

Otimização Dinâmica da Ordem de Roteamento e Atribuição de Espectro em EONs usando Aprendizado por Reforço

Alexandre L. Silva Jr^{*}, Matheus L. Santos^{*}, Gustavo Sousa Pavani[†], Hermano A. Cabral^{*} e Raul C. Almeida Jr.^{*}

^{*}Departamento de Eletrônica e Sistemas, Universidade Federal de Pernambuco, Recife, Brasil

[†]Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, Santo André, Brasil

e-mail: alexandre.lsilva2@ufpe.br

Resumo— O problema de Roteamento e Atribuição de Espectro (RSA) em Redes Ópticas Elásticas é frequentemente resolvido por heurísticas isoladas que priorizam o Roteamento (R-SA) ou o Espectro (SA-R). Este artigo propõe um mecanismo de Aprendizado por Reforço, baseado no algoritmo PPO (Proximal Policy Optimization), para a seleção dinâmica destas heurísticas. Operando sob observação restrita aos pares origem-destino, o agente treinado assimilou uma política adaptativa eficiente. Testes nas topologias NSFNet e Abilene comprovam que o PPO supera as heurísticas tradicionais, atingindo probabilidade de bloqueio equivalente à otimização via Algoritmos Genéticos (AG).

Palavras-Chave— Redes Ópticas Elásticas, Roteamento e Alocação de Espectro, Algoritmos Genéticos, Aprendizado por Reforço, Probabilidade de Bloqueio.

I. INTRODUÇÃO

O tráfego de dados globais tem experimentado um crescimento exponencial ano após ano, exigindo das redes de comunicação uma capacidade de resposta cada vez mais eficiente. As redes tradicionais baseadas em Multiplexação por Divisão de Comprimento de Onda (WDM), devido à sua grade espectral rígida e alocação estática, desperdiçam recursos valiosos e apresentam dificuldades para acomodar de forma eficiente essas demandas. Para superar essas limitações, as Redes Ópticas Elásticas (EONs - Elastic Optical Networks) foram desenvolvidas, permitindo uma alocação dinâmica e flexível do espectro, o que melhora substancialmente a utilização dos recursos da rede [1]. Nessas redes, a tarefa fundamental de acomodar as conexões e designar os recursos necessários é conhecida como o problema de Roteamento e Atribuição de Espectro (RSA - Routing and Spectrum Assignment).

Resolver os problemas de roteamento e atribuição de espectro de maneira simultânea apresenta uma alta complexidade computacional. Por esse motivo, a literatura frequentemente aborda o RSA de forma dissociada, utilizando estratégias de ordenação que priorizam uma das etapas do processo por vez [2]. A estratégia R-SA (Routing-Spectrum Assignment) prioriza a redução da ocupação global da rede, visto que busca alocar recursos primeiramente nas rotas mais curtas. Em contrapartida, a estratégia SA-R (Spectrum Assignment-Routing) prioriza uma organização mais eficiente do espectro e o balanceamento de carga, mesmo que isso implique na escolha de rotas com um maior número de saltos [2].

Embora ambas as estratégias apresentem benefícios sob diferentes aspectos topológicos da rede, a aplicação exclusiva de uma única heurística para toda a topologia restringe o desempenho geral. Visando combinar as vantagens de ambas as abordagens, a literatura recente propôs o HRSA (Hybrid Routing and Spectrum Assignment), uma estratégia que utiliza um Algoritmo Genético (AG) para determinar, de forma otimizada, qual ordenação (R-SA ou SA-R) deve ser aplicada a cada par de nós Origem-Destino (OD) [2]. Na abordagem de [2], a otimização ocorre de forma estática (offline): o AG gera indivíduos (cromossomos) que mapeiam a escolha binária de heurística para todos os pares OD da topologia simultaneamente. Para avaliar a aptidão (fitness) de cada solução candidata, o algoritmo necessita executar uma simulação completa e determinística do tráfego global da rede, o que demanda elevado custo computacional e de memória. Diante das limitações operacionais dessa avaliação em lote típica dos algoritmos evolutivos, surge a necessidade de investigar se modelos baseados em aprendizado sequencial e estocástico, que interagem com o ambiente de forma contínua e individualizada por requisição, são capazes de convergir para soluções de roteamento igualmente eficientes.

