

**UNIVERSIDADE FEDERAL DE ALAGOAS**

**INSTITUTO DE COMPUTAÇÃO**

**COORDENAÇÃO DE PÓS-GRADUAÇÃO EM INFORMÁTICA**

**MASTER'S THESIS**

**DEFINING OPTICAL AMPLIFIERS GAINS USING  
REINFORCEMENT LEARNING**

**MASTER'S STUDENT**

**JOSÉ CARLOS PINHEIRO FILHO**

**ADVISOR**

**ERICK DE ANDRADE BARBOZA**

**MACEIÓ, AL**

**JUNE - 2024**

**Catálogo na fonte**  
**Universidade Federal de Alagoas**  
**Biblioteca Central**  
**Divisão de Tratamento Técnico**

Bibliotecária: Girlaine da Silva Santos – CRB-4 – 1127

P654d Pinheiro Filho, José Carlos.

Defining optical amplifiers gains using reinforcement learning / José Carlos Pinheiro Filho. – 2024.

39 f. : il.

Orientador: Erick de Andrade Barboza.

Dissertação (Mestrado em Informática.) – Universidade Federal de Alagoas. Instituto de Informática. Programa de Pós-Graduação em Informática, Maceió, 2024.

Bibliografia: f. 36 - 39.

1. Amplificadores ópticos. 2. Inteligência artificial. 3. Comunicação óptica. I. Título.

CDU: 004.8: 621.375



**MINISTÉRIO DA EDUCAÇÃO**  
UNIVERSIDADE FEDERAL DE ALAGOAS  
INSTITUTO DE COMPUTAÇÃO  
Av. Lourival Melo Mota, S/N, Tabuleiro do Martins, Maceió - AL, 57.072-970  
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO (PROPEP)  
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

## **Folha de Aprovação**


**JOSE CARLOS PINHEIRO FILHO**

### **CONTROLE ADAPTATIVO DO PONTO DE OPERAÇÃO DOS AMPLIFICADORES ÓPTICOS UTILIZANDO APRENDIZAGEM POR REFORÇO**

### **DEFINING OPTICAL AMPLIFIERS GAINS USING REINFORCEMENT LEARNING**


Dissertação submetida ao corpo docente do  
Programa de Pós-Graduação em Informática  
da Universidade Federal de Alagoas e  
aprovada em 28 de junho de 2024.

#### **Banca Examinadora:**

Documento assinado digitalmente  
 **ERICK DE ANDRADE BARBOZA**  
Data: 28/06/2024 16:12:18-0300  
Verifique em <https://validar.it.gov.br>


---

**Prof. Dr. ERICK DE ANDRADE BARBOZA**  
UFAL – Instituto de Computação  
**Orientador**

Documento assinado digitalmente  
 **BALDOINO FONSECA DOS SANTOS NETO**  
Data: 28/06/2024 17:21:00-0300  
Verifique em <https://validar.it.gov.br>

---

**Prof. Dr. BALDOINO FONSECA DOS SANTOS NETO**  
UFAL – Instituto de Computação  
**Examinador Interno**

Documento assinado digitalmente  
 **JOAQUIM FERREIRA MARTINS FILHO**  
Data: 28/06/2024 16:27:45-0300  
Verifique em <https://validar.it.gov.br>

---

**Prof. Dr. JOAQUIM FERREIRA MARTINS FILHO**  
UFPE - Universidade Federal de Pernambuco  
**Examinador Externo**

## List of Figures

1	Basic concept of a communication system. Source: adapted from [25]. . . . .	4
2	Concept of a digital optical system. Source: adapted from [25] . . . . .	5
3	Concept of wavelength division multiplexing (WDM). Source: adapted from [13] . . . . .	6
4	"Example of a flat signal, with a tilt of 5 dB, and with a tilt of -5 dB. (Source: Provided by the advisor) . . . . .	8
5	Power mask of an optical amplifier, highlighting a specific parameter of the amplifier. (Source: [5]) . . . . .	10
6	Flow of how SB3 interacts with GNPY during the execution of a step. . . . .	22
7	Optical link used for the scenarios employed in the simulation. It comprises a transmitter (TX), 4 fiber spans of 110 Km followed by optical amplifiers each, and the receptor (RX). . . . .	23
8	Optical Signal spectrum on the first amplifier input for each scenario. . . . .	24
9	The reward as a function of the training steps considering five independent runs in the four simulations scenarios. . . . .	27
10	The distribution of final GSNR (dB) from fifty simulations of each model into the same scenario it was trained. The reference is the GSNR returned by the loss compensation. . . . .	29
11	GSNR Spectrum at the last link amplifier output when the best ACOP-RL model of each scenario is used and when the loss compensation (reference) is used. . . . .	30
12	GSNR distribution of fifty simulations from every best model of each scenario applied in every other scenario that was not trained with the reference value (loss compensation) for all of them. . . . .	31
13	Pareto front returned by the ACOP-MOO [6] and the positioning of the ACOP-RL solution (triangle) in the space formed by the minimum GSNR and the minimum ripple. . . . .	34

## List of Tables

1	SCENARIOS' PARAMETERS	24
2	MEAN GSNR MEDIAN REACHED BY EACH ACOP-RL MODEL TRAINED IN ONE SCENARIO AN TESTED IN ALL SCENARIOS, THE MEAN GSNR VALUE RETURNED BY THE LOSS COMPENSATION (REFERENCE) AND THE INCREASE IN THE MEAN GSNR WHEN ACOP-RL IS USED INSTEAD OF LOSS COMPENSATION. ALL VALUES ARE IN DB AND THE HIGHEST VALUES FOR EACH COLUMNS ARE HIGHLIGHTED.	32
3	MEAN OF TIME SPENT TO PREDICT EACH SCENARIO BY MODEL (SECONDS)	33

# Contents

<b>1 Introduction</b>	<b>1</b>
<b>2 Objectives</b>	<b>2</b>
2.1 General Objective	3
2.2 Specific Objectives	3
<b>3 Theoretical Foundation</b>	<b>4</b>
3.1 Optical Communication	4
3.1.1 Optical amplifiers	7
3.1.2 Signal quality	8
3.1.3 Power Mask	9
3.2 Reinforcement learning	10
3.2.1 Reinforcement Learning concept	10
3.2.2 Reinforcement learning elements	11
3.2.3 Reinforcement learning example	13
3.2.4 Policy Gradient Methods	13
3.2.5 Proximal Policy Optimization algorithm	14
<b>4 Related Works</b>	<b>15</b>
4.1 ACOP Techniques	15
4.2 Applications of Reinforcement Learning	17
4.3 Relevance of the Proposal	18
<b>5 Materials and methods</b>	<b>20</b>
5.1 Modeling the ACOP as a Reinforcement Learning Problem	20
5.2 Integrating RL with Optical Simulation	21
5.3 Amplifier Modeling	23
5.4 Scenarios	23
5.5 Training	25
5.6 ACOP Multiobjective Optimization (MOO) Implementation	25

<b>5.7 Evaluated metrics</b> . . . . .	26
<b>6 Results</b>	<b>27</b>
<b>6.1 Training</b> . . . . .	27
<b>6.2 Validation</b> . . . . .	28
<b>6.3 Comparison with ACOP-MOO</b> . . . . .	32
<b>7 Conclusions</b>	<b>35</b>

## **Abstract**

The dynamic nature of future optical networks requires that amplifiers autonomously adjust their gain in response to changing network conditions, such as the addition or removal of channels, to maintain signal power and General Signal-to-Noise Ratio (GSNR) across a cascade of amplifiers. This challenge is known as the Adaptive Control of Optical Amplifier Operating Point (ACOP). Solutions for the ACOP problem have been proposed using techniques such as cognitive learning, supervised learning, and evolutionary algorithms. Among these, the evolutionary approach has achieved the best results in terms of transmission quality. However, it has a relatively high response time, which is a significant drawback for operational deployment. On the other hand, reinforcement learning techniques are important in the field of artificial intelligence to solve problems in real-time with trained models. This work proposes the first modeling of the ACOP problem using reinforcement learning, specifically employing the Proximal Policy Optimization (PPO) algorithm integrated with the GNPY simulator. The objective is to improve signal quality by maximizing the GSNR through interaction with the gains of the amplifiers in the link. In four scenarios with varying numbers of channels, this approach achieved results close to the evolutionary approach, but with a speed-up of 300 times.



# 1 Introduction

According to [5], the massive utilization of applications and services that require a high transmission bandwidth has increased data traffic in telecommunication networks, with an emphasis on the Internet. The perspective is that data traffic will increase approximately three times in the coming years, as [5] states that in recent years, the emergence of new services such as video-on-demand, IPTV, among others, has led to an increase of over five times in global IP network traffic.

It is stated by [18] that optical networks constitute the fundamental physical infrastructure of all major providers' networks worldwide. According to [5], this is because optical networks offer many advantages, such as having minimal interference with low losses, which makes them suitable for long distances, and also having an enormous bandwidth of up to  $50THz$ . According to [18], there is no indication that a substitute technology may emerge in the near future.

According to [5], optical networks face numerous challenges in addressing the diverse quality requirements of various services due to their need to support different types of services. The work [18] highlights that diverse approaches have been explored to enhance the performance of optical networks, including routing, wavelength allocation, traffic management, and ensuring operational resilience. [5] elaborates that the quality standards demanded for optical networks may fluctuate over time, necessitating the evolution of adaptive network architectures. A key challenge lies in establishing effective management and control mechanisms for the integrated devices within these networks.

A major breakthrough in optical fiber transmission systems came with the introduction of optical amplifiers. However, these amplifiers come with certain imperfections that need consideration when designing an optical network. As noted in [17], it is vital to recognize that in addition to amplifying signals, they also introduce noise. In addition, frequency channels undergo varying gain levels, which can affect signal uniformity and cause distortion. Therefore, as highlighted by [17], with networks becoming increasingly dynamic, it is imperative to have a thorough understanding of the performance of the amplifier to ensure an optimal configuration.

In a dynamic scenario, setting the gain of an optical amplifier *a priori* can potentially

degrade its performance. According to [5], traditional techniques typically take losses from previous devices into account to configure the amplifier gain, which may not ensure the best performance for the overall link. Therefore, it is necessary for the amplifier to be adaptable and operate autonomously to achieve optimal transmission quality.

In [6] (ACOP-MOO) evolutionary algorithms for multiobjective optimization are used to jointly select optimal operating points for amplifiers, resulting in the best GSNR results found in the literature.

