

Controle Indireto por Ensemble DDPG para Escalonamento Multiobjetivo em HetNets 5G com Tráfego Misto eMBB/uRLLC

Gustavo C. Domingos e Éderson R. da Silva

Resumo— Este trabalho propõe um escalonador híbrido para redes 5G heterogêneas com tráfego misto eMBB/uRLLC. Um *ensemble* de agentes *Deep Deterministic Policy Gradient* (DDPG) ajusta dinamicamente parâmetros de equidade, latência e confiabilidade de um escalonador *Proportional Fair* subjacente. O controle indireto reduz a dimensionalidade do espaço de ação e favorece a convergência. Avaliações no *Vienna 5G System Level Simulator*, em 30 execuções independentes com intervalo de confiança de 95% e coeficiente de variação, indicam redução da violação de *deadline* uRLLC frente a *baselines* heurísticas e menor variabilidade, sugerindo o DRL como alternativa viável ao escalonamento estático.

Palavras-Chave— 5G, aprendizado por reforço profundo, DDPG, eMBB/uRLLC, escalonamento de recursos, HetNet, redes heterogêneas.

Abstract— This paper proposes a hybrid scheduler for 5G heterogeneous networks with mixed eMBB/uRLLC traffic. An ensemble of Deep Deterministic Policy Gradient (DDPG) agents dynamically adjusts fairness, latency, and reliability parameters of an underlying Proportional Fair scheduler. Indirect control reduces action-space dimensionality and improves convergence. Evaluations on the Vienna 5G System Level Simulator over 30 independent runs, using 95% confidence intervals and coefficient of variation, show reductions in uRLLC deadline violation relative to heuristic baselines, with lower run-to-run variability, suggesting DRL-based scheduling as a viable alternative to static resource allocation in heterogeneous networks.

Keywords— 5G, DDPG, deep reinforcement learning, eMBB/uRLLC, heterogeneous networks, HetNet, resource scheduling.

I. INTRODUÇÃO

A evolução das redes de comunicação móvel e a adoção de redes heterogêneas (HetNet) aumentaram exponencialmente a complexidade do gerenciamento de recursos [1], [2]. Nesse cenário, a coexistência de serviços *Enhanced Mobile Broadband* (eMBB) e *Ultra-Reliable Low-Latency Communications* (uRLLC) impõe um desafio de otimização multiobjetivo: satisfazer simultaneamente requisitos conflitantes de *throughput*, latência e equidade sob condições dinâmicas de canal e tráfego [3].

Como algoritmos heurísticos tradicionais, a exemplo do *Proportional Fair* (PF) e do *Best Channel Quality Indicator* (Best CQI), operam com parâmetros fixos que não se adaptam a

essas variações dinâmicas [4], a automação por meio de Aprendizado por Reforço Profundo (DRL) surge como uma alternativa essencial [5]. Entre as abordagens de DRL, o *Deep Deterministic Policy Gradient* (DDPG) destaca-se por ser adequado a espaços de ação contínuos, permitindo que agentes inteligentes aprendam políticas adaptativas [6].

Neste contexto, o trabalho apresenta três contribuições. Primeiro, propõe e valida um escalonador DRL híbrido para HetNet com tráfego misto eMBB/uRLLC, combinando controle indireto via DDPG com um escalonador PF subjacente, o que reduz a dimensionalidade do espaço de ação e facilita a convergência. Segundo, introduz uma estratégia de *ensemble* com seleção de agentes por *ranking*, com ganho mensurável de robustez frente à seleção aleatória. Terceiro, adota um padrão de avaliação com 30 execuções independentes, coeficiente de variação e Intervalo de Confiança (IC) de 95%, favorecendo comparação rigorosa e reprodutibilidade.

II. TRABALHOS RELACIONADOS

O escalonamento de recursos em redes 5G com tráfego misto tem sido amplamente estudado sob a perspectiva de otimização analítica e heurísticas [4], [7]. Métodos baseados em PF ponderado e preempção de pacotes uRLLC sobre eMBB [8] assumem conhecimento do estado do canal, limitando sua aplicabilidade em ambientes dinâmicos.