Para lidar com a complexidade e a dinamicidade das EONs, técnicas de Aprendizado de Máquina, especialmente o Aprendizado por Reforço Profundo (DRL - Deep Reinforcement Learning), emergiram como abordagens promissoras para a automação do controle e gerenciamento de recursos [3] [4]. Nesse cenário, o algoritmo Proximal Policy Optimization (PPO) tem se destacado por sua estabilidade de treinamento e eficácia em mitigar a fragmentação espectral e aumentar a utilização de recursos em longo prazo [5] [6].

Neste contexto, o presente trabalho propõe um mecanismo baseado em Aprendizado por Reforço (utilizando o algoritmo PPO) para o problema de seleção de ordenação R-SA/SA-R em Redes Ópticas Elásticas. O agente PPO é treinado a partir da interação contínua e estocástica com o ambiente de simulação. O objetivo deste estudo é investigar a capacidade do PPO em mapear requisições individuais (pares origem-destino) para as heurísticas R-SA e SA-R, avaliando seu desempenho computacional e a dinâmica de atribuição de recompensas em comparação ao limite ótimo das soluções estáticas.

II. METODOLOGIA

Para abordar o problema de seleção de ordenação no Roteamento e Atribuição de Espectro (RSA) em Redes Ópticas Elásticas (EONs), desenvolveu-se uma modelagem baseada em Processos de Decisão de Markov, substituindo a avaliação global e offline característica dos Algoritmos Genéticos por um modelo de Aprendizado por Reforço Profundo (DRL). A simulação estocástica de eventos discretos que sustenta o treinamento do agente foi implementada utilizando a interface padronizada da biblioteca Gymnasium [7].

A. Modelagem do Ambiente e Processo de Decisão de Markov (MDP)

O problema de seleção de heurísticas de alocação de recursos sob tráfego dinâmico é formulado como um Processo de Decisão de Markov (MDP), definido pela tupla $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, detalhada a seguir:

- Espaço de Observação ($\mathcal{S}$): No cenário restrito à identificação estrutural da requisição, o espaço de observação exposto ao agente a cada passo de tempo consiste puramente em um dicionário contendo o par ordenado de nós origem e destino da demanda corrente, representado matematicamente como um vetor discreto multidimensional $\vec{s} = [n_{\text{origem}}, n_{\text{destino}}]$, em que $n \in \{0, 1, \dots, N_{\text{nós}} - 1\}$.
- Espaço de Ações ($\mathcal{A}$): O espaço de ações é binário e unitário, definido por $\mathcal{A} = \{a\}$. A ação $a = 0$ determina que o plano de controle execute a alocação espectral utilizando a heurística R-SA. A ação $a = 1$ comuta o comportamento da rede para a heurística SA-R.
- Função de Transição ($\mathcal{P}$): A dinâmica de transição de estados é governada pela chegada estocástica de requisições de chamadas seguindo um processo de Poisson com taxa de chegada λ correlacionada à carga da rede em Erlangs, enquanto o tempo de retenção das conexões ativas apresenta uma distribuição exponencial de taxa $\mu = 1$.
- Função de Recompensa ($\mathcal{R}$): O sinal de feedback imediato quantifica o sucesso da política de alocação. A cada requisição que chega à rede atribui-se uma recompensa positiva +1 caso os slots espectrais contíguos e contínuos sejam reservados com sucesso em uma das rotas candidatas, e uma penalidade de -5 na ocorrência de bloqueio.
- Fator de Desconto (γ): Define a magnitude de importância das recompensas futuras em relação às imediatas no cálculo do retorno esperado. Configurou-se o valor $\gamma = 0.995$, induzindo o agente a mitigar ações míopes e priorizar a saúde espectral da rede no longo prazo para evitar cascatas de bloqueio.

B. Heurísticas de Alocação de Recursos Base: R-SA e SA-R

Independentemente da ação tomada pelo agente de DRL, o cálculo de caminhos baseia-se no algoritmo de Yen para determinação das K -rotas mais curtas em número de saltos (K-Shortest Paths). A atribuição de blocos de frequência adota a política First-Fit (FF) para garantir a compactação espectral. O diferencial entre as políticas reside na ordem de precedência dos laços de busca executados em matrizes vetorizadas:

- Mecanismo R-SA ($a = 0$): Avalia sequencialmente as rotas ordenadas pelo comprimento do caminho. Para cada rota examinada, busca-se o primeiro índice de slot livre capaz de acomodar a totalidade da banda requisitada. Essa heurística prioriza caminhos curtos e minimiza a ocupação física de enlaces.
- Mecanismo SA-R ($a = 1$): Inverte a hierarquia de busca estabelecendo um laço externo sobre os índices globais de slots espectrais disponíveis na rede. Para um determinado slot analisado, o algoritmo varre as rotas candidatas à procura da primeira correspondência livre. Essa abordagem favorece o balanceamento de carga e a preservação da continuidade espectral ao longo da topologia.

