Multi-band Environments for Optical Reinforcement Learning Gym for Resource Allocation in Elastic Optical Networks

Patricia Morales, Patricia Franco, Astrid Lozada, Nicolás Jara
Department of Electronics Engineering,
Universidad Técnica Federico Santa María
Valparaíso, Chile
patricia.morales@sansano.usm.cl

Felipe Calderón, Juan Pinto-Ríos, Ariel Leiva
School of Electrical Engineering,
Pontificia Universidad Católica de Valparaíso,
Valparaíso, Chile

***Abstract*—The use of additional fiber bands for optical communications -known as Multi-band or Band-division multiplexing (BDM) - allows to increase the traffic served in transparent optical networks. In recent years, many proposals have emerged as a solution for resource allocation in such multi-band architectures. This work presents a novel approach based on reinforcement learning (RL) techniques to accommodate multi-band elastic optical network resources. Two new environments were implemented and added to the Optical-RL-Gym toolkit considering four scenarios with different band availability. Six agents were tested in four real network topologies, contrasting their episode rewards on a large number of training steps. Results show Trust Region Policy Optimization (TRPO) as the best performing agent, with consistent output across all the scenarios and network topologies considered. In addition, we illustrate the blocking probability behavior in relation to the traffic load, and band usage distribution, allowing further discussions.**

***Index Terms*—multi-band optical networks, resource allocation, reinforcement learning**

I. Introduction

Due to the ever-growing Internet traffic, the maximum capacity of individual fibers will soon be reached, leading to what is commonly referred to as *capacity crunch* [1]. As a consequence, the business model of communication systems worldwide claims for new solutions to eke out the fiber capacity [2]. Flexibility on optical spectrum allocations in Elastic Optical Networks (EONs) architectures [3] and the expansion of C-band communications to C+L+S+E bands in Band-Division Multiplexing (BDM) technologies [4] offer complementary solutions to this impending problem, which in turn leads to new challenges to solve. Among them, solving the optical connection (lightpath) establishment problem in optical networks has aroused the interest of the scientific community to a large extent. In elastic optical networks, the resource assignment involves solving the routing, modulation level, and spectrum allocation (RMLSA) problems [3]. In this regard, to add a new dimension that corresponds with the band, the well-known RMLSA problem becomes the Routing, Band, Modulation Format, and Spectrum Allocation (RBMLSA) problem in multi-band EONs.

Recent proposals [5], [6] solve the RBMLSA problem through heuristics approaches, using an algorithm based on the degradation of Generalized optical Signal-to-Noise-Ratio (GSNR). Authors consider a set of available bands where these are used sequentially. They compile several scenarios considering C, L, S, E, and O bands to validate the results. Furthermore, [7] investigates the practical benefits of multi-band optical networks leveraging distance-adaptive resource allocation. Results show that the S+C+L-band scenario can successfully accommodate three times more traffic than a single-band scenario, even on a large-scale network. In the case of [8], [9] authors analyzed multi-core and multi-band solutions, concluding that both multiply the network capacity by the number of cores or bands available, making the performances of the two solutions both comparable and compatible.

On another note, Artificial Intelligence (AI) techniques have brought a fresh and potential vision when facing optical network-related issues [10]. In that sense, Reinforcement Learning (RL) techniques have shown cutting-edge performance in large-scale control tasks. As a result, the use of RL for solving different resource allocation problems in optical networks has started to attract significant research efforts [11]. In this context, Chen et.al. [12], [13] implemented an RL-based framework for Routing, Modulation Level, and Spectrum Assignment (RMSA) in EONs achieving blocking reduction compared to the commonly used KSP-FF (k-Shortest Path for routing, First-Fit allocation policy for spectrum slots) heuristic approach. This work was extended and included in an open-source toolkit published by Natalino and Monti [14], known as Optical RL-Gym. The toolkit, among its many capabilities, allows the application of RL techniques to solve optical network resource allocation by providing EON and wavelength-routed environments. However, none of these environments implement multi-band EON architectures, refraining to solve the Routing, Band, Modulation Level, and Spectrum Assignment (RBMLSA) problem under dynamic operation.