In [14], a comparison was made between the multi-objective technique proposed by [6]; however, no improvements were observed in terms of transmission quality. An initiative to reduce the response time taken by the multi-objective approach in [7], using supervised learning techniques, was successful; however, the data generated by the multi-objective technique was still required for model training.

According to [15], Reinforcement Learning (RL) has emerged as a critical research domain within machine learning, playing a pivotal role in the advancement of artificial intelligence over the past two decades. Its applications extend across various fields, such as robotics, computer vision, speech recognition, and natural language processing. However, [19] notes that despite its promising prospects, RL has not been extensively explored for optimizing the performance of optical networks.

Therefore, this work aims to contribute to the adaptive control problem of the optical amplifier operating point by proposing a novel ACOP algorithm that uses reinforcement learning. The aim is to achieve an outcome better than the approach that defines the gains of the amplifier to compensate the losses, and closer to the one returned by ACOP-MOO but in a shorter time than ACOP-MOO.

## 2 Objectives

This section addresses the objective of this work, which is to contribute to the improvement of the adaptive control technique for the operating point of optical amplifiers. Additionally, it outlines the specific objectives that will guide the path towards achieving the overall goal.

## **2.1 General Objective**

To contribute to solving the adaptive control problem of the operating point of optical amplifiers (ACOP) using reinforcement learning.

## **2.2 Specific Objectives**

1. Model the ACOP problem as a reinforcement learning problem.
2. Evaluate reinforcement learning technique in different scenarios of quantity of channels.
3. Compare with technique from the literature in an optical network scenario.

## 3 Theoretical Foundation

### 3.1 Optical Communication

According to [25], an optical fiber communication system is fundamentally similar to any other communication system. It consists of a transmitter or modulator connected to the information source and the transmission medium, and a receiver or demodulator at the destination point. According to [13], all telecommunication systems use some form of electromagnetic energy to transmit signals from one device to another. This electromagnetic energy, a combination of electric and magnetic fields, encompasses radio waves, microwaves, infrared light, visible light, ultraviolet light, X-rays, and gamma rays, each constituting a band of the electromagnetic spectrum. Figure 1 illustrates the basic concept of a communication system, where a transmitter is connected to the source of information and the medium, and a receiver is connected to the transmission medium and the destination.

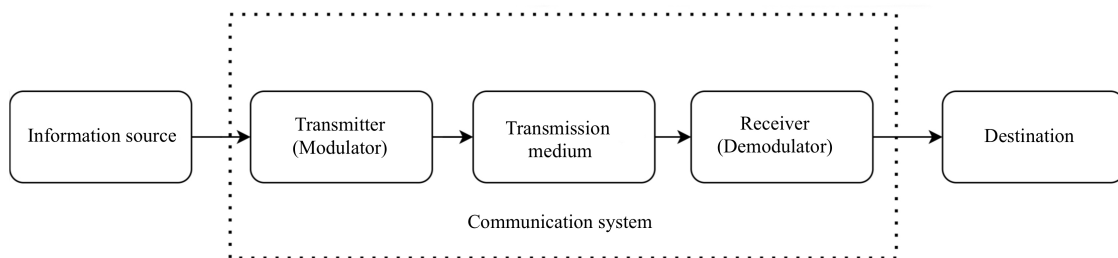


Figure 1: Basic concept of a communication system. Source: adapted from [25].

In a simplified description, [13] states that the function of an optical fiber link is to transport a signal originating from electrical equipment from one location to another with confidence and precision. [13] mentions that the components of an optical link include: transmitter, optical fiber, receiver, optical amplifiers, passive elements such as isolators and connectors, and active elements such as the optical amplifier itself and lasers. [25] depicts a concept of a digital optical system, illustrated in Figure 2, composed of a digital source, encoder, laser driver circuit responsible for modulation and laser in the transmission area. In the reception part, there is an avalanche photodiode (APD), a signal amplifier, a decoder, and finally the digital signal output. Optical fiber is used as the transmission medium.

It is explained by [13] that the electromagnetic spectrum band consisting of the optical

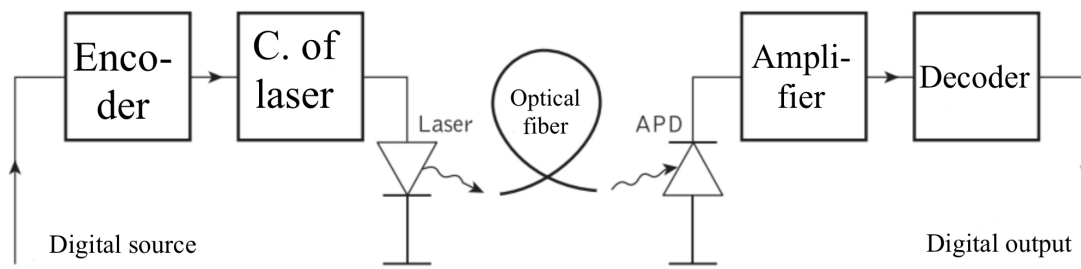


Figure 2: Concept of a digital optical system. Source: adapted from [25]

spectrum ranges from 5 nm (ultraviolet) to 1 mm (far infrared), and the region used for optical communications spans from 800 nm to 1675 nm. Furthermore, according to [13], the International Telecommunication Union has designated six transmission bands for medium and long distances, known by the letters O, E, S, C, L and U.

- *Original band* (band O): 1260 to 1360 nm
- *Extended band* (band E): 1360 to 1460 nm
- *Short band* (band S): 1460 to 1530 nm
- *Conventional band* (band C): 1530 to 1565 nm
- *Long band* (band L): 1565 to 1625 nm
- *Ultralong band* (band U): 1625 to 1675 nm

According to [25], optical carriers can be modulated both analogically and digitally, with analog modulation involving continuous variation of the light emitted by the source, while digital modulation uses discrete changes in light intensity in on-off pulses. However, [25] points out that analog modulation is limited, primarily because semiconductor optical sources cannot maintain the necessary linearity, especially at high frequencies. Therefore, they are only used for short distances and with bandwidth operations smaller than those using digital modulation.

Wavelength division multiplexing (WDM), according to [13], provides an additional boost in fiber transmission capacity. According to [25], WDM involves transmitting multiple optical signals of different wavelengths, or channels, in parallel over a single optical

fiber. Figure 3 illustrates the concept of WDM, where several optical signals of different wavelengths are combined using a multiplexer to be transported through a single optical fiber.

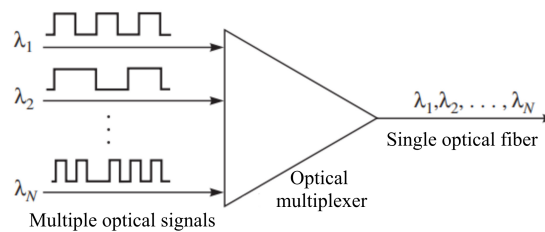


Figure 3: Concept of wavelength division multiplexing (WDM). Source: adapted from [13]

As stated in [13], the drive to accommodate more channels within a narrow spectral range led to the development of dense wavelength division multiplexing (DWDM). It is crucial to maintain sufficient spacing between optical channels to prevent interference. [25] explains that DWDM was initially designed to multiplex optical signals around the  $1.55 \mu\text{m}$  wavelength, utilizing erbium-doped fiber amplifiers (EDFAs) to increase the capacity of the system and reduce costs.

According to [28], there are cases in dynamic traffic networks where optical paths are added and removed, necessitating the use of heuristic methods to assign wavelengths (or channels) to these new paths. [28] mentions several heuristics that can be applied to channel allocation, two of which are *Random* and *First-Fit*, aimed at reducing the probability of blocking a new connection. These methods can also be implemented online and combined with different network architectures. *Random*, or random wavelength assignment, as described by [28], initially seeks the spacing between channels to determine all possible wavelengths that will be available. Whenever a new channel is added, the algorithm searches for free channels and selects one randomly, typically following a uniform probability distribution. On the other hand, *First-Fit*, outlined by [28], assigns numerical labels to all channels. When a new path is added, it selects the smallest available channel number. Unlike *Random*, *First-Fit* requires global information about available channels but has lower computational cost since it avoids frequent searches for free channels.

### 3.1.1 Optical amplifiers

According to [25], optical amplifiers operate exclusively in the optical domain without photon-to-electron conversion, and are placed at intervals along a fiber link to amplify transmitted optical signals. They are bidirectional and support multiplexing operations, particularly in single-mode fiber systems where signal dispersion effects can be minimal. Therefore, the primary limitation on repeater spacing becomes attenuation due to fiber losses. [13] states that optical amplifiers are essential for high-performance optical communication links. Furthermore, [13] mentions three basic amplifier technologies: semiconductor optical amplifiers (SOA), erbium-doped fiber amplifiers (EDFA) and Raman amplifiers.

As stated in [17], erbium-doped fiber amplifiers (EDFA), which are a type of doped fiber amplifier (DFA), use erbium-ionized fiber as a means of amplification. This fiber is pumped by lasers operating at 980 nm or 1480 nm wavelengths to maximize energy transfer efficiency between the pumping source and the signal. According to [17], the application of EDFAs covers the C band, which ranges from the wavelength of 1530 to 1560 nm.

The EDFA exhibits gain dependence on the signal wavelength, so in multichannel systems, different channels may be amplified with varying gains. Consequently, amplifying channels with different levels causes the amplifier's output spectrum to become non-flat, resulting in the accumulation of this slope or *tilt* [17, 1]. In a system where the signal traverses a series of optical amplifiers, signal distortion can become a critical factor for system quality, as some channels may arrive at the receiver with very low power, potentially causing reception failure. Another point mentioned by [13] is that in a chain of amplifiers within a link, amplified spontaneous emission noise (ASE) is so dominant that thermal noise from the receiver and shot noise can be neglected when calculating the signal-to-noise ratio of the link.