C. Agente de Otimização Baseado em Proximal Policy Optimization (PPO)

A otimização utiliza o Proximal Policy Optimization (PPO) sob uma arquitetura Ator-Crítico. O algoritmo maximiza uma função objetivo com corte (clipped surrogate objective) para impedir atualizações destrutivas de gradiente que desestabilizam a convergência. O modelo baseia-se no cálculo do estimador de vantagem generalizada (GAE) para computar de forma robusta o impacto de longo prazo das decisões tomadas curva a curva no ambiente de simulação dinâmico.

D. Otimização de Hiperparâmetros via Busca Automatizada

O desempenho e a estabilidade de convergência de modelos baseados em Aprendizado por Reforço Profundo são severamente influenciados pela parametrização dos algoritmos e pela topologia das redes neurais. Para mitigar vieses decorrentes de escolhas manuais por tentativa e erro, este trabalho adota uma abordagem de ajuste sistemático e automatizado de hiperparâmetros utilizando o framework de otimização de software Optuna [8].

O processo de sintonia fina baseia-se na execução de múltiplos ensaios (trials) independentes, nos quais o algoritmo mapeia e explora um espaço de busca multidimensional. Os hiperparâmetros submetidos ao escrutínio do algoritmo de otimização, bem como os parâmetros estruturais do modelo, englobam:

- Parâmetros de Atualização do Gradiente: A taxa de aprendizado (learning rate), o limite de corte do objetivo substituto (clip range) e a quantidade de épocas de otimização executadas por minilote (n_epochs).
- Regularização e Estocasticidade: O coeficiente de entropia (ent_coef), responsável por gerenciar o balanço entre a exploração de novas estratégias ópticas e o aproveitamento de caminhos conhecidos.
- Estrutura Temporal e de Amostragem: O fator de desconto (gamma), o parâmetro de suavização do estimador de vantagem (gae_lambda), o horizonte de passos por ambiente (n_steps) e a dimensão do minilote de processamento (batch size).
- Capacidade Arquitetônica e Função de Valor: A arquitetura base da rede neural (policy), definida como uma rede perceptron multicamadas (Padrão), e o coeficiente da

função de valor (vf_coef), que ajusta o impacto do erro da rede do Crítico na função de perda total.

A função objetivo do algoritmo de busca visa a minimização da Probabilidade de Bloqueio (PB) obtida pelo modelo. Para neutralizar o risco de convergência para mínimos locais e evitar o overfitting a um padrão de tráfego específico, implementou-se um mecanismo de avaliação multissemente assíncrono. Cada política candidata gerada pelo framework é submetida a testes em ambientes inicializados com sementes pseudoaleatórias inéditas e distintas daquelas utilizadas no processo de treinamento. Esse rigor garante que a métrica de aptidão retornada ao otimizador reflita a real capacidade do agente PPO em generalizar suas decisões sequenciais de roteamento sob regimes dinâmicos e flutuações sazonais de tráfego.

Os valores ideais encontrados ao final do processo de otimização estão consolidados na Tabela I. Com o intuito de acelerar o treinamento e a coleta de dados, a simulação foi executada utilizando processamento paralelo vetorial distribuído em múltiplos ambientes simultâneos (N núcleos), rodando diretamente em processamento central (device CPU). Por fim, o treinamento encerra-se ao atingir a estabilidade da política, definida por um limite máximo de amostras processadas (total_timesteps).

III. SIMULAÇÃO E RESULTADOS

Para avaliar o desempenho do mecanismo de seleção de ordenação baseado em Aprendizado por Reforço (PPO), desenvolveu-se um simulador de eventos discretos integrado à interface do Gymnasium. Os resultados obtidos pelo agente PPO foram comparados com as heurísticas estáticas individuais (R-SA e SA-R) e com o modelo de seleção estática gerado pelo Algoritmo Genético (HRSA) proposto na literatura [2].