In this paper, we report two new environments developed

978-3-903176-33-1

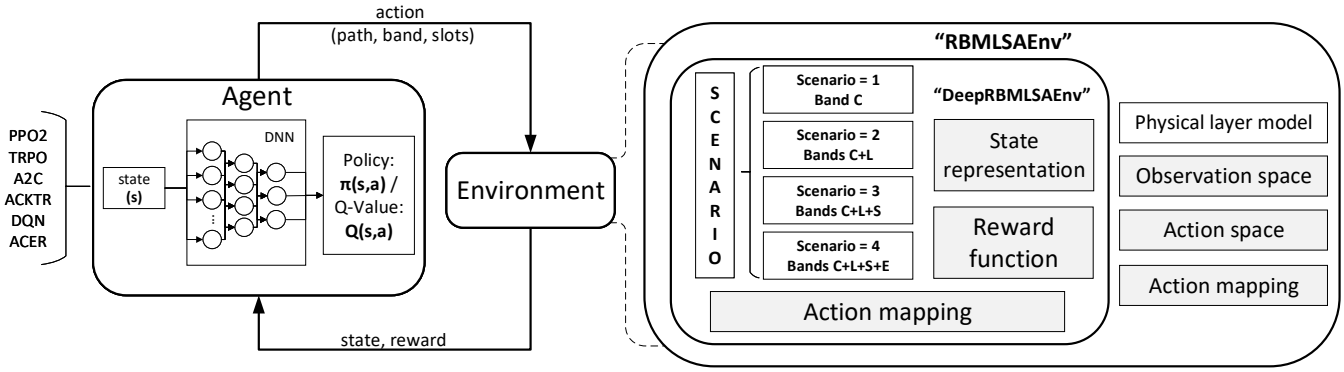


Fig. 1. Schematic of the developed RL framework. Left: the main components and their interactions. Right: a breakdown of the new two environments and their principal entities.

for Optical RL-Gym: `RBMLSAEnv` and `DeepRBMLSAEnv` [1], proposed to solve the RBMLSA problem. The system comprises scenarios with a different number of available bands. To demonstrate our new environments' potential, we contrast the performance of six agents working in the scenario with the higher number of bands (C+L+S+E bands available). We perform this experiment in four real network topologies. Using the best performing agent, we analyze the blocking probability behavior for a set of traffic loads with a large number of training steps for the implemented scenarios. Also, the band usage distribution is shown for the four topologies considered. Therefore, we expect this implementation helps the research community widen the range of techniques to successfully solve the resource allocation problem in dynamic elastic optical networks.

II. Multi-band environments

The Optical-RL-Gym [14] is a four-level hierarchy of environments following the principles established by the OpenAI Gym [15], the *de-facto* standard for RL environments. Analogously to the `RMSAEnv` and `DeepRMSAEnv` environments already included in Optical RL-Gym, we incorporate `RBMLSAEnv`, which includes the functionalities related to the RBMLSA problem, and `DeepRBMLSAEnv`, which implements the specific use-case to solve the RBMLSA problem in dynamic multi-band EONs. Figure 1 shows an overview of the system functionalities. The left side shows the two main components (agent and environment) and their interaction. The agent is the central axis of an RL model since it is in charge of the decision-making process, sending commands to the environment in the form of actions. Thus, we can find two classes or types of RL algorithms:

1) Q learning: the network learns the Q function, refraining to pass it as ground truth. Thus, the highest Q value is chosen and the corresponding action is executed.
2) Policy learning: the algorithm directly learns a policy, avoiding the use of the Q function as intermediary.

[1]The new environments presented in this paper are available at: https://gitlab.com/IRO-Team/optical-rl-gym-multiband/

An agent takes as input the different states "s" returning the Q values (in Q learning) or a probability distribution (in case of policy learning), for each possible action "a" to be executed on a given state. Thus, the action with the highest Q value or probability is chosen.