EDFA faces challenges such as the spectral flatness of the signal and the low noise figure [17]. One metric used to measure amplifier distortion is the signal *tilt*. The tilt is the angular coefficient of the line segment representing the channels in the output power spectrum of the amplifier. Essentially, it is calculated by subtracting the power value of the lowest-frequency signal from that of the highest-frequency signal. Figure 4 illustrates three signals: a flat signal where power levels are equal, a signal with positive tilt where the power

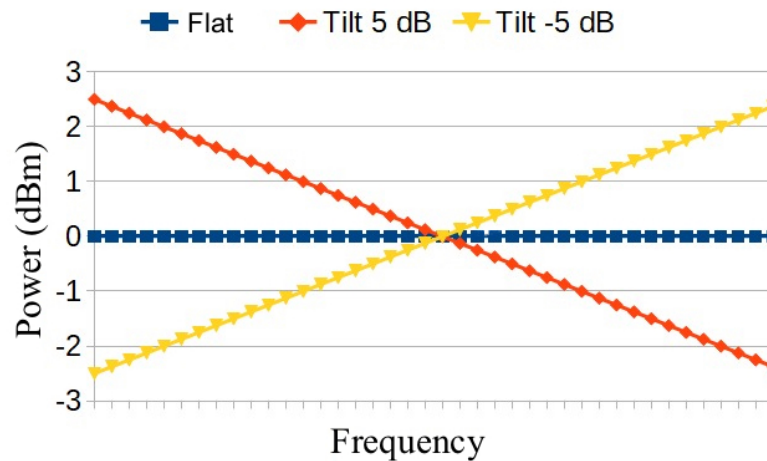


Figure 4: "Example of a flat signal, with a tilt of 5 dB, and with a tilt of -5 dB. (Source: Provided by the advisor)

at the higher frequency is lower than at the lower frequency, and a signal with negative tilt where the power at the higher frequency is higher than at the lower frequency. [5] mentions a metric similar to tilt, which is also important to define the quality of the amplifier, known as *ripple*. The ripple is calculated by subtracting the power value of the lowest channel from the highest channel in the spectrum.

### 3.1.2 Signal quality

For [13], one of the main challenges in operating WDM networks is ensuring proper system functionality. Therefore, [13] emphasizes the critical importance of intelligently monitoring each channel to meet network reliability requirements and ensure agreed-upon Quality of Service (QoS) standards with clients. Finally, [13] mentions that the key performance parameters to monitor include wavelength, optical power, Optical Signal-to-Noise Ratio (OSNR), and Bit Error Rate (BER).

According to [13], the OSNR is a metric used to plan, install, and verify the health of optical networks and individual optical channels. The OSNR depends solely on the ratio of the average signal power  $P_{signal}$  to the average noise power  $P_{noise}$ , and is independent of factors such as data format, pulse shape, or bandwidth filters. The OSNR can be calculated using these two powers and a reference spectral bandwidth  $L_{ref}$ , typically set to 0.1 nm. The equation for OSNR takes into account the spectral bandwidth of the noise measurement



equipment  $L_b$ , as seen in Equation [1].

$$OSNR = 10 \log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) + 10 \log_{10}\left(\frac{L_{ref}}{L_b}\right) \quad (1)$$

Consonant [10], a metric that has been proven to be unique and effective in evaluating the quality of modern WDM optical transmission, modulated at various levels, is the Generalized Signal-to-Noise Ratio (GSNR). [10] states that GSNR encompasses both the Optical Signal-to-Noise Ratio (OSNR) and the Nonlinear Signal-to-Noise Ratio ( $SNR_{NL}$ ). Equation [2] demonstrates how the GSNR can be calculated according to [17], and can be performed for each channel or as an average of the signal.

$$\frac{1}{GSNR} = \frac{1}{OSNR} + \frac{1}{SNR_{NL}} \quad (2)$$

### 3.1.3 Power Mask

In a dynamic network, it is essential that amplifiers operate satisfactorily within a range of input power levels to accommodate varying channel loads. These channels can dynamically change, and input power may also fluctuate due to factors such as the gain of the preceding amplifier [17].

The power mask is defined by [5] as the operational region of optical amplifiers, encompassing the minimum and maximum input and output powers, thus defining the maximum and minimum gain of the devices. The performance of the amplifier within the mask is rarely known and can be obtained by laboratory characterization. According to [17], this characterization involves varying the input power and/or channel loading for each gain setting to obtain output power values and parameters such as the noise figure, the gain control accuracy, and the spectral gain flatness, thus describing the dynamic behavior of the optical amplifier. Finally, [17] states that the power mask can be used in dynamic networks to predict amplifier performance with varying channel loading. Figure [5] illustrates the power mask of an optical amplifier with a specific highlighted parameter.

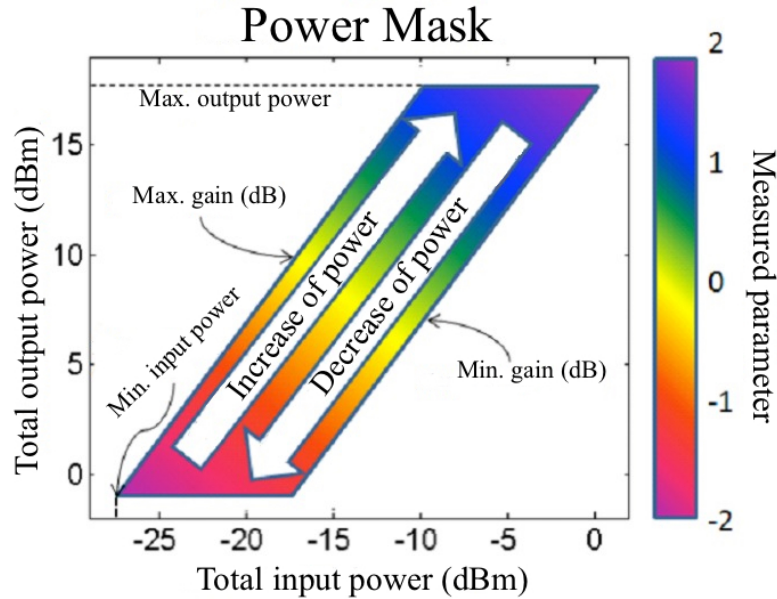


Figure 5: Power mask of an optical amplifier, highlighting a specific parameter of the amplifier. (Source: [5])

## 3.2 Reinforcement learning

From the perspective of [23], reinforcement learning may be the only viable approach in certain complex domains for training a program to make good decisions. For example, they illustrate the difficulty that a human would face in teaching an agent to pilot a helicopter, which would require an enormous amount of data covering all possible scenarios encountered while flying. However, by providing appropriate negative rewards for crashing, oscillating, or deviating from a defined course, an agent can learn to fly autonomously.

### 3.2.1 Reinforcement Learning concept

Reinforcement learning differs from supervised learning, which is the type of learning studied in most current machine learning research. Supervised learning involves learning from a training set of labeled examples provided by a qualified external supervisor. Each example consists of a description of a situation along with a specification of the correct action the system should take in that situation, often identifying a category to which the situation belongs [26].

In [26] the authors explain that the goal of supervised learning is for the system to ex-

trapolate, or generalize, its responses so that it can act correctly in situations not present in the training set. They also indicate that, while important, this type of learning alone is not suitable for learning from interactions. In interactive problems, it is often impractical to obtain examples of desired behavior that are correct and representative of all the situations in which the agent must act. Therefore, in unknown territory, where learning would be most beneficial, an agent must be able to learn from its own experience.

The reinforcement learning problem can be formulated mathematically as a Markov Decision Process. A Markov Decision Process is defined by [23] as a problem involving a sequence of decisions with an observable environment, a set of states, and a set of actions, where the state changes based on the chosen action and a reward function.

In reinforcement learning, the purpose of the agent is the reward, the signal passed from the environment to the agent on every interaction or time step. At each time step, the reward will be a number and the agent's goal is to maximize the total amount of reward in the long run. There is a special state called the terminal state, and the set of time steps until the agent reaches the terminal state is called episode, but there are some cases where the end of the episode is not so clear. So, it is possible to define that the main goal of the agent is to reach the maximum reward at the end of an episode.

It is pointed out in [26], that an important aspect of reinforcement learning is the balance between exploration and exploitation. To achieve the best results, a reinforcement learning agent must exploit actions that it has tried in the past and found to be effective in producing rewards. However, to discover such actions, it must explore actions it has not selected before. This creates a dilemma where one should not focus entirely on exploration or exploitation, as pure exploitation leads to getting stuck in a routine, while pure exploration leads to never putting knowledge into practice, as stated by [23]. An agent should try out a variety of actions and progressively favor those that appear to be better.

### 3.2.2 Reinforcement learning elements

- Agent: The learner of the reinforcement learning [26]. Periodically make decisions, observe the results, and then automatically adjust its strategy [15].
- Environment: The element which the agent interacts and everything outside the agent.

The agent and the environment interact continuously [26].

- **Policy:** Defines the way an agent behaves at a given time; broadly speaking, it is a mapping from perceived states of the environment to actions to be taken when in those states. It can be a simple function, a table, or something more computationally extensive, such as a search process. The policy is the core of the agent and is sufficient to dictate the agent's behavior. In general, policies can be stochastic, specifying probabilities for each action [26].
- **Reward Signal:** Defines the objective of a reinforcement learning problem. In each interaction, the environment sends the reinforcement learning agent a single number called the reward. The unique objective of the agent is to maximize the total reward received in the long term, thus defining what constitutes a bad action and how good an action is. Rewards are the immediate and defining features of the problem facing the agent. The reward signal is the primary basis for altering the policy. If an action selected by the policy is followed by a low reward, then the policy may be changed to select a different action in that situation in the future. Finally, reward signals can be stochastic functions of the state of the environment and the actions taken [26].
- **Value function:** Specifies what is beneficial in the long term. The value of an environment state is the total amount of reward an agent can expect to accumulate in the future, starting from that state. Although rewards determine immediate desirability, values indicate long-term desirability by considering the states likely to follow and the rewards available in those future states. For example, a state may consistently yield a low immediate reward but still have a high value because it is often followed by other states that yield high rewards. Conversely, a state may have a high immediate reward, but a low value if it leads to states with low rewards.
- **Episode:** The set of time steps or actions-states until an agent reaches a determined final state which will return the accumulated reward of the actions chosen by the agent. [26].

### 3.2.3 Reinforcement learning example

In [26] the authors give an example of an application of reinforcement learning of a mobile robot that needs to decide when to go to a new room or go back to the battery recharging station. It makes decisions based on the current level of the battery and how easy and fast it has been able to find the recharging station before.

A way to model this problem for reinforcement learning could be:

- Agent: The robot.
- Environment: The house and the battery.
- Episode: the set of time steps, which are the choices of go to another room or go back to the battery station and the final of the episode can be the end of the battery or clear all the house.
- Reward: A higher value if the robot clear the house with less backs to the battery station or negative if the battery is over before the house be totally clean.