A aplicação de DRL à alocação de recursos foi sistematizada em [5], e o uso de DDPG para espaços de ação contínuos foi proposto em [6]. Li e Zhang [14] aplicaram DDPG ao escalonamento conjunto de eMBB e uRLLC, otimizando alocação de banda e sobreposição de tráfego uRLLC. Tseng *et al.* [15] propuseram uma arquitetura DDPG modularizada para 5G NR, na qual a ação combina algoritmos de escalonamento existentes. Gu *et al.* [16] desenvolveram um DDPG assistido por QoS (K-DDPG), reduzindo o tempo de convergência em 30–50%. Hsu e Liao [17] introduziram um mecanismo DRL *event-driven* com escalonadores separados para eMBB e uRLLC, e o DDPG também foi explorado em redes com fatiamento [9]. Em HetNet, trabalhos recentes investigam DRL para coordenação de interferência [10], *beamforming* adaptativo [11] e aprendizado federado para adaptação distribuída [12].

Diferentemente dos trabalhos anteriores, nos quais o agente DRL atua diretamente na seleção de recursos, o que implica espaços de ação de alta dimensionalidade e convergência mais lenta, a abordagem proposta emprega controle indireto: o agente ajusta apenas três parâmetros (equidade alvo, urgência de latência e peso de confiabilidade) de um escalonador PF subjacente, que permanece responsável pela alocação RB a RB em todos os *Transmission Time Intervals* (TTIs). Essa

formulação reduz a dimensionalidade do espaço de ação, acelera a convergência e preserva o comportamento estrutural da heurística mesmo sob políticas ainda não treinadas. Adicionalmente, nenhum dos trabalhos relacionados emprega *ensemble* de agentes com seleção por *ranking* nem avaliação estatística com múltiplas execuções independentes, lacunas que este trabalho busca tratar.

III. SISTEMA PROPOSTO

A. Cenário HetNet e Ambiente de Simulação

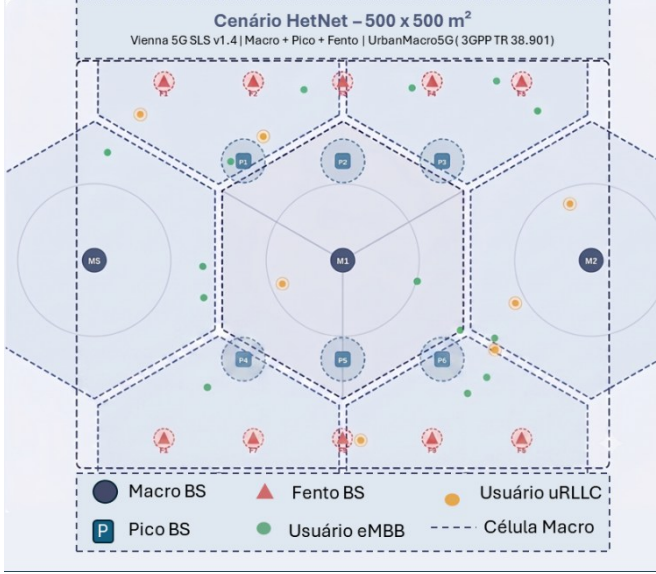


Fig. 1. Cenário HetNet.

O cenário HetNet, ilustrado na Fig. 1, é implementado no *Vienna 5G System Level Simulator* (SLS) v1.4 [18] sobre uma área de $500 \times 500 \text{ m}^2$, com três tipos de estações base. A célula macro adota topologia HexRing com 7 células e 3 setores cada, distância intersite de 250 m, altura de 25 m e potência de 40 W. A célula pico é composta por 6 pico *Base Stations* (BSs) em posições predefinidas, com altura de 5 m e potência de 12 W. A célula femto compreende 10 femto BSs em posições fixas, com altura de 1,5 m e potência de 1 W. Todas as BSs operam com antena *Single Input Single Output* (SISO) e adotam o modelo UrbanMacro5G (3GPP TR 38.901), com componentes com Linha de Visada (LOS) e Sem Linha de Visada (NLOS).

Todas as BSs operam com frequência central de 2 GHz (FR1, Sub-6 GHz) e largura de banda de 10 MHz, com Numerologia 0 (espaçamento de subportadora de 15 kHz e duração de *slot* de 1 ms). Com 12 subportadoras por RB e 90% de banda utilizável, essa configuração resulta em 50 RBs na dimensão de frequência e 100 RBs por *slot*, sem agregação de portadoras.

