendizado que orienta a atualização da política [7].

Uma das principais vantagens do PPO em ambientes com espaços de ação contínuos é que o crítico pode ajustar a função valor-ação diretamente, sem a necessidade de o ator buscar a ação ótima de forma explícita. Essa característica torna os métodos *actor-critic*, como o PPO, particularmente adequados para tarefas de otimização contínua.

O PPO pode utilizar duas formas distintas de função objetivo para substituir a função original: *clipping* ou penalidade baseada na divergência KL[12]. No SB3, é adotada a

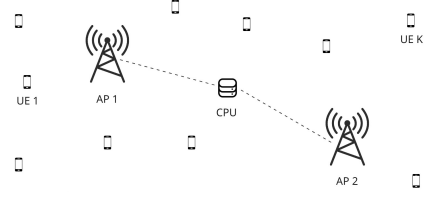


Fig. 1. Sistema Cell-Free Massive MIMO

abordagem com *clipping*, que limita o quanto a nova política pode se desviar da política anterior, assim visando otimizar uma política estocástica de forma estável e eficiente. A função utilizada é descrita como [12]

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)], \quad (19)$$

onde ϵ é um hiperparâmetro, normalmente utilizado o valor de $\epsilon = 0,2$, e $r_t(\theta)$ é a razão de probabilidade denotada por $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$, e o termo *clip* restringe $r_t(\theta)$ nos intervalos de $1 - \epsilon$ e $1 + \epsilon$ e $\hat{A}_t$ é um estimador para a função de vantagem.

Além disso, conforme destacado em [12], quando se utilizam arquiteturas neurais que compartilham parâmetros entre a política e a função de valor, é comum adotar uma função de perda composta. Essa função inclui tanto o termo da perda da política (*surrogate loss*) quanto um termo que penaliza o erro na estimativa da função de valor.

A combinação desses termos resulta em uma função objetivo que é maximizada, ainda que de forma aproximada, a cada iteração do treinamento.

$$L_t^{CLIP+VF+S}(\theta) = \hat{\mathbb{E}}_t[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)], \quad (20)$$

onde c_1 é o coeficiente de entropia (*ent_coef*), e c_2 o coeficiente da função de valor, e S é o bônus de entropia, e L_t^{VF} é a perda do erro quadrático $(V_\theta(s_t) - V_t^{targ})^2$.

A maximização dessa função objetivo permite que o agente aprenda políticas mais eficazes e robustas, equilibrando atualização estável da política, aprendizado da função de valor e comportamento exploratório — características essenciais em ambientes com ações contínuas, como na alocação de potência em sistemas CF Massive MIMO.

VI. RESULTADOS

Para os resultados, foi considerado a rede com $M = 2$ e $K = 10$, onde cada AP e UE são distribuídos de forma aleatória em um cenário com tamanho $250 \text{ km} \times 250 \text{ km}$, representado na Figura 1, com a frequência da onda portadora de 1,9 GHz. Os parâmetros utilizados são descritos na Tabela I, conforme descritos em [1].

Para verificar o desempenho do algoritmo PPO, foi realizada uma comparação com uma abordagem do *baseline*, na qual a alocação de potência prioriza os usuários com os melhores canais (*baseline* canal-forte). O treinamento do agente PPO foi conduzido utilizando 1 milhão de passos por meio da biblioteca SB3. Foram mantidos os hiperparâmetros padrão do algoritmo, com exceção de algumas modificações descritas na

TABELA I
PARÂMETROS DO CENÁRIO.

Parâmetros	Valor
Frequência da portadora	1,9 GHz
Largura de banda	20 MHz
h_{AP}	10 m
h_u	1,65 m
Ruído	9 dB
d_1, d_0	50 m, 10 m
ρ_d, ρ_u	100 mW, 10 mW

Tabela II onde foram obtidas através de testes. A rede utiliza camadas completamente conectadas com tamanhos 128×128 tanto para a rede ator quanto para a rede crítica.

TABELA II
HIPERPARÂMETROS DO AGENTE.

Parâmetros	Valor
ent_coef	0,01
$Batch_size$	128
n_steps	1024
n_epochs	5

As modificações nos hiperparâmetros apresentadas na Tabela II foram realizadas com o objetivo de aprimorar o desempenho do agente. Entre elas, destaca-se o ajuste do coeficiente de entropia (ent_coef), cuja função é incentivar uma exploração mais ampla do espaço de ações. Esse ajuste busca evitar a convergência prematura para políticas excessivamente determinísticas, promovendo o aprendizado de estratégias mais robustas e com maior capacidade de generalização.