### 3.2.4 Policy Gradient Methods

In reinforcement learning, there is a class of methods used when there is an infinite number of possible observations in an environment, known as Approximate Solution Methods. These problems, with a large number of possible states, necessitate generalization so that new states can be evaluated based on previous states, given the impossibility of storing all past states. Supervised techniques, such as neural networks and pattern recognition, can be used as approximation functions for these methods. Approximation methods aim to receive examples of the desired input-output behavior of the function they are trying to approximate, such as the value function [26].

As a subset of approximate solution methods, there are policy gradient methods [26]. In policy gradient methods, the policy can be parameterized in any way, as long as it is different with respect to its parameters. This is an advantage because, with continuous policy, the action probabilities change smoothly as a function of the learning. The policy gradient methods work by computing an estimator of the policy gradient and putting it into a stochastic gradi-

ent ascent algorithm [24]. As part of the policy gradient methods, there is a set of algorithms called actor critic which uses the estimators for both the policy and the value function [26].

### 3.2.5 Proximal Policy Optimization algorithm

A proximal policy optimization (PPO) algorithm is described as one that uses fixed length trajectory segments where in each iteration, each of  $N$  agents collects  $T$  steps of data [24]. Then a surrogate loss is constructed on these  $NT$  steps of data and optimized with minibatch stochastic gradient descent, for  $K$  epochs. This method has stability and reliability in a simple way to implement, requiring only few changes to a vanilla policy gradient implementation with a better overall performance.

## 4 Related Works

In this chapter, various proposals from the literature that address the ACOP problem will be discussed. A brief description of these proposals will be provided in Section 4.1. Furthermore, Section 4.2 will present a concise overview of how reinforcement learning is being applied in other fields. Finally, the relevance of the current proposal will be elaborated in Section 4.3.

### 4.1 ACOP Techniques

Oliveira et al. proposed a technique for adaptive definition of the operating point of EDFA amplifiers in [20], known as Adaptive Gain Control (AdGC). In AdGC, the objective is to select the best operating point for an amplifier in terms of noise figure and ripple, considering the total input power and the points within the power mask. The best point is defined as the one with the smallest Euclidean distance. If multiple points have the same distance, the angle that the line to this point makes with the origin is used as a tie-breaker.

The AdGC has a problem where, due to small variations in ripple, a point with the best noise figure was not chosen. To address this, an improvement was presented in [8], which used weighting factors to give more importance to the noise figure in the selection process. This approach led to a decrease in signal flatness; however, the noise figure experienced a significant reduction in its value within the link.

Experiments conducted using a mesh network in [27] demonstrated the influence of power equalization, wavelength allocation, and ACOP strategies on OSNR. The results indicated that ACOP had a greater impact on transmission quality compared to power equalization and wavelength allocation strategies. When using AdGC to set gains, improvements of up to 1.6dB were observed compared to compensating for losses with fixed gains.

Em [16], a technique called AcCBR (Adaptive Cognitive Case-Based Reasoning) was developed using Case-Based Reasoning (CBR). This technique determines the gains of amplifiers along a specific optical path for each new path that arises, aiming to optimize the OSNR of connections. AcCBR is executed whenever there is a request to establish a new optical path, initiating the CBR process to determine optimal amplifier gains.

The CBR process used in [16] operates in 4 stages: retrieval, reuse, revision, and retention. With the aid of a database, in the retrieval stage, information is sought from previously established optical paths that share similar characteristics with the new path being set up, such as total input power, link losses, and number of amplifiers. In the reuse stage, changes are made to the values that define the gains in the scenarios retrieved in the previous stage, generating new gain settings. The review stage is used to evaluate the new possibility and assign an OSNR, and finally, in retention, this new possibility and its outcome are stored in the database. Finally, AcCBR considers the highest OSNR among the options presented, which may not necessarily be the one just created.

In [4], the first ACOP strategy was proposed that considers the non-linear effects of the fiber to define the amplifier operating point, in order to maximize OSNR and minimize ripple. The model used a Gaussian noise model to account for non-linear effects [21]. The inclusion of these effects had a beneficial impact compared to models that only compensated for losses.

With the aim of reducing the impact of interference from non-linear effects, a new approach was proposed in [6] (ACOP-MOO), where a Variable Optical Attenuator (VOA) is used at the output of each amplifier in the optical path. Additionally, a strategy based on multi-objective optimization with evolutionary algorithms was proposed to maximize the signal-to-noise ratio and minimize its ripple, considering the non-linear effects throughout the optical path. The addition of the VOA provided a benefit by offering more solutions that result in better transmission quality. This allows for better performance in scenarios where amplifiers operate at points with high noise figures. Furthermore, the proposed strategy yielded better results compared to previous models when considering two different commercial amplifier models.

In [14], a comparison was made between the multi-objective strategy and a single-objective strategy based on swarm intelligence. As a result, it was observed that although they performed similarly in scenarios with few amplifiers, when reaching eight amplifiers, the multiobjective strategy outperforms the swarm-based strategy. The explanation provided by [14] for these results is that, by attempting to reduce the ripple of the signal-to-noise ratio, the multi-objective technique achieves better overall performance across scenarios when considering transmission quality.



According to [7], despite giving good results, the multi-objective model based on evolutionary algorithms has a drawback when used in real-time applications, it takes a considerable amount of time to complete all iterations. Therefore, [7] conducted a study using four different machine learning techniques to develop a model aimed at replacing the multi-objective approach. This model learns from patterns without requiring many iterations, enabling it to be used to define the gains and losses of VOAs in the optical path during network operation.

The machine learning techniques used in [7] for evaluation were: *Bayesian Ridge Regressor*, *Random Forest Regressor*, *Decision Tree Regressor*, *Lasso CV*, and *SimpleMeanRegressor*. The errors returned by the techniques did not differ from each other, leading to the conclusion that the *SimpleMeanRegressor* would already suffice for the problem. When using it, an improvement in execution time of 120 times was observed compared to the multiobjective approach.

## 4.2 Applications of Reinforcement Learning

Reinforcement learning has been used to automate various problems. According to [11], with the growth of urbanization, it is anticipated that future transportation systems will feature autonomous traffic management, akin to autonomous drivers. Consequently, [11] notes that researchers are increasingly opting for reinforcement learning mechanisms not only for autonomous traffic management but also for autonomous driving. In addition, reinforcement learning is utilized for other applications, such as controlling road conditions, setting permissible speed limits, and managing vehicle energy for improved battery performance.

The continuous growth of the data network can lead to network congestion, resulting in a decrease in transfer rates, as explained by [9].

In [9], the authors proposed a solution to the congestion problem based on multiagent reinforcement learning, where each router intelligently chooses the next router based on its local observation, then receives the reward and observation from the chosen router. This approach enables routers to work in a distributed manner to reduce computational complexity compared to a single agent, thus reducing the probability of congestion and improving network performance.

In [12], it is pointed out that the congestion control of the TCP protocol, widely used

in network technologies, is not optimized for video streaming applications. The high transmission rate coupled with low latency requirements poses a challenge to traditional congestion control methods that aim to maintain quality of the user experience. To address this issue, [12] proposed a reinforcement learning framework aided by transfer learning that learns to operate within the optimal congestion window. This approach distinguishes between congestion-induced losses and other types of losses, thereby achieving high throughput and low end-to-end delay. Furthermore, [12] used a multi-agent reinforcement learning technique in the context of congestion control, surpassing conventional rule-based TCP congestion control methods in terms of throughput and delay performance, despite having a simple state space.

According to [15], there is a trend towards increasingly decentralized and autonomous networks, as evidenced by the growth of solutions based on the concept of Internet of Things. Furthermore, according to [15], in these networks, entities need to make local decisions to maximize performance in an uncertain network environment. It is demonstrated in [15] that in complex and large-scale networks, with a large state and action space, the technique currently used to mitigate computational time in finding solutions is deep reinforcement learning, which combines deep learning with reinforcement learning. Some applications of communication and networks mentioned in [15] that use deep reinforcement learning include: network access, adaptive rate control, proactive caching, connectivity preservation, network security, traffic control, resource scheduling, and data collection.

Given the lack of reinforcement learning applications in optical networks and the growing interest in this technique, in [19] the authors developed an open source tool that facilitates its application in optical network problems. They demonstrated its functionality by successfully solving two different service provisioning problems.

### **4.3 Relevance of the Proposal**

This work aims to contribute to the adaptive control problem of the optical amplifier operating point by proposing a novel ACOP algorithm that uses reinforcement learning. The aim is to achieve an outcome better than the approach that defines the gains of the amplifier to compensate the losses, and closer to the one returned by ACOP-MOO but in a shorter time than ACOP-MOO.

---

Also, this work contributes to reinforcement learning, because as pointed by [19], there is no much applications on optical networks until now, and the ACOP problem were not modeled to this technique.

## 5 Materials and methods

### 5.1 Modeling the ACOP as a Reinforcement Learning Problem

In order to solve the ACOP problem using reinforcement learning (ACOP-RL), initially, the problem needed to be modeled as an environment where an agent can interact. Thus, the first methodological decision was to define what variables the agent can observe (observation space). The variables chosen were the mean GSNR at the output of the last amplifier, the gain in each amplifier, the input power in the link, and the number of channels in the optical signal.

The variables that the agent can change are the amplifier gains. To simplify the implementation, was define a discrete set of choices, the possible choices for the agent are:

- Do nothing
- Increase in 1 dB the gain of one of the link amplifiers
- Increase in 2 dB the gain of one of the link amplifiers
- Decrease in 1 dB the gain of one of the link amplifiers
- Decrease in 2 dB the gain of one of the link amplifiers

Initially, was only considered a step of 1 dB to increase or decrease the gain. However, by adding the option of 2 dB, it was observed that the agent needed less steps to improve the GSNR.

Since it is not possible to determine exactly the end of an episode, because the model can interact infinitely increasing and decreasing the gains, was defined the end of the episode to happen when the agent gives a total of 10 steps, every step being a choice if it will increase, decrease, or do nothing. This decision was empirically tested, and what was observed was that using a high number of steps until the end of the episode only made the agent not reach a good result or just spent much more time to reach the same result as found with 10 steps.