Os usuários são distribuídos segundo processo espacial de Poisson bidimensional, com densidades de $30 \times 10^{-6} \text{ u/m}^2$ para eMBB e $10 \times 10^{-6} \text{ u/m}^2$ para uRLLC, resultando em aproximadamente 7 e 2 usuários, respectivamente, por célula. A associação segue o critério de *Reference Signal Received Power* (RSRP) sem *Cell Range Expansion*. Usuários eMBB operam em *Full Buffer*; usuários uRLLC seguem processo de chegada Poisson com *deadline* de 60 *slots* [8].

B. O Escalonador

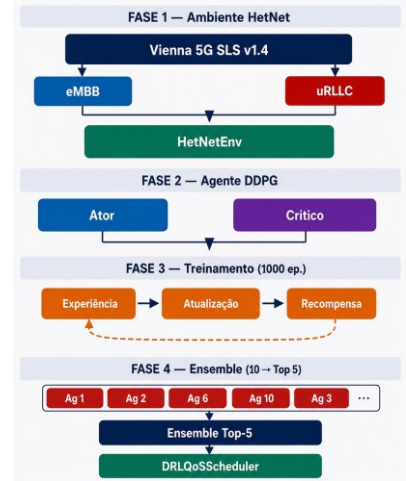


Fig. 2. Fluxograma do sistema proposto.

A Fig. 2 resume o processo em quatro fases: modelagem do ambiente HetNet no Vienna 5G SLS (Fase 1), definição do agente DDPG com redes ator e crítico (Fase 2), treinamento pelo ciclo (Fase 3) e seleção dos agentes para compor o *ensemble* (Fase 4). O agente DDPG seleciona, a cada TTI, o vetor de ação $a = [J, \alpha_{lat}, \alpha_{rel}] \in [0,1]^3$, correspondente ao índice de Jain alvo, peso de urgência de latência e peso de confiabilidade. O espaço de observação possui 8 dimensões: *throughput* normalizado, índice de Jain atual, taxa de violação de latência, *Block Error Rate* (BLER) média, eficiência do escalonador e os três parâmetros da ação anterior.

A função de recompensa multiobjetivo, dada pela Eq. (1), combina quatro termos: o *throughput* normalizado (thr_{norm}), que é a razão entre o *throughput* médio instantâneo e o de pico da célula; o índice de Jain (Jain), que mede a equidade entre usuários; a taxa de violação de *deadline* uRLLC (latViol), representando a fração de pacotes que excedem o *deadline*; e um termo de regularização sobre o vetor de ação a , que penaliza desvios do ponto central $[0,5; 0,5; 0,5]$.

$$r = 0,35 \cdot \text{thr}_{\text{norm}} + 0,40 \cdot \text{Jain} + 0,25 \cdot (1 - \text{latViol}) - 0,05 \cdot \|a - 0,5\|^2 \quad (1)$$

O último termo da Eq. (1) não mede equidade, papel do índice de Jain, mas regulariza o espaço de ação: com peso reduzido (0,05), desencoraja a saturação das ações nos extremos de $[0,1]$, onde o escalonador opera em regimes degenerados, e suaviza a variação dos parâmetros entre TTIs, favorecendo a estabilidade do treinamento.

O escalonador PF estendido descrito a seguir calcula, para cada par usuário e bloco de recurso (RB), um *score* que combina a função utilidade do PF com urgências de latência e confiabilidade baseadas em funções exponenciais do *deadline* residual e da BLER, respectivamente. Uma correção iterativa aproxima o índice de *fairness* obtido do valor alvo J .

O papel do agente DDPG restringe-se a ajustar dinamicamente os três parâmetros $[J, \alpha_{lat}, \alpha_{rel}]$ consumidos por esse escalonador que executa a alocação RB a RB em todos os TTIs e é detalhado a seguir, o que justifica a denominação de controle indireto.