On the other hand, we encapsulate an event-driven simulator in the environments. Two events are essential: the request arrival and the departure of provisioned demands. However, the agent only decides in the first event: the request arrival. Upon request reception, the agent selects an action defined by a path, selected out of k possible pre-computed routes, the band, and the identification of the first j available slots-block to allocate the request. Such selection is then sent to the environment. When the environment receives the agent's action, it selects the best modulation format, and the corresponding number of slots required. Next, it checks whether the request can be established and - based on the request establishment's success - sends back an observation on the new system state and a reward. If the request is accepted, the reward is equal to 1. Otherwise, it is equal to -1. Requests are rejected if there are no enough spectrum resources in the path and band selected by the agent, i.e., if the request can only be established using slots from two contiguous bands, it is also rejected.

The observation sent back to the agent uses the same structure proposed in [12], extended to include information of the different bands. The number of available bands ranges from 1 (C band only) to 4 (C, L, S, and E bands), represented by *Scenario* input variable, as shown on the right side of Figure 1. The observation method is therefore responsible for building the representation of the current state of the network. Such representation is the one that will be presented to the agent. In our environments, the representation is usually composed of how the resources are currently allocated (free/used) in the network, e. g, source/destination/bitrate of the request. Although each of these different components is created following their own shape (usually a matrix), in the end, these are concatenated together and reshaped as a vector, following neural networks common input structure.

In this work, the impact of physical impairments on the

TABLE I
MAXIMUM ACHIEVABLE REACH (MAR) PER MODULATION FORMAT, FOR A BER_{TH} VALUE EQUAL TO $4.7 \cdot 10^{-3}$.

Modulation	Net Bit-rate [Gb/s]	Maximum achievable reach [# spans]									
		Scenario 1	Scenario 2		Scenario 3			Scenario 4			
		C	C	L	C	L	S	C	L	S	E
BPSK	23	199	197	167	174	167	148	130	144	102	31
QPSK	46	99	99	84	87	84	74	65	72	51	15
8-QAM	69	54	54	46	47	46	41	35	39	29	9
16-QAM	92	27	14	22	23	22	20	17	19	14	4
32-QAM	115	13	13	11	12	11	10	8	9	7	2
64-QAM	140	7	7	6	6	6	5	4	5	3	1
256-QAM	186	1	1	1	1	1	1	1	1	0	0

quality-of-transmission of an optical route is taken into account by determining the maximum reach of optical signals as a function of the modulation format for a given bit-error-rate threshold (BER_{th}), as proposed in [16]. Tables I and II shows the parameters considered in this work based on [16]. In the first column of table I are listed the modulation formats available for each optical communication. The second column of this table refers to the bit-rate capacity available on a single slot for each modulation format (net Bit-rate). The rest of table I shows the maximum achievable reach (MAR) in number of spans (1 span = 100 km) for the available bands in the four possible scenarios considered. Consequently, based on this table, a modulation format assigned will be the one that best accommodates the request's maximum achievable distance. In the case of table II, it shows the total number of slots per band considered.

TABLE II
TOTAL SLOTS PER BAND

Band	Frequency (THz)	Bandwidth (BW)	Slots (BW/12.5 GHz)
L	185.7 -191.7	6	480
C	191.7 - 196	4.3	344
S	196 - 205.5	9.5	760
E	205.5 - 219.7	14.2	1136
Total	185.7 - 219.7	34	2720

III. AGENT TRAINING AND PERFORMANCE EVALUATION

Six different agents from `stable_baselines` library [17] were studied: Proximal Policy Optimization version 2 (PPO2), Trust Region Policy Optimization (TRPO), Synchronous Advantage Actor-Critic (A2C), Actor Critic using Kronecker-Factored Trust Region (ACKTR), Deep Q Learning (DQN) and Actor-Critic with Experience Replay (ACER). For every agent, $2 \cdot 10^3$ training steps were configured with an episode length of 50. Thus, a total of 10^5 time steps were carried out. The learning rate and discount factor were set to 10^{-5} and 0.95, respectively. This parameters allow computing the discounted reward which reflects the agents' behavior through learning curves. The discounted reward is obtained after completing an episode as the summation of multiplying the discount factor and the action reward in each time step.