A Figura 2 apresenta a função de distribuição acumulada (CDF) da taxa por usuário no cenário *orthogonal-multiple-access* (OMA), comparando a política inicial com a aprendida pelo agente PPO. Observa-se que o desempenho associado ao PPO apresentou melhorias significativas para os usuários medianos e os piores usuários. Em particular, os maiores ganhos ocorrem entre os usuários com pior desempenho: por exemplo, a taxa correspondente a 10% dos usuários mais desfavorecidos aumentou de aproximadamente 0,03 para 0,15 bits/s/Hz — uma melhoria de cinco vezes. Além disso, para o usuário mediano (CDF = 0,5), a taxa aumentou de 0,10 para 0,18 bits/s/Hz, representando um ganho relativo de cerca de 80%. Destaca-se também que, para uma SE de 0,1 bits/s/Hz, a CDF utilizando o PPO foi de aproximadamente 32%, ou seja, há 32% de chance de que um usuário selecionado aleatoriamente tenha SE menor ou igual a esse valor. Já com o *baseline*, essa CDF sobe para cerca de 53%, indicando que mais da metade dos usuários tem desempenho inferior ou igual a 0,1 bits/s/Hz. Esses resultados indicam que a política aprendida melhora não apenas a SE média, mas também favorece os usuários em pior situação, promovendo maior equidade na alocação de recursos.

VII. CONCLUSÕES

Neste trabalho, foi investigado o desempenho do agente PPO na alocação de potência em um ambiente de Cell-Free MIMO, uma abordagem que, até o momento, não havia

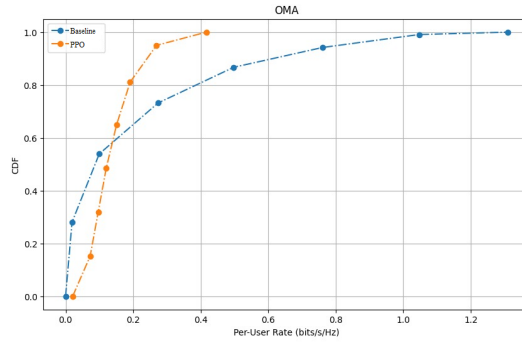


Fig. 2. CDF da eficiência espectral por usuário

sido explorada em outros estudos nessa área. Embora o PPO tenha sido utilizado com sucesso em outras aplicações de aprendizado por reforço, dedicou-se neste estudo a adaptá-lo especificamente para o cenário de alocação de potência. Os resultados mostraram uma melhoria modesta na equidade da alocação de potência segundo o critério max-min, mas o desempenho do agente ainda foi limitado pela tendência de se fixar em uma política subótima. Esse comportamento sugere a necessidade de ajustes no processo de treinamento ou na modelagem do ambiente para evitar que o agente se prenda a soluções não ideais.

REFERÊNCIAS

- [1] 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, pp. 1834-1850, Março 2017.
- [2] E. Nayebi, A. Ashikhmin, T. L. Marzetta and B. D. Rao, "Performance of cell-free massive MIMO systems with MMSE and LSFD receivers," *2016 50th Asilomar Conference on Signals, Systems and Computers*, pp. 203-207, 2016.
- [3] S. Buzzi, C. -L. I, T. E. Klein, H. V. Poor, C. Yang and A. Zappone, "CA Survey of Energy-Efficient Techniques for 5G Networks and Challenges Ahead," *IEEE Journal on Selected Areas in Communications*, v. 34, pp. 697-709, Abril 2016.
- [4] N. C. Luong et al., "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Communications Surveys & Tutorials*, v. 21, pp. 3133-3174, 2019.
- [5] Zhenyuan Feng and Bruno Clerckx, "Deep Reinforcement Learning for Multi-user Massive MIMO with Channel Aging," *arXiv*, 2023.
- [6] S. Chakraborty and B. R. Manoj, "Power Allocation in a Cell-Free MIMO System using Reinforcement Learning-Based Approach," *2023 National Conference on Communications*, pp. 1-6, 2023.
- [7] L. Luo, J. Zhang, S. Chen, X. Zhang, B. Ai and D. W. K. Ng, "Downlink Power Control for Cell-Free Massive MIMO With Deep Reinforcement Learning," *IEEE Transactions on Vehicular Technology*, v. 71, pp. 6772-6777, Julho 2022.
- [8] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson and T. L. Marzetta, "Cell-Free Massive MIMO: Uniformly great service for everyone," *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications*, pp. 201-205, 2015.
- [9] Ao Tang, JiXian Sun and Ke Gong, "Mobile propagation loss with a low base station antenna for NLOS street microcells in urban area," *IEEE VTS 53rd Vehicular Technology Conference*, v. 1, pp. 333-336, 2001.
- [10] Z. Wang, E. K. Tameh and A. R. Nix, "Joint Shadowing Process in Urban Peer-to-Peer Radio Channels," *IEEE Transactions on Vehicular Technology*, v. 57, pp. 52-64, Janeiro 2008.
- [11] Y. Zhao, I. G. Niemegeers and S. M. H. De Groot, "Dynamic Power Allocation for Cell-Free Massive MIMO: Deep Reinforcement Learning Methods," *IEEE Access*, v. 9, pp. 102953-102965, 2021.
- [12] John Schulman and Filip Wolski and Prafulla Dhariwal and Alec Radford and Oleg Klimov, "Proximal Policy Optimization Algorithms," *arXiv*, 2017.