C. Arquitetura do Agente DDPG

O DDPG é um algoritmo de aprendizado por reforço profundo para espaços de ação contínuos [6], [19]. Diferentemente do *Deep Q-Network* (DQN), que opera em espaços discretos, emprega uma política determinística (ator) que mapeia estados diretamente para ações contínuas, adequada às ações $a = [J, \alpha_{lat}, \alpha_{rel}] \in [0,1]^3$ [6], [15]. O agente possui quatro redes: ator $\pi(s)$, crítico $Q(s,a)$ e suas redes-alvo π' e Q' . A cada passo: (i) o ator seleciona a ação com ruído de Ornstein-Uhlenbeck ($\sigma: 0,2 \rightarrow 0,02$); (ii) o ambiente retorna a recompensa e o estado s' ; (iii) a transição (s, a, r, s') é armazenada no *replay buffer* (capacidade 20.000); e (iv) um *mini-batch* de 128 experiências atualiza o crítico (erro de Bellman) e o ator (gradiente de política). As redes-alvo usam *soft update* $\theta' \leftarrow \tau\theta + (1 - \tau)\theta'$ com $\tau = 0,005$ [6], [16], [19].

TABELA I. HIPERPARÂMETROS DO TREINAMENTO DDPG.

Hiperparâmetro	Valor
Episódios de treinamento	1.000
Passos por episódio	20
Capacidade do replay buffer	20.000
Batch size	128
Taxa de aprendizado — ator	1×10^{-4}
Taxa de aprendizado — crítico	1×10^{-3}
Fator de desconto (γ)	0,90
Fator de suavização (τ)	0,005
Ruído OU — desvio padrão inicial	0,20
Ruído OU — desvio padrão mínimo	0,02
Gradiente threshold (ator/crítico)	1,0

O ator é uma rede *feedforward* $8 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 3$, com ativação *Rectified Linear Unit* (ReLU) nas camadas ocultas e *Sigmoid* na saída, mapeando as ações para $[0,1]$. O crítico emprega arquitetura *dual-path*: observações (8D) e ações (3D) seguem caminhos paralelos: camadas de 256 e 128 neurônios para as observações, antes de serem somadas e condensadas em 64 neurônios até o valor Q escalar [16]. A Tabela I resume os hiperparâmetros, definidos empiricamente em testes preliminares no cenário base, sem otimização automática.

D. Convergência e Ensemble

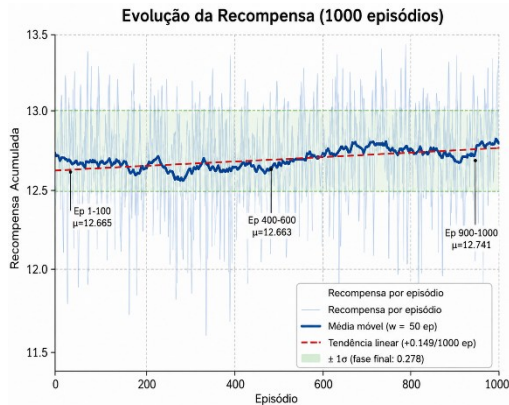


Fig. 3. Recompensa por episódio.

A Fig. 3 apresenta a curva de recompensa ao longo dos 1.000 episódios de treinamento, com média móvel de janela 50 e distribuição por fase. A recompensa média na fase inicial (ep 1–100) é de $12,67 \pm 0,34$, frente a $12,74 \pm 0,20$ na fase final (ep 900–1000), com ganho relativo de +0.60%. A média móvel varia

no intervalo $[12,55; 12,82]$ ao longo de todo o treinamento, indicando que o espaço de ação reduzido pelo controle indireto facilita a exploração inicial e resulta em convergência rápida para uma política estável.

A pequena variação absoluta da recompensa média é consequência direta do controle indireto: como o escalonador PF subjacente garante comportamento funcional mesmo sob parâmetros subótimos, o agente parte de um piso de recompensa elevado e estável. Assim, o aprendizado manifesta-se menos no crescimento da recompensa e mais na redução da variabilidade entre episódios, o desvio-padrão cai de $\pm 0,34$ para $\pm 0,20$ (~41%), e nas métricas finais de rede, sobretudo na violação de *deadline* uRLLC, reduzida em até 13,4 p.p. frente às *baselines* (Seção IV). Isso contrasta com a seleção direta de recursos, em que a recompensa parte de valores baixos e cresce, porém sem o piso de desempenho garantido pela heurística.

Dez agentes DDPG foram treinados com *seeds* distintas e avaliados em 5 episódios cada. Os cinco com maior recompensa média $S_{média}$ foram selecionados para o *ensemble*, cuja ação final é obtida pela média aritmética simples das ações dos cinco agentes. O *ensemble* top-5 supera 98.2% dos *ensembles* de cinco agentes selecionados aleatoriamente entre os dez treinados, confirmando que a seleção por *ranking* agrega robustez à política final.