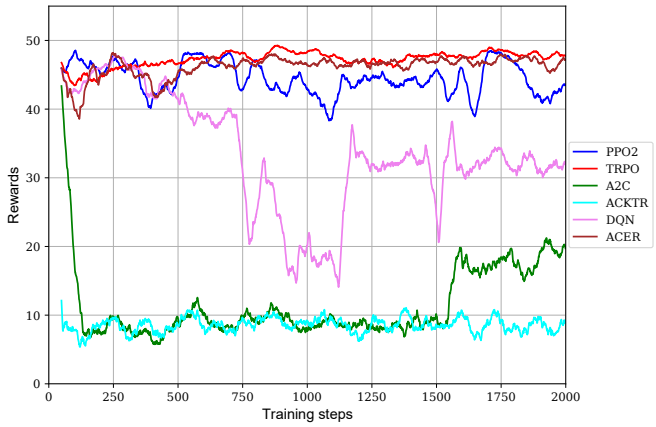
TABLE III
NETWORK TOPOLOGIES PARAMETERS USED IN THIS WORK

Topology	Nodes	Links	node-pairs
NSFNet	14	42	82
UKNet	21	78	420
USNet	46	152	2070
Eurocore	11	50	110

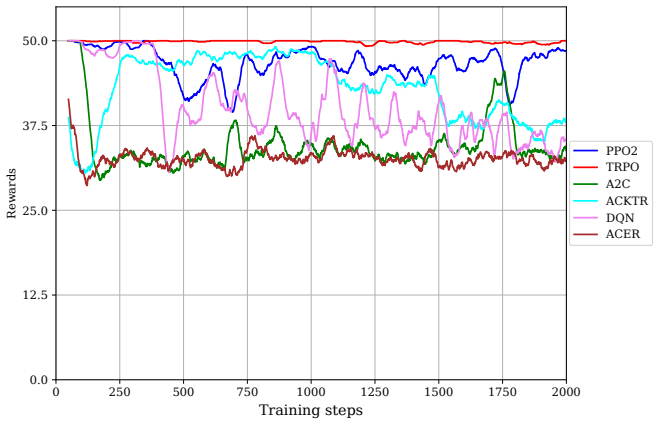
We execute our environmnents for the *NSFNet*, *Eurocore*, *USNet* and *UKNet* networks. Table III contains the number of nodes, links and node-pairs requesting communication for each topology. In case of the *NSFNet* the configuration corresponds to the one described in [12].

The number of possible routes of node-pairs requesting communication were set to 5 (k=5) and the j value equal to 1, meaning that the agent will choose 1 of 5 routes and the first suitable slots-block. The maximum achievable reach for each modulation format and the spectrum capacity available on the different bands were computed based on [16] using the parameters in Table I for a BER_{th} of $4.7 \cdot 10^{-3}$ and a slot spectral width of 12.5 GHz. The modulation formats considered in this study are binary phase-shift keying (BPSK), quadrature phase-shift keying (QPSK), and Λ-quadrature amplitude modulation (Λ-QAM), where Λ takes the values 8, 16, 32, and 64 (first column of Table I). Bit rates are randomly selected among 10, 40, 100, 400, and 1000 Gbps, and the connection requests are uniformly distributed among all node pairs. As an example, if a request with 5000 km distance and 1000 Gbps bit-rate arrives in scenario 2, the modulation formats that best suit the request's requirements will be 8-QAM (from 1400 to 5400 km) in the C band and QPSK (from 4600 to 8400 km) in the L band. The number of slots necessaries in the C band will be given by the division of the *Bit-rate* and *Net Bit-rate* of the corresponding modulation format: $\frac{1000}{69} \approx 15$ slots in this case. On the contrary, in the L band, the number of slots will be $\frac{1000}{46} \approx 22$ slots.