A. Cenário de Simulação e Parâmetros

Os testes de validação foram conduzidos considerando duas topologias de rede de referência com características estruturais distintas: a NSFNet (14 nós e 21 enlaces bidirecionais) e a rede Abilene (11 nós e 14 enlaces bidirecionais). A NSFNet apresenta um grau médio nodal mais elevado, oferecendo maior diversidade de rotas alternativas, enquanto a Abilene possui uma conectividade mais restrita, o que impõe desafios mais severos de continuidade espectral e propensão a gargalos.

O tráfego de rede foi modelado de forma dinâmica, em que a chegada de requisições de conexão segue um processo de Poisson, e o tempo de retenção (duração da chamada) é governado por uma distribuição exponencial. Para cada enlace da topologia, assumiu-se uma capacidade total de 320 slots espectrais. As conexões geradas requisitaram bandas uniformemente distribuídas pertencentes a classes de demanda variando entre 6 e 12 slots de frequência ($C \in \{6, 7, 8, 9, 10, 11, 12\}$). O roteamento físico baseou-se no cálculo prévio dos $K = 3$ caminhos mais curtos (K-Shortest Paths) para cada par origem-destino. Com o intuito de garantir a estabilidade estatística dos resultados, cada episódio de simulação processou um total de 100.000 requisições sequenciais. As simulações foram executadas variando a carga total da rede (em Erlangs) para

submeter os algoritmos a diferentes regimes de estresse espectral e fragmentação em ambas as topologias.

TABELA I: Parâmetros de treinamento do RL

	Abilene	NSFNet
N núcleos	16	16
learning_rate	$3,46 \cdot 10^{-4}$	$3,71 \cdot 10^{-4}$
n_steps	8192	8192
batch_size	4096	4096
n_epochs	5	10
gamma	0,9903	0,9906
gae_lambda	0,9438	0,9654
clip_range	0,1	0,1
ent_coef	$5,12 \cdot 10^{-3}$	$5,75 \cdot 10^{-3}$
vf_coef	0,5	0,5
total_timesteps	100007936	58982400
policy	Padrão	Padrão
device	cpu	cpu

B. Análise da Probabilidade de Bloqueio (PB)

A principal métrica de eficiência avaliada foi a Probabilidade de Bloqueio (PB), definida pela razão entre conexões rejeitadas e o total requisitado. A Figura 1 apresenta as curvas de Probabilidade de Bloqueio em função da carga da rede para as quatro abordagens avaliadas nas topologias NSFNet e Abilene.

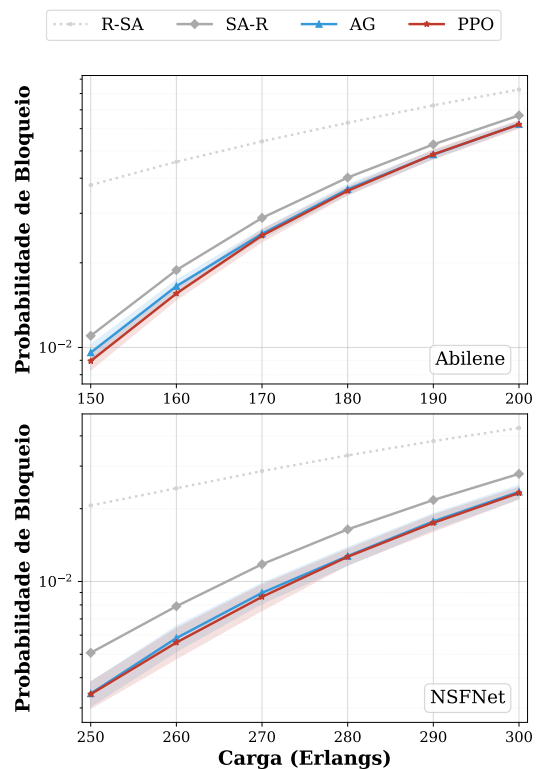


Fig. 1: Comparação da Probabilidade de Bloqueio das heurísticas.

A análise dos gráficos revela que a adoção exclusiva da heurística R-SA resulta no pior desempenho geral em ambos os cenários. Ao priorizar estritamente o menor caminho, o

R-SA tende a esgotar rapidamente os recursos dos enlaces centrais, gerando fragmentação severa sob altas cargas de tráfego. Esse efeito é ainda mais pronunciado na topologia Abilene, onde a escassez de rotas alternativas disjuntas entre si para diferentes pares origem-destino da rede amplifica o estrangulamento da rede. Por outro lado, a política SA-R, ao realizar uma varredura espectral primária, promove uma melhor compactação do uso dos slots da rede e um maior espalhamento do tráfego nos enlaces da rede, resultando em reduções consistentes na probabilidade de bloqueio em comparação ao R-SA, independentemente da topologia.