E. Baselines de Comparação

Quatro algoritmos são avaliados como referência. O *Round Robin* (RR) aloca os RBs em rodízio igualitário entre os usuários, sem considerar condições de canal ou requisitos de qualidade de serviço (QoS), produzindo alta equidade por construção, mas ignorando urgências de latência. O *Best CQI* aloca cada RB ao usuário com maior indicador de qualidade de canal (CQI) instantâneo, maximizando a eficiência espectral ao custo de forte desigualdade na distribuição de recursos.

O QoS-Aware emprega as mesmas funções de urgência de latência e confiabilidade do escalonador DRL, exponenciais do *deadline* residual e da BLER, porém com parâmetros pré-configurados e imutáveis: $J = 0,7$, $\alpha_{lat} = 1,1$ e $\alpha_{rel} = 2,0$, definidos empiricamente para o cenário base. A comparação com o DRL-QoS é, assim, metodologicamente rigorosa: ambos compartilham a mesma lógica subjacente, diferindo apenas no ajuste dos parâmetros, estático no QoS-Aware e dinâmico via DDPG no DRL-QoS, o que isola a contribuição do aprendizado. O DRL-QoS é o escalonador proposto neste trabalho.

IV. RESULTADOS

A avaliação foi conduzida em 30 execuções independentes no Vienna 5G SLS v1.4, com posicionamento aleatório de usuários. Os resultados são reportados em média e IC de 95% (t de Student, $n = 30$). O coeficiente de variação (CV%) é empregado como medida de variabilidade relativa entre execuções.

A. Desempenho Global

Os resultados da Tabela II evidenciam o *trade-off* clássico entre *throughput* e *fairness*. O *Best CQI* atinge o maior *throughput* médio (1,609 Mbps) ao custo do menor índice de Jain (0,367), refletindo alocação enviesada em favor de usuários com melhores condições de canal, efeito amplificado pela disparidade de potência entre células em ambiente HetNet.

TABELA II. CENÁRIO HETNET BASE (N = 30, IC 95%).

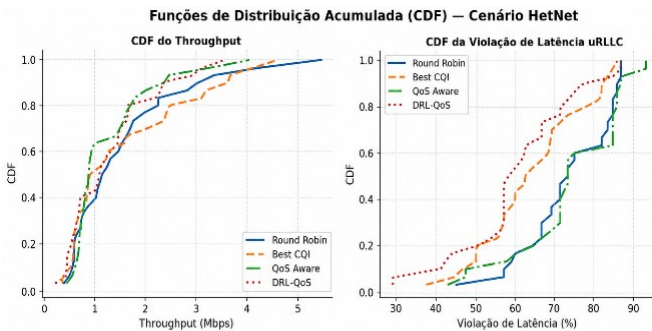
Scheduler	Thr (Mbps) IC95	CV% Thr	Jain IC95	LatViol (%) IC95	BLER
Round Robin	1,578 ±0,444	75,3	0,423 ±0,060	73,59 ±4,12	0,0003
Best CQI	1,609 ±0,460	77,2	0,367 ±0,060	64,54 ±4,85	0,0004
QoS-Aware	1,278 ±0,310	66,8	0,474 ±0,080	74,05 ±4,93	0,0002
DRL-QoS	1,308 ±0,323	66,1	0,403 ±0,064	60,63 ±5,43	0,0003

O RR obtém maior equidade relativa que o DRL-QoS (Jain = 0,423 frente a 0,403), porém com elevada violação de latência (73,59%). O QoS-Aware alcança o maior Jain (0,474) e o menor BLER (0,0002), mas apresenta o menor *throughput* médio (1,278 Mbps) e maior taxa de violação (74,05%).

O DRL-QoS apresenta a menor taxa de violação de *deadline* uRLLC (60,63%), com redução de 13,0 p.p. frente ao RR, 13,4 p.p. frente ao QoS-Aware e 3,9 p.p. frente ao *Best CQI*, mantendo *throughput* e *fairness* em posição intermediária. Notavelmente, o Jain do DRL-QoS (0,403) é inferior ao do RR (0,423): o agente descobriu autonomamente que sacrificar equidade é necessário para reduzir a violação de latência, priorizando usuários uRLLC. Esse *trade-off* emergente é consistente com a função de recompensa, que pondera 0,40 à *fairness* e 0,25 ao controle de latência num ambiente com requisitos uRLLC restritivos. O método apresenta ainda o menor CV de *throughput* (66,1%) e o menor IC de Jain ($\pm 0,064$), indicando maior estabilidade entre execuções.