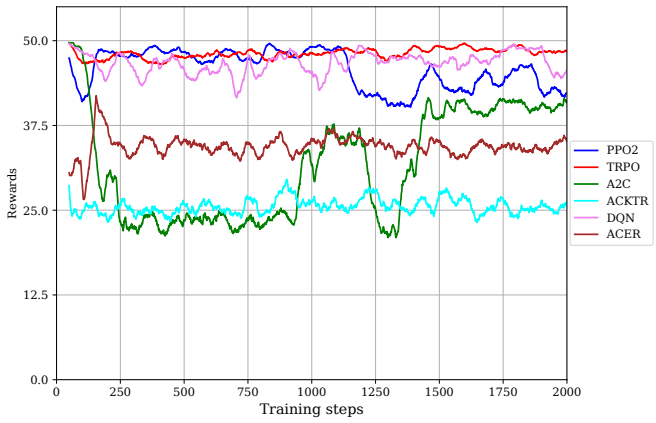
The arrival requests are modeled as a Poisson process, with a mean arrival rate equal to λ. Mean holding times are exponentially distributed with an average of $1/\mu$ = 200s. Different traffic loads - equal to λ/μ - are obtained by varying the arrival rate.



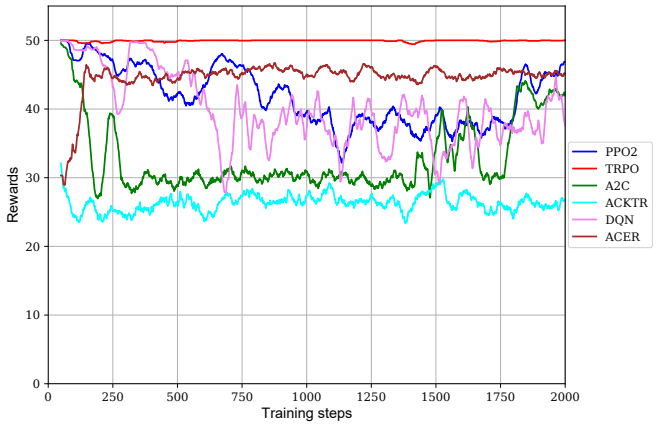
(a) Episode Rewards vs Training Steps (NSFNet)



(b) Episode Rewards vs Training Steps (Eurocore)

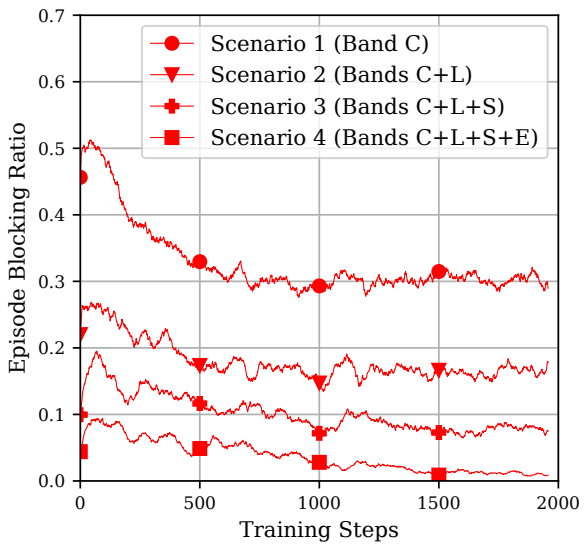


(c) Episode Rewards vs Training Steps (USNet)

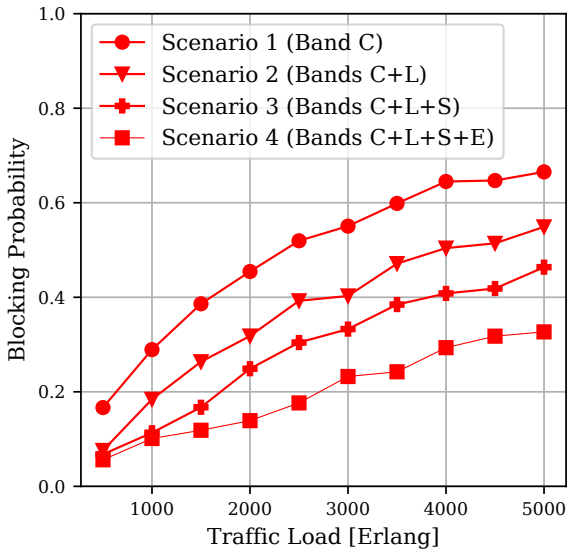


(d) Episode Rewards vs Training Steps (UKNet)

Fig. 2. Reward accumulated by the six agents in Scenario 4 with a traffic load = 1000 Erlang.