O agente PPO demonstrou a capacidade de assimilar o comportamento dinâmico da rede, superando o desempenho de ambas as heurísticas isoladas em todos os pontos de carga simulados, tanto na NSFNet quanto na Abilene. Esse resultado comprova que o modelo de Aprendizado por Reforço conseguiu mapear, com base unicamente na identificação dos pares origem-destino, uma política válida que alterna entre R-SA e SA-R de maneira mais eficiente do que a escolha arbitrária de uma única regra. O sucesso do PPO em duas topologias de naturezas distintas atesta a capacidade de generalização do algoritmo frente a diferentes restrições topológicas.

Adicionalmente, nota-se que o agente PPO alcançou um desempenho equivalente ao limite estabelecido pelo AG em ambas as redes. Esse resultado valida a robustez do aprendizado por reforço dada a diferença fundamental entre os métodos: enquanto o AG avalia a qualidade das decisões de forma global, analisando o tráfego apenas após a conclusão de toda a simulação, o agente PPO atinge o mesmo patamar de eficiência operando sob uma lógica de passo a passo, tomando decisões sequenciais e independentes a cada nova requisição de conexão.

C. Análise do Custo Computacional e Eficiência de Amostragem

A comparação de desempenho entre os métodos de otimização global (AG) e sequencial (PPO) exige a quantificação do volume de dados processados durante a fase de treinamento. Como as dinâmicas de busca operam em métricas de evolução distintas — o AG avança em lotes por gerações (Figura 2), enquanto o PPO atualiza sua política em passos temporais contínuos (Figura 3) —, o referencial utilizado para equipará-los de forma justa é o número total de interações exigidas com o ambiente. A Figura 2 ilustra inicialmente o comportamento de convergência da função de aptidão (fitness) para o Algoritmo Genético nas duas topologias avaliadas.

Para a rede Abilene, o processo evolutivo foi concluído na geração 129, empregando populações de 100 indivíduos. Na topologia NSFNet, a convergência deu-se em 145 gerações com 100 indivíduos. Considerando que a avaliação de aptidão de cada indivíduo exige o processamento de um episódio completo de simulação (10^5 requisições), o processo evolutivo consumiu aproximadamente 1,29 bilhões e 1,45 bilhões de interações absolutas na Abilene e NSFNet, respectivamente.

Em contrapartida, conforme detalhado na Tabela I de hiperparâmetros, o treinamento do agente PPO atingiu a estabilização da política após processar cerca de 100 milhões e 59 milhões de passos de tempo em cada rede. A comparação

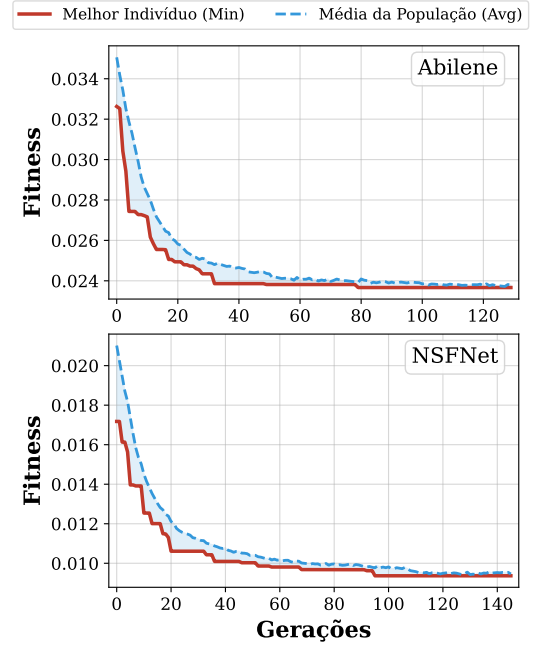


Fig. 2: Evolução da fitness média ao longo das gerações do Algoritmo Genético nas topologias Abilene e NSFNet.

direta desses volumes absolutos de requisições processadas indica que a arquitetura de DRL obteve resultados de bloqueio equivalentes consumindo um volume de amostras de dados estatisticamente menor do que o algoritmo evolutivo puro, evidenciando uma maior eficiência na exploração do espaço estatístico de tráfego.