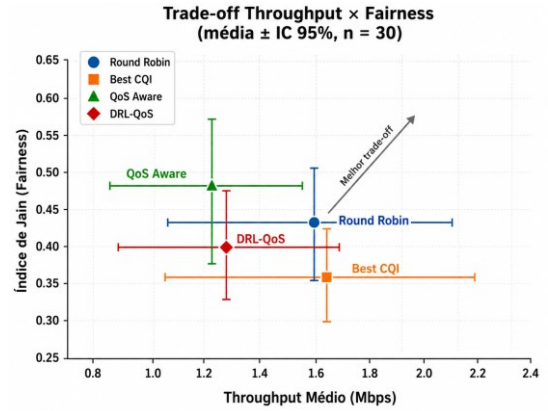
As taxas absolutas de violação, elevadas para todos os métodos, e os valores conservadores de *throughput* decorrem do dimensionamento propositalmente severo do cenário (*deadlines* curtos, alta densidade, banda estreita, antenas SISO e interferência entre camadas sem coordenação inter-tier), adotado para evidenciar diferenças entre escalonadores com o objetivo de comparação relativa sob condições idênticas, não a maximização do *throughput* absoluto.

B. Análise por CDF e Trade-off


 Fig. 4. Funções de distribuição acumulada (CDF) de (a) *throughput* e (b) violação de latência.

A Fig. 4 apresenta as funções de distribuição acumulada (*Cumulative Distribution Function* — CDF) de *throughput* e violação de latência. O DRL-QoS e o QoS-Aware exibem distribuições mais concentradas, com menor probabilidade de casos extremos. Na CDF de violação, a curva do DRL-QoS

desloca-se para menores valores em toda a sua extensão frente ao RR e ao QoS-Aware, confirmando que a redução não se restringe às médias; a proximidade entre DRL-QoS e *Best CQI* reflete a menor diferença absoluta entre eles.


 Fig. 5. Trade-off *throughput* × Jain (esquerda) e coeficiente de variação por métrica (direita).

A Fig. 5 confirma que nenhum algoritmo domina simultaneamente *throughput* e *fairness*: o *Best CQI* ocupa o extremo de alto *throughput* e baixo Jain, e o QoS-Aware o oposto. O DRL-QoS posiciona-se próximo ao QoS-Aware em *throughput*, característica da priorização do tráfego uRLLC, com menores ICs em ambas as dimensões ($\pm 0,323$ Mbps e $\pm 0,064$ em Jain) e os menores CVs de *throughput* (66,1% e 66,8%), indicando maior consistência. O maior CV de latência do DRL-QoS (24,0%) é esperado em políticas adaptativas, que exploram diferentes prioridades e apresentam maior dispersão nessa métrica, ainda que com menor média.

Em *throughput* e Jain, os ICs do DRL-QoS se sobrepõem aos das *baselines*, esperado, dada a alta variabilidade espacial do cenário. O ganho concentra-se na métrica-alvo uRLLC, a violação de *deadline*, na qual o DRL-QoS obtém a menor média e menor variabilidade. O benefício do aprendizado não é dominar todas as métricas, mas deslocar o ponto de operação do *trade-off* rumo ao requisito mais crítico. O tempo de inferência de 0,604 ms (Seção IV-D) é compatível com o TTI de 1 ms, e o treinamento ocorre uma única vez, *off-line*.

C. Avaliação Exploratória — Cenários Variantes HetNet

Foram conduzidas avaliações preliminares em três cenários variantes: *Loaded* (2× usuários, *deadline* 30 slots para uRLLC), *Mobile* (mobilidade, *deadline* 60 slots para uRLLC) e *Stress* (4× usuários, *deadline* 20 slots para uRLLC). Cada cenário foi executado 30 vezes de forma independente. Dado o tamanho amostral reduzido, os resultados devem ser interpretados como indicativos de tendências.

TABELA III. CENÁRIO LOADED (2× USUÁRIOS, DEADLINE = 30 SLOTS).