(a) Episode Blocking Ratio vs Training Steps



(b) Blocking Probability vs Traffic Load

Fig. 3. (a) Blocking ratio of episodes versus the number of training steps exhibited by the best performing agent (TRPO) for the four scenarios, (b) Blocking probability obtained by TRPO in all scenarios as a function of the traffic load.

Figure 2 shows the learning curves obtained when training six different agents in Scenario 4 (C+L+S+E bands). This scenario corresponds to the largest action space. Therefore, a high discounted reward is expected after completing all training steps. The curves were obtained for a traffic load of 1000 Erlang. Although results show that almost all agents' performance varies on the different topologies, consistently TRPO maintains a more stable behavior, translating into a less fluctuating reward. TRPO achieves the best results converging faster than the five remaining agents to a value of reward around 50. This performance can be related to the monotonic improvement that TRPO guarantees, optimizing the policies that the agent acquires in the environment [18]. We can also highlight PPO2, which in some cases reaches the highest value (50) of reward but with a fluctuating behavior. In opposition, the ACKTR agent exhibits inferior performance in all network topologies. Consequently, the best agent (TRPO) was then used in the following experiments.

Figure 3.a shows the episode blocking ratio's performance of the TRPO agent as a function of the training steps for the different scenarios under a traffic load of 1000 Erlang in the *NSFNet*. Remark that the scenarios 1 to 4 are defined in terms of band availability, thus Band C, Band C+L, Band C+L+S, Band C+L+S+E respectively. As expected from the results shown in Figure 2, the higher the reward, the lower the blocking probability value. As the number of training steps increases, the episode blocking probability diminishes, converging to a low value, which is expected behavior for an appropriate learning process. Furthermore, if the number of available bands increases, the blocking probability decreases due to the larger network capacity. The most significant blocking reduction is obtained when migrating from the C-band to the scenario C+L bands available.

To test the robustness of the two proposed environments, we studied the network blocking performance for different traffic loads ranging from 500 to 5000 Erlang in steps of 500 in *NSFNet* network topology. The results are illustrated in Figure 3.b as it can be seen, for the same traffic load, as the number of available bands increases (i.e., network capacity increases), the blocking probability decreases significantly. For example, for 5000 Erlang, the blocking probability for Scenario 1 (C band only), exceeds 0.66. This situation means that more than half of the requests are being blocked. The situation changes as the number of bands increases, reaching a blocking probability of 0.32 in Scenario 4. On the other hand, for the same scenario and the same amount of available bands, the blocking probability increases with the traffic load. This performance is also reasonable since higher traffic translates into a higher system occupation and less available spectrum resources for incoming connection requests. Current analysis demonstrates that exploiting multi-band transmissions strongly reduces the blocking probability, as shown by previous researches [5], [6], [19].

Notice that both experiments shown in Figure 3 were also performed in the three remaining topologies displayed in Figure 2. However, these results only remark the facts described above. Given this situation, in addition to the lack of space, we prefer to keep the results out of the scope of this paper.

Finally, Figure 4 shows the Band Usage Distribution (BUD) for Scenario 4 on the four network topologies illustrated in Figure 2, for a traffic load of 1000 Erlang. This figure indicates how the agent distributes the use of each band along the complete training process. The given percentage is computed following Eq. 1. The numerator corresponds with the number of times a given band is assigned over the total assigned requests (denominator).

$$\text{BUD } (\%) = \frac{\text{Request per Band}}{\text{Total Assigned Requests}} \cdot 100 \qquad (1)$$

As we can see in Figure 4, the use of the bands is more balanced in the *USNet* network topology. As an interesting fact, this corresponds with the largest topology studied in this work in terms of nodes, links and node-pairs requesting communication.