Equation 3 shows the reward definition considered in this work. The  $GSNR_{current}$  is the mean GSNR after the choice (step), the  $GSNR_{initial}$  is the mean GSNR at the start of the training, and  $GSNR_{highest}$  is the highest mean GSNR found.

$$Reward(step) = \begin{cases} \frac{GSNR_{current}}{GSNR_{initial}} - 1 & \text{if } GSNR_{current} \leq GSNR_{highest} \\ \frac{GSNR_{current}}{GSNR_{initial}} & \text{if } GSNR_{current} > GSNR_{highest} \\ -0.001 & \text{if } GSNR_{current} \text{ is out of bounds.} \end{cases} \quad (3)$$

Was decide that a larger reward should be given when the choice drives the agent to the highest GSNR until that time. Moreover, a protection was added to not let the agent go out of the bounds of the gain values, and it has a punishment when the agent does this. The punishment is low ( $-0.001$ ) to not discourage the agent from trying choices that can lead to a high GSNR even closer to the search space limits.

The reward was the hardest aspect to model. Since the main goal is to increase the GSNR, the reward is a function of it. Before the reward definition presented in equation 3 were tried other approaches. First, was defined the reward as the difference between the GSNR after the choice and the GSNR before the choice. But it made the agent run in a loop, because the reward for having just a small improvement makes the option "do nothing" too good for the agent. To avoid this problem, instead of considering the GSNR before the choice, was defined the reward to consider the initial GSNR, and if the GSNR after the choice (current) is bigger than the initial, it is verified if the GSNR is the best until that point. If the current GSNR was the best until then, the reward would be the division of the current GSNR by the initial GSNR. Otherwise, the reward would be the division of the current GSNR by the initial minus one.

## 5.2 Integrating RL with Optical Simulation

Was used the Stable-baselines3 (SB3) version 2.0.0[22] framework because it has a good number of algorithms for Reinforcement Learning already implemented. Was used the Proximal Policy Optimization (PPO) algorithm, because it is easy to implement and has good performance compared to other RL algorithms[24]. The implementation of the PPO in the SB3 also supports the modeling of the ACOP-RL described above.

To configure the PPO, the frequency of policy update is at every end of the episode, which means that for every 10 steps a roll-out will start and the learning rate has a linear decrease starting in 0.003.

GNPy is an open source tool that its core engine is a coherent transmission estimator quality for multiplexed optical networks with wavelength division [10]. Thus, it was the chosen simulator to obtain the GSNR of a configuration that the agent could choose, using the version 2.6.0. GNPy uses a JSON file to set up the link components, such as span length, amplifiers with their respective gains, transmitter, receiver, and their sequence in the link. In addition, another JSON file is used to define the maximum frequency, minimum frequency, and channel spacing of the signal. Consequently, by providing these two JSON files as input and executing the simulation, GNPy outputs the GSNR for each signal channel at the link end. The mean GSNR is calculated as the arithmetic mean of the GSNR values of the channels.

In order to adapt the GNPy to work together with SB3, the simulation was defined as a function that has as parameters the location path for both JSON files and returns an array with the GSNR for each channel in the signal at the end of the link. In this way, for every modification the agent makes in a step, the GNPy simulation is executed, and the agent will have the GSNR that will be used in the calculation of the reward. Figure 6 illustrates the flow of how SB3 interacts with GNPy during the execution of a step, starting with the agent's choice and ending with the agent receiving the reward.

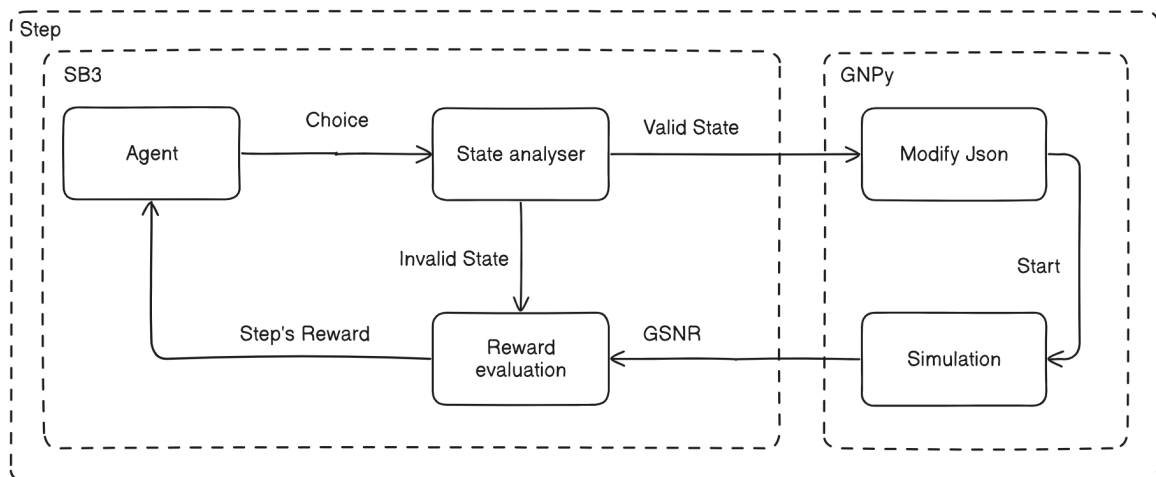


Figure 6: Flow of how SB3 interacts with GNPy during the execution of a step.

### 5.3 Amplifier Modeling

In [3] the authors proposed an estimator that uses power masks to predict the output power of the EDFA amplifier's spectrum based on the input signal's power spectrum. This estimator relies on linear interpolation and uses a base of power masks. The algorithm consults this base to find a signal with characteristics that match the input signal. If such a signal exists in the base, its output is returned. Otherwise, the algorithm searches for similar signals and estimates the output signal of the amplifier by interpolating their outputs. In this work, was used a modified version of the GNPY that uses the estimator proposed in [3] to calculate the signal and noise power of each optical channel.

### 5.4 Scenarios

Were considered four simulation scenarios. The main idea with these scenarios is to emulate the addition and removing of channels in signal that is passing through the link, which simulates a dynamic optical network in which clients are added and removed frequently. The scenarios are all point-to-point link with four EDFA amplifiers, a fiber span of  $110\text{km}$  between them, the minimum frequency is  $191,65\text{Thz}$  with a spacing of  $50\text{GHz}$ . Figure 7 illustrates the optical link.

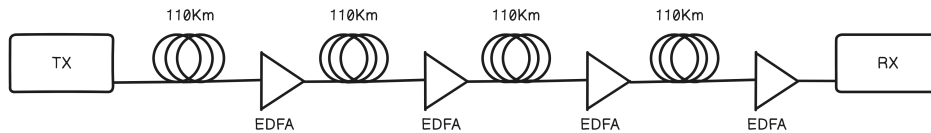


Figure 7: Optical link used for the scenarios employed in the simulation. It comprises a transmitter (TX), 4 fiber spans of 110 Km followed by optical amplifiers each, and the receptor (RX).

For each scenario, was considered a reference GSNR value ( $GSNR_{ref}$ ). The reference value of the GSNR is the mean GSNR at the end of the link when the scenario is simulated considering that each amplifier gain was defined to compensate for the losses of the previous fiber link, which means that all amplifier gains were set to  $23\text{dB}$  as the initial state. The specifications of the scenarios can be seen in Table 1, where it has the maximum frequency,  $GSNR_{ref}$  and the number of channels for each scenario. Figure 8 represents the optical

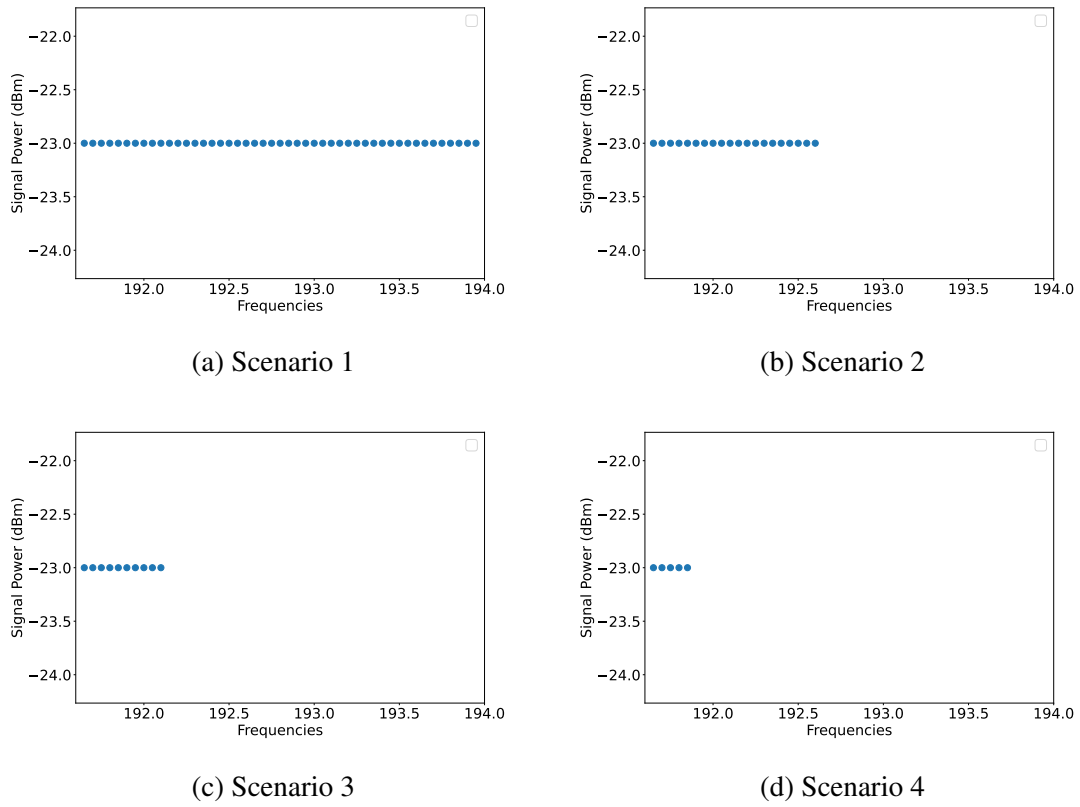


Figure 8: Optical Signal spectrum on the first amplifier input for each scenario.

signal at the input of the first link amplifier in each scenario. It is possible to see that the number of channels decreases from one scenario to another, but the signal power per channel is the same.

Table 1: SCENARIOS' PARAMETERS

Scenario	Maximum Frequency (THz)	$GSNR_{ref}$ (dB)	Number of channels
Scenario 1	194	17.55	47
Scenario 2	192.65	17.26	20
Scenario 3	192.15	17.02	10
Scenario 4	191.9	16.70	5



## 5.5 Training

To evaluate the performance of RL when trained in different scenarios, specifically to determine whether training in a scenario with fewer channels yields better results or one with more channels is preferable, five models were trained for each of the scenarios mentioned above. All models were configured identically.