Scheduler	Thr (Mbps)	Jain	LatViol (%)	BLER
Round Robin	0,860 ±0,202	0,388 ±0,072	53,31 ±9,57	0,0027
Best CQI	1,005 ±0,204	0,351 ±0,074	56,47 ±10,80	0,0019
QoS-Aware	0,782 ±0,184	0,397 ±0,065	50,23 ±9,38	0,0017
DRL-QoS	0,868 ±0,137	0,391 ±0,085	53,39 ±10,40	0,0016

TABELA IV. CENÁRIO MOBILE (DEADLINE = 60 SLOTS).

Scheduler	Thr (Mbps)	Jain	LatViol (%)	BLER
Round Robin	1,691 ±0,397	0,380 ±0,049	54,66 ±10,92	0,0011
Best CQI	1,611 ±0,330	0,339 ±0,062	54,28 ±10,24	0,0015
QoS-Aware	1,585 ±0,357	0,376 ±0,051	54,76 ±10,75	0,0013
DRL-QoS	1,524 ±0,344	0,401 ±0,049	52,03 ±9,56	0,0013

TABELA V. CENÁRIO STRESS (4× USUÁRIOS, DEADLINE = 20 SLOTS).

Scheduler	Thr (Mbps)	Jain	LatViol (%)	BLER
Round Robin	0,435 ±0,063	0,149 ±0,052	57,06 ±9,90	0,0068
Best CQI	0,503 ±0,066	0,145 ±0,029	63,90 ±11,07	0,0052
QoS-Aware	0,412 ±0,069	0,145 ±0,038	60,03 ±10,42	0,0028
DRL-QoS	0,467 ±0,063	0,140 ±0,041	54,93 ±9,43	0,0034

No cenário *Loaded* (Tabela III), o QoS-Aware apresenta a menor violação de latência ($50,23 \pm 9,38\%$), seguido do DRL-QoS ($53,39 \pm 10,40\%$), com intervalos de confiança amplamente sobrepostos. Esse desempenho inferior deve-se a dois fatores: os agentes foram treinados apenas no cenário base e avaliados no *Loaded* sem re-treinamento (generalização fora da distribuição, com o dobro de usuários e *deadline* pela metade); e os parâmetros fixos do QoS-Aware ($J = 0,7$; $\alpha_{lat} = 1,1$; $\alpha_{rel} = 2,0$) privilegiam fortemente a latência, coincidindo com esse regime. Ainda assim, o DRL-QoS mantém o menor BLER (0,0016) e o o intervalo de confiança de throughput mais estreito ($\pm 0,137$ Mbps), indicando maior consistência entre execuções: o re-treinamento nos cenários variantes fica como trabalho futuro.

No cenário *Mobile* (Tabela IV), o DRL-QoS obtém a menor violação ($52,03 \pm 9,56\%$) e o maior Jain ($0,401 \pm 0,049$), sugerindo boa adaptação a ambientes com mobilidade; seu BLER (0,0013) iguala o do QoS-Aware, ambos inferiores aos demais. No cenário *Stress* (Tabela V), apresenta a menor violação ($54,93 \pm 9,43\%$) e o menor BLER (0,0034), com redução de 8,97 p.p. frente ao *Best CQI* e 5,10 p.p. frente ao *RR*.

D. Tempo de Inferência

O tempo isolado da rede neural, medido em 1.000 chamadas consecutivas após aquecimento do compilador JIT (Intel Core i5-1135G7, 16 GB RAM, sem GPU), é de 0,604 ms por decisão de *ensemble* — inferior ao TTI de 1 ms da Numerologia 0 do 5G NR ($\mu = 0$) quando considerado o contexto de janelas de 1.000 slots (3GPP TS 38.211).

V. CONCLUSÃO

Este trabalho propôs e avaliou um escalonador multiobjetivo baseado em DDPG para redes 5G heterogêneas com tráfego misto eMBB/uRLLC. O DRL-QoS reduziu a violação de latência uRLLC em 13,0 p.p. frente ao *RR* e 13,4 p.p. frente ao QoS-Aware no cenário base ($n = 30$, IC 95%), mantendo *throughput* e *fairness* intermediários e o menor coeficiente de variação de *throughput* entre os escalonadores. Nos cenários variantes, o agente foi consistente: no *Mobile*, menor violação (52,03%) e maior Jain (0,401); no *Stress*, menor violação (54,93%) e menor BLER (0,0034). A metodologia de *ensemble*

com seleção por *ranking* garantiu robustez, com inferência de 0,604 ms por decisão, compatível com o TTI de 1 ms. Como trabalhos futuros, destacam-se coordenação inter-tier, *Cell Range Expansion*, re-treinamento em densidades variáveis de pequenas células e a extensão a redes 6G [13].