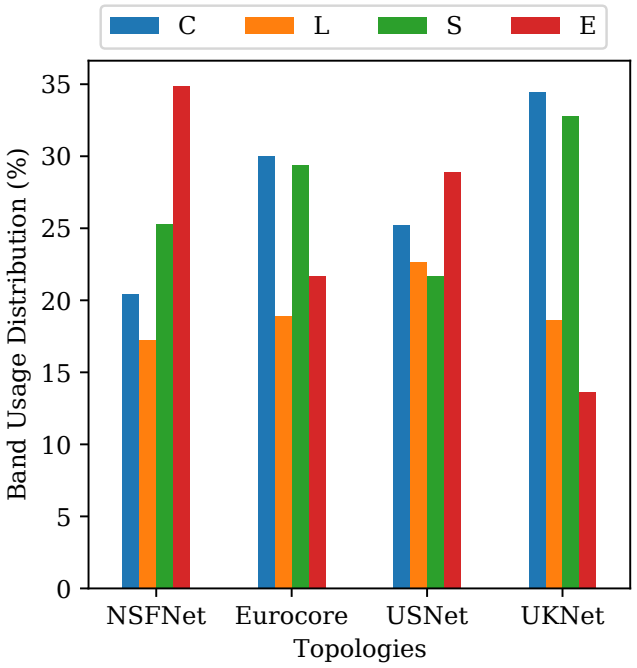


Fig. 4. Band Usage Distribution

IV. Final Remarks and Future Work

In this work, two new environments were implemented and added to the Optical RL-Gym toolkit to solve the routing, band, modulation level, and spectrum assignment problem in dynamic EONs. Results show the environments' feasibility, as a significant blocking reduction was obtained with the number of available bands represented by the scenario performed. The results evidence TRPO as the agent with more stable performance across the different network topologies tested. The given analysis shows that RL techniques can achieve

good results in terms of blocking probability and become an attractive solution to solve the RBMLSA problem. The variability in the action space has always been a cause of a problem in network-related RL. This happens since not all actions are available, unlike other environments where all actions are always available. Composing the observation vector is challenging because it should be as simple as possible (i.e., have the lowest dimension possible) but still representing the state of the network. Working with multiple bands expands the observation space significantly. When the observation space increases, this usually needs to be combined with more neurons/layers in the neural network architecture. As discussed, since some options are not available all the time, it is interesting to define an action space that is as simple as possible but still allows the RL agent to make "intelligent" decisions. The neural network architecture (number of layers/neurons at each layer) usually needs to follow the complexity of the observation/action space. Therefore, if the observation/action space increases, a more complex neural network structure is likely to be needed.

Future work will explore optimizing the system performance, expanding the number of training topologies, performing a sensitivity analysis on the training hyper-parameters, and comparing the RL results with those obtained via different heuristic algorithms [14], [20]–[22]. Since the objective of this work was to present a novel tool for allocating resources in EONs, the reward function remains unvaried with respect to the environments presented in [14]. On going research is studying a way of improving the reward function based on not only rewarding the allocation or not of connection requests, but instead prizing the allocation which result in the most optimized allocations, based on whatever optimization criterion is being pursued (e.g., resource utilization). This way, the agent's learning process will be also more efficient. Another modification understudy is building a system at scale, giving to the agent a shorter action space (i.e. reducing the number of slots and the other parameters instead of using the original values given by the physical layer model). A system at scale will allow to establish a fair comparison with the aforementioned state-of-the-art heuristic algorithms, in addition to reflect in a clear manner the learning process.

V. Acknowledgements

Financial support from projects USM PI LII_2020_74, STI-CAMSUD 19STIC-01, DI-PUCV 039.382/2021, y USM PIIC 021/2021, ANID 21200588 and ANID FONDECYT Iniciación 11201024 are gratefully acknowledged.

References

[1] A. D. Ellis, N. Mac Suibhne, D. Saad, and D. N. Payne, "Communication networks beyond the capacity crunch," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2062, 2016.

[2] H. Waldman, "The Impending Optical Network Capacity Crunch," *Sbfoton Conference*, pp. 1–3, 2018.

[3] B. Chatterjee and E. Oki, *Elastic Optical Networks: Fundamentals, Design, Control, and Management*. CRC Press, 2020.

[4] J. K. Fischer, M. Cantono, V. Curri, R.-P. Braun, N. Costa, J. Pedro, E. Pincemin, P. Doaré, C. Le Bouëtté, and A. Napoli, "Maximizing the capacity of installed optical fiber infrastructure via wideband transmission," in *2018 20th International Conference on Transparent Optical Networks (ICTON)*, pp. 1–4, IEEE, 2018.