For training, the maximum number of episodes was defined as 20,000, but to avoid overfitting, there are two callbacks in SB3 that can be used, *EvalCallback* and *StopTrainingOnNoModelImprovement*. *EvalCallback* is tasked with assessing the performance of the model during training, which occurs every 10 episodes. This evaluation involves applying the current model state across 5 episodes and then computing the average of the obtained rewards. *StopTrainingOnNoModelImprovement* operates in conjunction with *EvalCallback*, tracking whether the model shows progress. If no improvement is observed, it stops the training. Specifically, it begins monitoring after 10 evaluations by *EvalCallback* and requires a lack of enhancement over 10 consecutive episodes before terminating the training.

## 5.6 ACOP Multiobjective Optimization (MOO) Implementation

In [6] the authors did not use GNPY as a simulation engine. Therefore, for better comparison, the algorithm proposed in [6] (ACOP-MOO) was implemented in this work to use GNPY. We used the Pymoo library [2] to model the ACOP-MOO problem. This library was used because it is implemented in Python like GNPY, which facilitates the integration process. Moreover, this library has the implementation of the nondominated sorting genetic algorithm II (NSGA-II), which was used in [6] as the optimization algorithm.

ACOPM-MOO uses the NSGA-II to find amplifier gains that maximize the minimum GSNR, minimize signal ripple, and minimize power tilt. The minimum GSRN is the lowest GSNR among all signal channels. The signal ripple is the difference between the power of the channel with the highest power and the power of the channel with the lowest power. The power tilt is the power difference between the first and last channels. The ACOP-MOO proposed in [6] also optimizes the losses of variable optical attenuators (VOA) present at the amplifier output. However, in this work, was adapt the ACOP-MOO to optimize only the gains of the amplifiers because the current version of the ACOP-RL only considers the gains

in the learning process. Moreover, we adapt ACOP-MOO to only maximize the minimum GSNR and minimize the ripple because the ripple minimization is proportional to the power tilt minimization in the scenarios defined. In all ACOP-MOO simulations, 50 generations were considered, with a population of 75 individuals, a crossover rate of 75% and a mutation rate of 14%. The values of these parameters follow the ones defined in [6].

## **5.7 Evaluated metrics**

To ensure that the models learned, after the training, each of the five models trained in a scenario were tested in the same scenario 50 times (simulations), and then the GSNR distribution of these 50 results was compared with the reference value for the specific scenario. Also, with this comparison, it is possible to identify which model is the best among the five.

Subsequently, the optimal model chosen for a specific scenario underwent testing across alternative scenarios where it had not been trained. The evaluation involved 50 simulations, and similarly to previous tests, the GSNR distributions were evaluated against the benchmarks of each corresponding scenario.

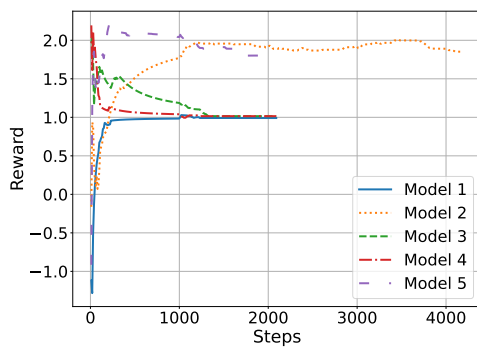
Furthermore, we recorded the computational time for each test, enabling an understanding of the model response time in practical applications. The simulations were performed on a machine with Intel(R) Core(TM) i5-1135G7 processor which operates at 2.40GHz, and with 16GB of RAM memory.

## 6 Results

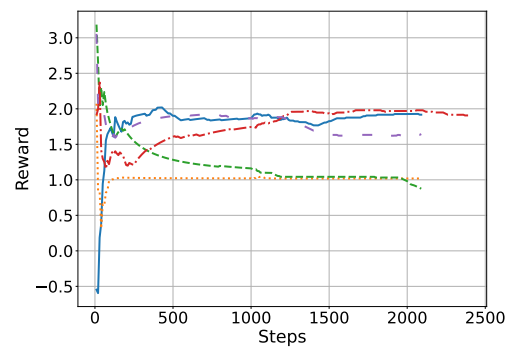
In this section, the results obtained during training, validation, and comparison between the ACOP-RL and ACOP-MOO approaches will be presented.

### 6.1 Training

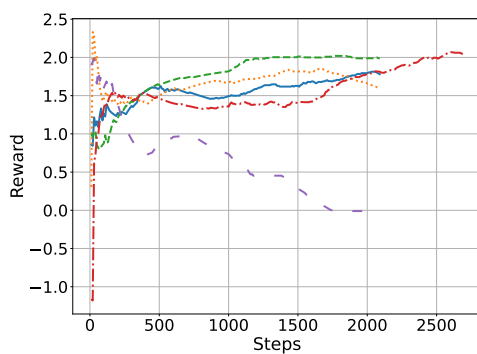
Figure 9 shows the reward as a function of the models training steps considering five distinct runs. One can see that in the majority cases the reward is converging. Even having the maximum step number set to 20,000, the training stops earlier to avoid overfitting.



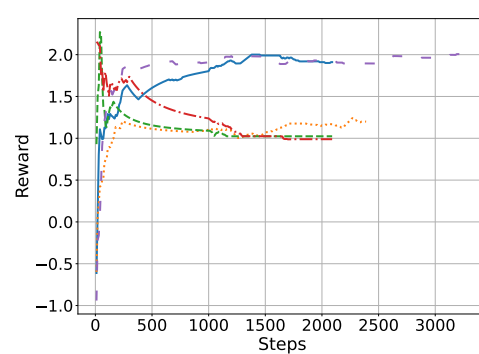
(a) Scenario 1



(b) Scenario 2



(c) Scenario 3



(d) Scenario 4

Figure 9: The reward as a function of the training steps considering five independent runs in the four simulations scenarios.

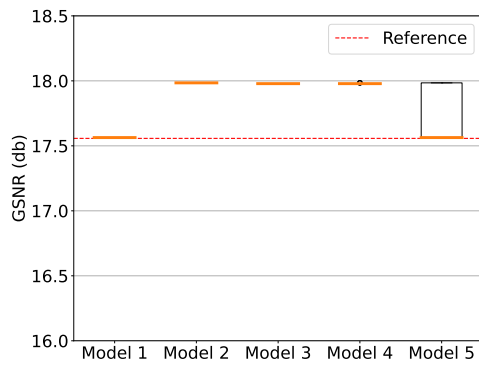
## 6.2 Validation

Figure 10 shows the distribution of the GSNR value at the end of the link for 50 independent simulations considering each simulation scenario and the five models returned in the five runs of the training phase (Fig. 9), the dashed red line is the GSNR value returned when the gains of the amplifiers are set to compensate the link losses. Considering scenario 1 (Fig. 10a), one can see that the best model was the one returned in Model 2 with the median reaching  $17.98dB$ , whereas the reference for that scenario is  $17.55dB$ . Considering scenario 2 (Fig. 10b), the best model was the one returned in the training Model 5 with a median of  $17.74dB$ , whereas the reference is  $17.26dB$ . Considering scenario 3 (Fig. 10c), the reference is  $17.02dB$  and the best model was the one returned in the training Model 3 with a median of  $17.58dB$ . Considering scenario 4 (Fig. 10d), the best model was the one returned in training Model 4 with a median of  $17.26dB$ , whereas the reference is  $16.70dB$ .

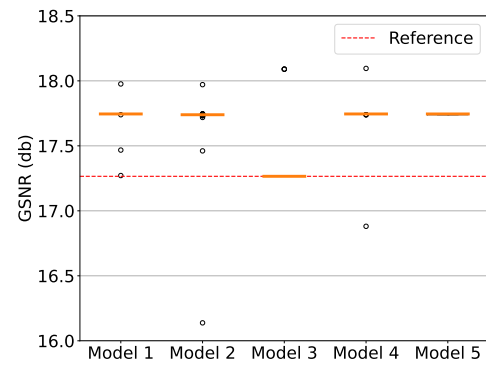
For all scenarios, one can see that the proposed model achieved a better result than the respective reference in at least one training run. Moreover, it should be noted that even the worst run of each scenario is better than or equal to the reference.

Figure 11 shows the GSNR spectrum of the optical signal at the output of the last link amplifier when the best ACOP-RL model (best training run) was used in each simulation scenario, and the same spectrum when loss compensation (reference) is used. Since for each ACOP-RL model we had 50 signals due to the 50 simulations, we considered the signal returned by the simulation with the GSNR value equal to the median value of the 50 simulations. One can see that in all scenarios, the channels had improvements in the GSNR value. Looking at Figs. 11c and 11d, one can see that the scenarios with fewer channels had a higher increase in the GSNR than those with more channels.