REFERÊNCIAS

- [1] M. Bennis, M. Debbah e H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.
- [2] X. Lin et al., "5G New Radio: Unveiling the Essentials of the Next Generation Wireless Access Technology," *IEEE Communications Standards Magazine*, vol. 3, no. 3, pp. 30–37, Sep. 2019.
- [3] G. Andrade Queiroz e E. Rosa da Silva, "A new Channel and QoS Aware Scheduler algorithm for real time and non real time traffic in 5G heterogeneous networks," *IEEE Latin America Transactions*, vol. 22, no. 8, pp. 659–669, Aug. 2024.
- [4] A. Anand, G. de Veciana e S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 28, no. 2, pp. 477–490, Apr. 2020.
- [5] O. Naparstek e K. Cohen, "Deep Multi-User Reinforcement Learning for Distributed Dynamic Spectrum Access," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 310–323, Jan. 2019.
- [6] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," in *Proc. ICLR*, San Juan, Puerto Rico, 2016.
- [7] H. Yin, L. Zhang e S. Roy, "Multiplexing URLLC Traffic Within eMBB Services in 5G NR: Fair Scheduling," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 1080–1093, Feb. 2021.
- [8] P. Popovski et al., "Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.
- [9] H. H. de Souza Lopes et al., "Deep Deterministic Policy Gradient-Based Resource Allocation Considering Network Slicing and Device-to-Device Communication in Mobile Networks," *Sensors*, vol. 24, no. 18, p. 6079, Sep. 2024.
- [10] O. Giwa et al., "Optimisation of resource allocation in heterogeneous wireless networks using deep reinforcement learning," *arXiv:2509.25284*, 2025.
- [11] S. Shukry e Y. Fahmy, "Modified prioritized DDPG algorithm for joint beamforming and RIS phase optimization in MISO downlink systems," *Scientific Reports*, vol. 16, 5942, 2026.
- [12] I. Djemai, M. Sarkiss e P. Ciblat, "Federated reinforcement learning for scheduling-offloading policies in multi-cluster NOMA systems," *EURASIP Journal on Advances in Signal Processing*, vol. 2025, no. 1, Aug. 2025.
- [13] A. Hazra et al., "6G-Enabled Ultra-Reliable Low Latency Communication for Industry 5.0: Challenges and Future Directions," *IEEE Communications Standards Magazine*, vol. 8, no. 2, pp. 36–42, Jun. 2024.
- [14] J. Li e X. Zhang, "Deep Reinforcement Learning-Based Joint Scheduling of eMBB and URLLC in 5G Networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1543–1546, set. 2020.
- [15] S.-C. Tseng, Z.-W. Liu, Y.-C. Chou e C.-W. Huang, "Radio Resource Scheduling for 5G NR via Deep Deterministic Policy Gradient," in *Proc. IEEE ICC Workshops*, Xangai, China, mai. 2019, pp. 1–6.
- [16] Z. Gu, C. She, W. Hardjawana, S. Lumb, D. McKechnie, T. Essery e B. Vucetic, "Knowledge-Assisted Deep Reinforcement Learning in 5G Scheduler Design: From Theoretical Framework to Implementation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1904–1919, jul. 2021.
- [17] Y.-H. Hsu e W. Liao, "eMBB and URLLC Service Multiplexing Based on Deep Reinforcement Learning in 5G and Beyond," in *Proc. IEEE WCNC*, Austin, TX, EUA, abr. 2022, pp. 1–6.
- [18] L. Huang, T. Chen, Z. Gao, M. Luo e Z. Liu, "System Level Simulation for 5G Ultra-Reliable Low-Latency Communication," in *Proc. Int. Conf. Commun., Comput., Cybersecurity Inform. (CCCI)*, Beijing, China, out. 2021, pp. 1–5.
- [19] E. H. Sumiea et al., "Deep Deterministic Policy Gradient Algorithm: A Systematic Review," *Heliyon*, vol. 10, no. 9, art. e30697, mai. 2024.