Além da eficiência de amostragem durante a fase de treinamento, é fundamental destacar a superioridade do agente PPO quanto aos requisitos de memória e tempo de inferência em tempo de execução (online). Enquanto a abordagem via AG exige o armazenamento de populações inteiras em memória e a reexecução custosa da simulação sempre que a topologia ou matriz de tráfego sofre alterações, o modelo PPO treinado consiste em uma rede neural profunda leve (pesando poucos megabytes). Na prática, a tomada de decisão do agente para uma nova requisição requer apenas uma propagação direta (forward pass) na rede neural, resolvendo a escolha entre R-SA e SA-R na ordem de milissegundos ou microssegundos.

D. Dinâmica de Aprendizado, Convergência e Interpretabilidade da Política

Para compreender a evolução da política do agente DRL e validar a estabilidade do processo de aprendizado, avaliaram-se as métricas de convergência extraídas durante o treinamento do algoritmo. A Figura 3 agrupa o comportamento do treinamento, apresentando de forma conjunta a evolução da recompensa episódica média (ep_rew_mean) e a divergência estatística da política (approx_kl) ao longo de milhões de passos de tempo (timesteps).

Analisando a curva de recompensa média, observa-se uma fase de convergência acelerada nos primeiros 30 milhões de passos. Contudo, após atingir seu pico de desempenho,

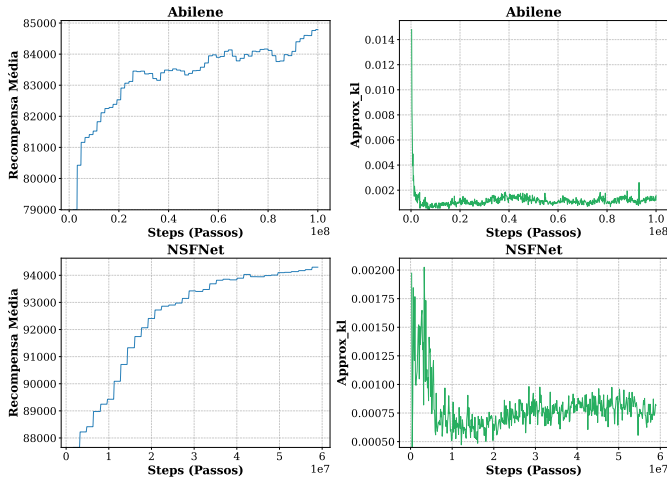


Fig. 3: Evolução da recompensa média e approx_kl.

a curva de aprendizado exhibe ligeiras oscilações contínuas. Esse fenômeno é um reflexo direto da alta estocasticidade do ambiente simulado, onde as flutuações dinâmicas de chegada e partida de tráfego injetam um ruído natural na avaliação episódica. Apesar dessa variação nos retornos, o gráfico de divergência de Kullback-Leibler (approx_kl) ilustra uma queda abrupta inicial seguida por uma estabilização consistente com variações contidas. Esse comportamento atesta a eficácia da função objetivo do PPO, que restringe atualizações abruptas nos pesos da rede neural e garante que a política de roteamento aprendida permaneça estável frente ao ruído do ambiente.

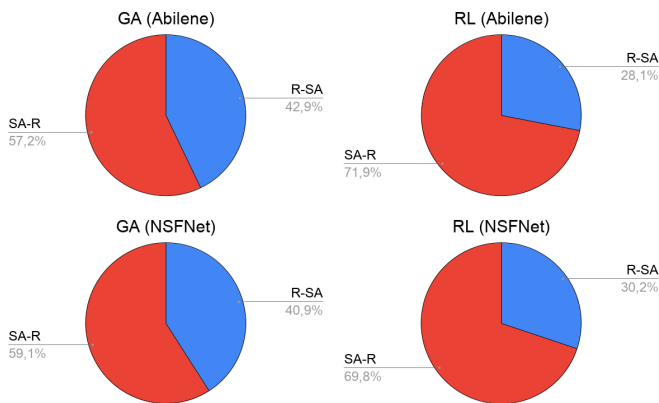


Fig. 4: Distribuição percentual das escolhas de ordenação (R-SA e SA-R) realizadas pelo agente PPO nas topologias NSFNet e Abilene.