[5] M. Mehrabi, H. Beyranvand, and M. J. Emadi, "Multi-band elastic optical networks: Inter-channel stimulated raman scattering-aware routing, modulation level and spectrum assignment," *Journal of Lightwave Technology*, pp. 1–1, 2021.

[6] N. Sambo, A. Ferrari, A. Napoli, N. Costa, J. Pedro, B. Sommerkorn-Krombholz, P. Castoldi, and V. Curri, "Provisioning in multi-band optical networks," *Journal of Lightwave Technology*, vol. 38, no. 9, pp. 2598–2605, 2020.

[7] M. Nakagawa, H. Kawahara, K. Masumoto, T. Matsuda, and K. Matsumura, "Performance evaluation of multi-band optical networks employing distance-adaptive resource allocation," in *2020 Opto-Electronics and Communications Conference (OECC)*, pp. 1–3, IEEE, 2020.

[8] E. Virgillito, R. Sadeghi, A. Ferrari, A. Napoli, B. Correia, and V. Curri, "Network performance assessment with uniform and non-uniform nodes distribution in c+ l upgrades vs. fiber doubling sdm solutions," in *2020 International Conference on Optical Network Design and Modeling (ONDM)*, pp. 1–6, IEEE, 2020.

[9] E. Virgillito, R. Sadeghi, A. Ferrari, G. Borraccini, and V. Curri, "Network performance assessment of c+ l upgrades vs. fiber doubling sdm solutions," in *Optical Fiber Communication Conference*, pp. M2G–4, Optical Society of America, 2020.

[10] J. Mata, I. de Miguel, R. J. Duran, N. Merayo, S. K. Singh, A. Jukan, and M. Chamania, "Artificial intelligence (ai) methods in optical networks: A comprehensive survey," *Optical switching and networking*, vol. 28, pp. 43–57, 2018.

[11] Y. Zhang, J. Xin, X. Li, and S. Huang, "Overview on routing and resource allocation based machine learning in optical networks," *Optical Fiber Technology*, vol. 60, p. 102355, 2020.

[12] X. Chen, B. Li, R. Proietti, H. Lu, Z. Zhu, and S. B. Yoo, "Deeprmsa: A deep reinforcement learning framework for routing, modulation and spectrum assignment in elastic optical networks," *Journal of Lightwave Technology*, vol. 37, no. 16, pp. 4155–4163, 2019.

[13] X. Chen, J. Guo, Z. Zhu, R. Proietti, A. Castro, and S. B. Yoo, "Deep-rmsa: A deep-reinforcement-learning routing, modulation and spectrum assignment agent for elastic optical networks," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, pp. 1–3, IEEE, 2018.

[14] C. Natalino and P. Monti, "The optical rl-gym: An open-source toolkit for applying reinforcement learning in optical networks," in *2020 22nd International Conference on Transparent Optical Networks (ICTON)*, pp. 1–5, IEEE, 2020.

[15] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[16] E. Paz and G. Saavedra, "Maximum transmission reach for optical signals in elastic optical networks employing band division multiplexing," *arXiv preprint arXiv:2011.03671*, 2020.

[17] A. Hill, A. Raffin, M. Ernestus, A. Gleave, A. Kanervisto, R. Traore, P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, "Stable baselines." https://github.com/hill-a/stable-baselines, 2018.

[18] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 1889–1897, PMLR, 07–09 Jul 2015.

[19] N. Sambo, A. Ferrari, A. Napoli, N. Costa, J. Pedro, B. Sommerkohrn-Krombholz, P. Castoldi, and V. Curri, "Provisioning in multi-band optical networks: A c+l+s-band use case," in *45th European Conference on Optical Communication (ECOC 2019)*, pp. 1–4, 2019.

[20] P. Goransson, C. Black, and T. Culver, *Software defined networks: a comprehensive approach*. Morgan Kaufmann, 2016.

[21] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[22] J. Wu, S. Chen, and X. Liu, "Efficient hyperparameter optimization through model-based reinforcement learning," *Neurocomputing*, vol. 409, pp. 381–393, 2020.