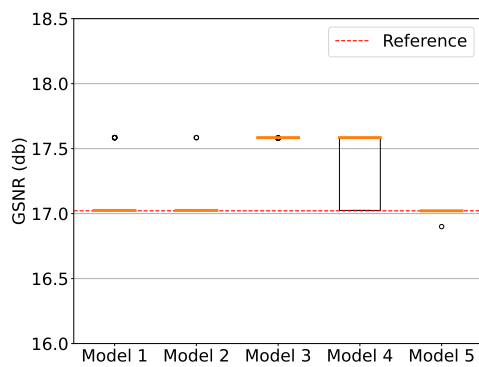
Figure 12 shows the distribution of the GSNR after fifty simulations using the best model of each scenario applied in every other scenario that was not trained, and the GSNR returned by the loss compensation in all of them. The best model trained in scenario 1 returned the median of  $17.74dB$  for scenario 2,  $17.58dB$  for scenario 3 and  $17.26dB$  for scenario 4, which can be seen in Figure 12a. Meanwhile, the best trained in scenario 2 has the median of  $17.98dB$  for scenario 1, in scenario 3 it reached the median of  $17.58dB$  and for scenario 4 it has  $17.26dB$  as shown in Figure 12b. In Figure 12c, the best model trained in scenario



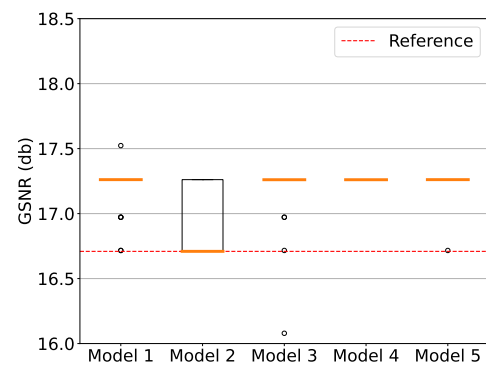
(a) Scenario 1



(b) Scenario 2

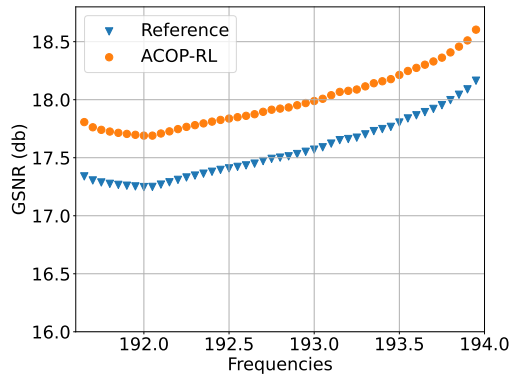


(c) Scenario 3

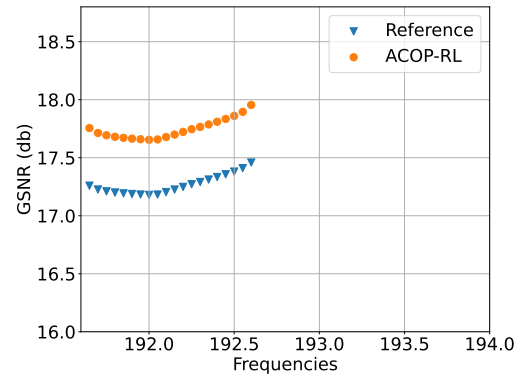


(d) Scenario 4

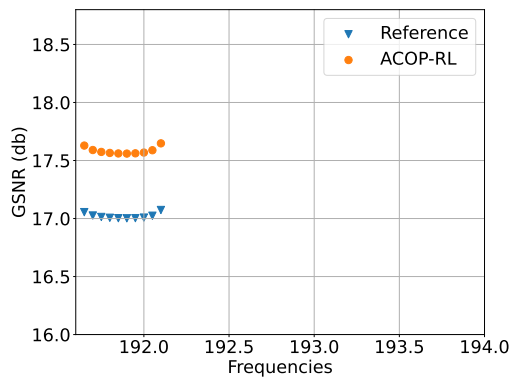
Figure 10: The distribution of final GSNR (dB) from fifty simulations of each model into the same scenario it was trained. The reference is the GSNR returned by the loss compensation.



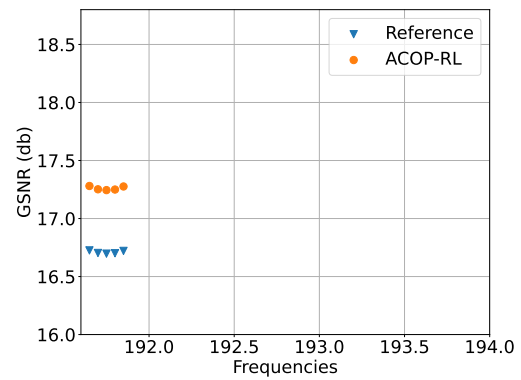
(a) Scenario 1, with Model 2



(b) Scenario 2, with Model 5



(c) Scenario 3, with Model 3



(d) Scenario 4, with Model 4

Figure 11: GSNR Spectrum at the last link amplifier output when the best ACOP-RL model of each scenario is used and when the loss compensation (reference) is used.

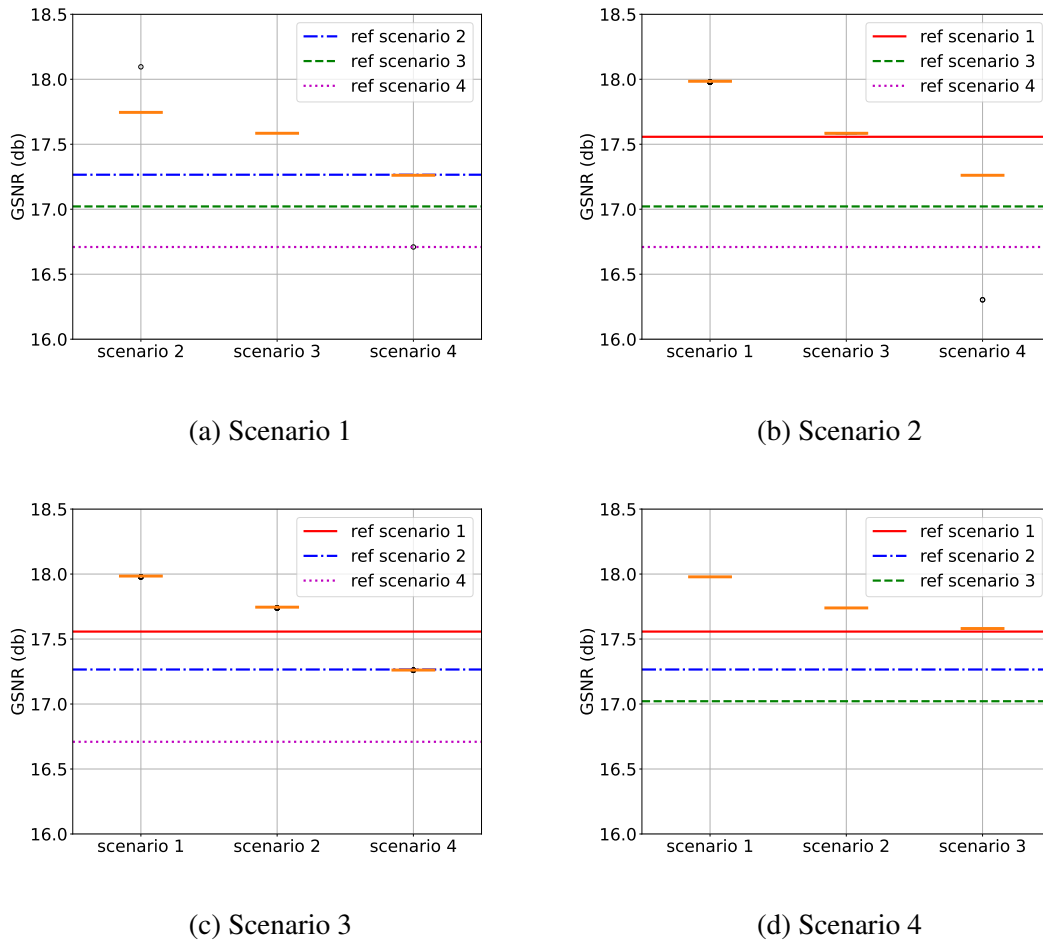


Figure 12: GSNR distribution of fifty simulations from every best model of each scenario applied in every other scenario that was not trained with the reference value (loss compensation) for all of them.

3 returned the median of  $17.98dB$ ,  $17.74dB$ , and  $17.26dB$ , respectively, for scenarios 1, 2, and 4. In Figure 12d, the best model trained in scenario 4 reached the median for scenario 1 of  $17.97dB$ , when applied to scenario 2 it got  $17.73dB$  and for scenario 3 the median was  $17.58dB$ . All models returned good results in the other scenarios, regardless of whether channels are added or removed from the spectrum. As can be seen in Table 2, which shows the median reached by every model in each scenario and the reference value, all models have almost the same result in every scenario. Moreover, one can see that the ACOP-RL models are better than the reference in all scenarios with an increase of at least  $0.43$  dB in the mean GSNR.

Table 3 shows the mean time each model spent predicting each scenario, the row shows

Table 2: MEAN GSNR MEDIAN REACHED BY EACH ACOP-RL MODEL TRAINED IN ONE SCENARIO AN TESTED IN ALL SCENARIOS, THE MEAN GSNR VALUE RETURNED BY THE LOSS COMPENSATION (REFERENCE) AND THE INCREASE IN THE MEAN GSNR WHEN ACOP-RL IS USED INSTEAD OF LOSS COMPENSATION. ALL VALUES ARE IN DB AND THE HIGHEST VALUES FOR EACH COLUMNS ARE HIGHLIGHTED.

		<b>Test Scenario</b>			
		1	2	3	4
<b>Trained Scenario</b>	1	<b>17.98</b>	<b>17.74</b>	<b>17.58</b>	<b>17.26</b>
	2	<b>17.98</b>	<b>17.74</b>	<b>17.58</b>	<b>17.26</b>
	3	<b>17.98</b>	<b>17.74</b>	<b>17.58</b>	<b>17.26</b>
	4	17.97	17.73	<b>17.58</b>	<b>17.26</b>
Reference		17.55	17.26	17.02	16.70
<i>Increase</i>		<i>0.43</i>	<i>0.48</i>	<i>0.56</i>	<i>0.56</i>

the scenario in which the model was trained, and the column shows the scenario in which the model was applied. It is possible to see that every model spent less than 1.68 seconds to reach out a configuration to the link. On the other hand, the ACOP-MOO spent between 12 and 13 minutes to achieve a solution for each Scenario considered. Therefore, the ACOP-RL is 360 times faster than the ACOP-MOO to define the amplifier gains. It is important to note that ACOP-MOO does not need a training phase, whereas ACOP-RL does. However, the ACOP-RL training times for scenarios 1, 2, 3, and 4 were, respectively, 16.92, 8.81, 11.08 and 7.71 minutes, and it is spent only once.