A Figura 4 apresenta o comparativo da distribuição percentual das ações executadas pelo modelo treinado, evidenciando a frequência de uso das heurísticas R-SA e SA-R nas topologias NSFNet e Abilene. O gráfico mostra a capacidade de adaptação estrutural do agente DRL. O modelo não converge para uma solução trivial (como a escolha exclusiva de uma única heurística), mas sim calibra a proporção de uso entre R-SA e SA-R e em que pares de nós dispõem a depender das restrições topológicas que enfrenta. Diferenças no grau de conectividade

nodal e na disponibilidade de rotas alternativas entre a NSFNet e a Abilene forçam o agente a rebalancear sua preferência, priorizando a mitigação de gargalos específicos de cada rede. Essa capacidade de ajustar o balanço entre conservação de saltos (R-SA) e organização espectral e distribuição de carga (SA-R) comprova que o PPO conseguiu inferir e explorar a correlação entre as estruturas de roteamento puramente a partir dos pares de origem e destino avaliados.

IV. CONCLUSÕES

Este trabalho investigou o uso do algoritmo PPO para a seleção sequencial de heurísticas RSA em EONs, utilizando uma modelagem restrita aos pares origem-destino. Os resultados experimentais indicam que o agente de DRL superou a aplicação exclusiva das heurísticas R-SA e SA-R nas topologias NSFNet e Abilene, alcançando níveis de probabilidade de bloqueio equivalentes aos obtidos pela otimização global do Algoritmo Genético. Além do pareamento em desempenho, o modelo de aprendizado por reforço demonstrou uma eficiência de amostragem superior, convergindo com uma demanda de interações com o simulador de uma a duas ordens de grandeza menor do que o processo evolutivo. Como trabalho futuro, planeja-se expandir o espaço de observação do agente para incorporar variáveis dinâmicas do ambiente óptico, tais como a taxa de ocupação instantânea dos enlaces e vetores locais de disponibilidade espectral, permitindo tomadas de decisão cientes da fragmentação em tempo real.

AGRADECIMENTOS

Gostaríamos de agradecer à FACEPE e ao CNPq o apoio financeiro concedido, bem como à UFPE e à UFABC o apoio institucional.

REFERÊNCIAS

- [1] Zarea, S., Khonsari, A., Rahbar, A. G., Moradian, M. Reinforcement learning-driven multicast routing in Elastic Optical Networks: A multi-objective cost optimization framework. *Optical Fiber Technology*, 95, 104444. 2025.
- [2] Dinarte, H. A., Correia, B. V. de A., Chaves, D. A. R., Almeida Jr., R. C. Routing and spectrum assignment: A metaheuristic for hybrid ordering selection in elastic optical networks. *Computer Networks*, 197, 108287. 2021.
- [3] Bermúdez, J. A., Morales, P., Pempelfort, H., Araya, M., Jara, N. Understanding deep reinforcement learning: Enhancing explainable decision-making in optical networks. *ICT Express*, 11(6), 969-973. 2025.
- [4] Xu, L., Huang, Y.-C., Xue, Y., Hu, X. Hierarchical Reinforcement Learning in Multi-Domain Elastic Optical Networks to Realize Joint RMSA. *Journal of Lightwave Technology*, 41(8), 2276-2287. 2023.
- [5] Morales, P., Franco, P., Lozada, A., Jara, N., Calderón, F., Pinto-Ríos, J., Leiva, A. Multi-band Environments for Optical Reinforcement Learning Gym for Resource Allocation in Elastic Optical Networks. In *International Conference on Optical Network Design and Modeling (ONDM)*. IEEE. 2023.
- [6] Prakash, V., Katiyar, S., Rai, R. K., Chatterjee, B. C. PPO-RMSA: Proximal Policy Optimization-Based Routing, Modulation, and Spectrum Allocation in Elastic Optical Networks. In *25th Anniversary International Conference on Transparent Optical Networks (ICTON)*. IEEE. 2025.
- [7] Towers, M., Terry, J. K., Kwiatkowski, A., Balis, J. U., Cola, G. d., Deleu, T., Goulao, M., Kallinteris, A., KG, A., Krimmel, M., Perez-Vicente, R., Pierre, A., Schulhoff, S., Tai, J. J., Shen, A. T. J., and Younis, O. G. Gymnasium. *Advances in Neural Information Processing Systems* 38. 2025.
- [8] Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019.