### 6.3 Comparison with ACOP-MOO

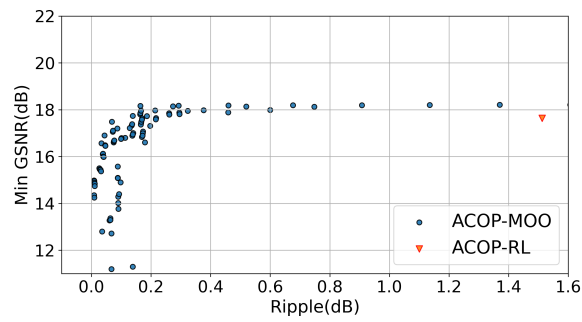
Figure 13 shows the Pareto front returned by the ACOP-MOO [6] after optimizing the gains in the scenarios defined in this work. Figure 13 also shows where the solution returned by ACOP-RL (depicted in Fig. 11) is in the space formed by the ACOP-MOO optimization objectives. It is possible to see that the solution returned by the ACOP-RL is close to the solutions with highest GSRN returned by the ACOP-MOO (the difference is less than 0.2



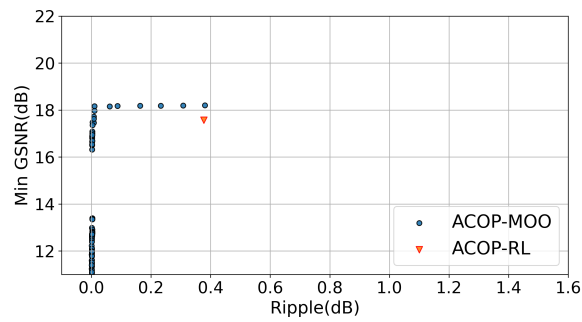
Table 3: MEAN OF TIME SPENT TO PREDICT EACH SCENARIO BY MODEL (SECONDS)

		<b>Test Scenario</b>			
		1	2	3	4
<b>Trained Scenario</b>	1	1.67	1.42	1.01	0.97
	2	1.67	1.52	1.53	1.48
	3	1.11	1.12	1.12	1.08
	4	1.65	1.45	1.47	1.49

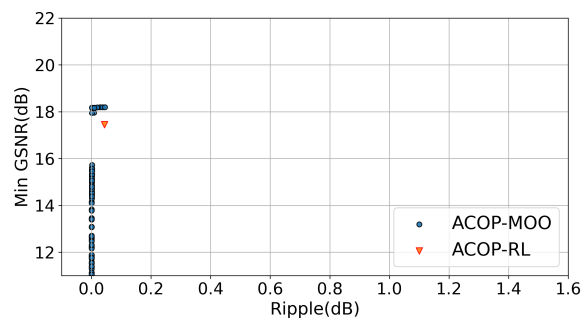
dB). Moreover, the ripple of the ACOP-RL is close to the ACOP-MOO solutions with the highest ripple. This positioning of the ACOP-RL solution in this space is expected, since the ACOP-RL only considered the GSNR as a reward.



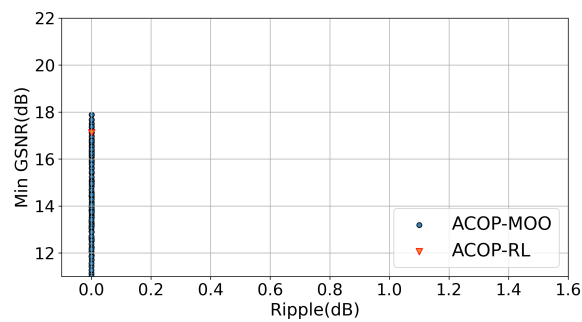
(a) Scenario 1



(b) Scenario 2



(c) Scenario 3



(d) Scenario 4

Figure 13: Pareto front returned by the ACOP-MOO [6] and the positioning of the ACOP-RL solution (triangle) in the space formed by the minimum GSNR and the minimum ripple.

## 7 Conclusions

In this work, a solution for the ACOP problem was proposed using reinforcement learning for the first time. The ACOP problem was modeled as an environment where an agent could interact with the amplifier gains, choosing if will increase or decrease the value, and learn to reach the best mean GSNR at the end of a cascade of amplifiers. To validate the algorithm through simulations, we integrate the GNPpy simulator with the RL framework SB3.

The proposed ACOP algorithm using RL outperformed the loss compensation strategy and returned similar results to the evolutionary technique found in the literature considering mean GSNR. Considering speed, the proposed algorithm returns solutions in less than 2 seconds, whereas the evolutionary algorithm takes 12 minutes. These results were achieved regardless of the scenario in which the model was trained or applied.

Although it returned good results, it is necessary to consider that in this work the effort was to propose the algorithm and validate its performance. Therefore, we only consider one optical amplifier model and one link configuration. Moreover, it did not take into account other reinforcement learning algorithms for comparison with PPO, and would not performed a more in-depth study on the optimization of the hyperparameters of the RL model.

In future work, as a suggestion, we will make comparisons with other reinforcement learning algorithms. In addition, to use other optical amplifier models in the link and perform tests in more complex network scenarios.

In conclusion, the first modeling of the ACOP problem using reinforcement learning was presented, and it opens the possibility for further refinement of this technique to achieve even better results and the potential to combine it with other techniques for dynamic configuration of the optical link with good quality and reasonable time.

Also, this work contributes to reinforcement learning, because as pointed by [19], there is no much applications on optical networks until now, and the ACOP problem were not modeled to this technique.

## References

- [1] C. J. Bastos-Filho, E. d. A. Barboza, and J. F. Martins-Filho. Estimating the spectral gain and the noise figure of edfa using artificial neural networks. *19th International Conference on Transparent Optical Networks (ICTON)*, pages 1 – 4, 2017.
- [2] J. Blank and K. Deb. pymoo: Multi-objective optimization in python. *IEEE Access*, 8:89497–89509, 2020.
- [3] A. A. B. Da Silva, E. d. A. Barboza, C. J. Bastos-Filho, and J. F. Martins-Filho. Adapting optical amplifier response estimation to consider non-flat input signals. In *2021 IEEE Latin-American Conference on Communications (LATINCOM)*, pages 1–6. IEEE, 2021.
- [4] E. de A. Barboza, M. J. da Silva, L. D. Coelho, C. J. A. Bastos-Filho, and J. F. M. Filho. Amplifier adaptive control of operating point considering non-linear interference. *PHOTONICS TECHNOLOGY LETTERS*, 30:573–576, 2018.
- [5] E. de Andrade Barboza. *Amplificadores Ópticos Autônomos: desenvolvimento e análise de técnicas*. PhD thesis, Universidade Federal de Pernambuco, 2017.
- [6] E. de Andrade Barboza, C. J. A. Bastos-Filho, and J. F. M. Filho. Adaptive control of optical amplifier operating point using voa and multi-objective optimization. *Journal of Lightwave Technology*, 37:3994–4000, 2019.
- [7] L. M. de Freitas, E. de A. Barboza, C. J. A. Bastos-Filho, and J. F. Martins-Filho. Surrogate model for adaptive control of optical amplifier operating point based on machine learning. *2021 IEEE Latin-American Conference on Communications (LATINCOM)*, pages 1–6, 2021.
- [8] U. C. de Moura, J. R. F. Oliveira, J. C. R. F. Oliveira, and A. C. Cesar. Edfa adaptive gain control effect analysis over an amplifier cascade in a dwdm optical system. *IEEE Microwave & Optoelectronics Conference (IMOC)*, pages 1–5, 2013.

- [9] R. Ding, Y. Yang, J. Liu, H. Li, and F. Gao. Packet routing against network congestion: A deep multi-agent reinforcement learning approach. *2020 International Conference on Computing, Networking and Communications (ICNC)*, 2020.
- [10] A. Ferrari, M. Filer, K. Balasubramanian, Y. Yin, E. L. Rouzic, J. Kandrát, G. Grammel, G. Galimberti, and V. Curri. Gnpv: an open source application for physical layer aware open optical networks. *Journal of Optical Communications and Networking*, 12:31–40, 2020.
- [11] A. Haydari and Y. Yilmaz. Deep reinforcement learning for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23:11–32, 2020.
- [12] S. K. Kasi, S. Das, and S. Biswaso. Tcp congestion control with multiagent reinforcement and transfer learning. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, 2021.
- [13] G. Keiser. *optical communications essentials*. McGraw-hill, 2003.
- [14] F. C. N. O. Lima, C. J. A. Bastos-Filho, E. A. Barboza, and J. F. Martins-Filho. Osnr ripple and tilt: Comparison between pso and moo acop techniques for edfas links. *International Microwave and Optoelectronics Conference (IMOC)*, pages 1–3, 2021.
- [15] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Communications Surveys & Tutorials*, 21:1–6, 2019.
- [16] U. Moura, M. Garrich, H. Carvalho, M. Svolenski, A. Andrade, A. C. Cesar, J. Oliveira, and E. Conforti. Cognitive methodology for optical amplifier gain adjustment in dynamic dwdm networks. *Journal of Lightwave Technology*, pages 1971–1979, 2016.
- [17] U. C. Moura, J. R. F. Oliveira, R. L. Amgarten, G. E. R. Paiva, and J. C. R. F. Oliveira. Caracterizador automatizado de máscara de potência de amplificadores Ópticos para redes wdm reconfiguráveis. *Simpósio Brasileiro de Telecomunicações*, 2012.

- [18] F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini, and M. Tornatore. An overview on application of machine learning techniques in optical networks. *Journal of Lightwave Technology*, pages 1383 – 1408, 2018.
- [19] C. Natalino and P. Monti. The Optical RL-Gym: an open-source toolkit for applying reinforcement learning in optical networks. In *International Conference on Transparent Optical Networks (ICTON)*, page Mo.C1.1, July 2020.
- [20] J. Oliveira, A. Caballero, E. Magalhães, U. Moura, R. Borkowski, G. Curiel, A. Hirata, L. Hecker, E. Porto, D. Zibar, et al. Demonstration of edfa cognitive gain control via gmpfs for mixed modulation formats in heterogeneous optical networks. pages OW1H–2, 2013.
- [21] P. Poggiolini. The gn model of non-linear propagation in uncompensated coherent optical systems. *Journal of Lightwave Technology*, pages 3857 – 3879, 2012.
- [22] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.
- [23] S. J. Russel and P. Novig. *Artificial Intelligence A Modern Approach third Edition*. Pearson Education, 2010.
- [24] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal Policy Optimization Algorithms. *arXiv:1707.06347*, 2017.
- [25] J. M. Senior and M. Y. Jamro. *Optical Fiber Communications Principles and Practice*. Pearson education, 2009.
- [26] R. S. Sutton and A. G. Barto. *Reinforcement Learning An Introduction second edition*. The MIT Press, 2018.
- [27] X. Wang, Y. Fei, M. Razo, A. Fumagalli, M. Garrich, A. D. Andrade, M. S. Svolenski, and H. S. Carvalho. Effects of signal power control strategies and wavelength assignment algorithms on circuit osnr in wdm networks. *Photonic Network Communications*, pages 404–417, 2016.

- 
- [28] H. ZANG, J. P. JUE, and B. MUKHERJEE. A review of routing and wavelength assignment approaches for wavelengthrouted optical wdm networks. *OPTICAL NETWORKS MAGAZINE*, 12:47–60, 2000.