UNIVERSITY OF OTTAWA

OTTAWA-CARLETON INSTITUTE FOR MECHANICAL AND AEROSPACE ENGINEERING

Reinforcement Learning Application in Wavefront Sensorless Adaptive Optics System

Author:
Runnan Zou

A thesis submitted in partial fulfillment of the requirements for the Master of Applied Science degree
in
Mechanical Engineering

uOttawa

February 13, 2024

Reinforcement Learning Application in Wavefront Sensorless Adaptive Optics System

Runnan Zou

Abstract

With the increasing exploration of space and widespread use of communication tools worldwide, near-ground satellite communication has emerged as a promising tool in various fields such as aerospace, military, and microscopy. However, the presence of air and water in the atmosphere causes distortion in the light signal, and thus, it is essential for the ground base to retrieve the original signal from the distorted light signal sent from the satellite.

Traditionally, Shack-Hartmann sensors or charge-coupled devices are integrated in the system for distortion measurement. In our pursuit of a cost-effective system establishment with optimal performance and enhanced response speed, sensors and charge-coupled devices have been replaced by a photodiode and a single mode fiber in this project. Since the system has limited observation capability, it requires a powerful controller for optimal performance. To address this issue, we have implemented an off-policy reinforcement learning framework, the soft actor-critic, in the adaptive optics system controller. This integration results in a model-free online controller capable of mitigating wavefront distortion. The soft actor-critic controller processes the acquired data matrix from the photodiode and generates a two-dimensional array control signal for the deformable mirror, which corrects the wavefront distortion induced by the atmosphere, and refocusing the signal to maximize the incoming power.

The parameters of the soft actor-critic controller have been tuned to achieve optimal system performance.Simulations have been conducted to compare the performance of the proposed controller with respect to wavefront sensor-based methods. The training and verification of the proposed controller have been conducted in both static and semi-dynamic atmospheres, under different atmospheric conditions. Simulation results demonstrate that, in severe atmospheric conditions, the adaptive optics system with the soft actor-critic controller achieves more than 55% and 30% Strehl ratio on average in static and semi-dynamic atmospheres, respectively. Furthermore, the distorted wavefront's power can be concentrated at the center of the focal plane and the fiber, providing an improved signal.

Dedication

To my mom Qiuyan Du and my wife Yanrui Dong, who support me in pursuing truth and living happily.

Acknowledgments

I am incredibly grateful to my thesis supervisor, Dr. Davide Spinello, for his continuous support, invaluable advice, and patience throughout my master program. His guidance and support have been crucial in the development of my research topic and methodology, and he has created a supportive research atmosphere that has allowed me to explore all kinds of possibilities and seek out advice when needed. His mentorship has helped me become a more competent and mature researcher in the pursuit of truth.

I am also extremely thankful to my co-supervisors, Dr. Ross Cheriton and Dr. Colin Bellinger, for their valuable contributions to my thesis. Dr. Cheriton provides valuable suggestions for the establishment of the adaptive optics system and the formulation of my paper, and his professional expertise in optics is invaluable in helping me formulate my research problems. Dr. Bellinger patiently assists me with the building and training of artificial intelligence algorithms, which are essential to my research. Without their help, this thesis would not have been possible.

I am grateful to have had the opportunity to work with Mr. Payam Parvizi, a PhD candidates, who provided me with helpful assistance and advice throughout my thesis research. I also want to express my gratitude to my friends Hao Yan, You Huang, Zhiyan Qu, Ruikun Zhou, Zheng Tan, and Zebo Pan, who have supported me in both my personal and academic endeavors, especially when I was new to Canada.

Lastly, I want to thank my mother, Qiuyan Du, and my wife, Yanrui Dong, for their love and unwavering support. They have been my foundation as I venture into the unknown in the field of science.

Il n'ya qu'un héroïsme au monde: *c'est de voir le monde tel qu'il est et de l'aimer.*

Romain Rolland

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Communication plays a crucial role in the aerospace and military fields, driving the development of near-ground satellite communication. However, the presence of air, water and temperature difference in the atmosphere introduces distortion to communication signals which poses a significant challenge in the operation of free-space satellite-to-ground communication systems. Therefore, it is necessary to develope adaptive optics systems to correct the original wavefront from the disturbed wavefront. Adaptive optics is a technology that dynamically corrects the distorted wavefront in a feedback loop by using the command of a real-time controller which generates action based on wavefront measurement from wavefront sensors. The resulting actions are then applied to a distribution of actuators that act on the surface of a deformable mirror, which is a crucial component of an adaptive optics system used for wavefront correction.

Wavefront sensor-based (WFS-based) adaptive optics systems utilize Shack-Hartmann sensors [1], charge-coupled devices, and complementary metal-oxide-semiconductor image capture devices [2] as wavefront sensors for turbulence measurement. However, a significant fraction of the cost arises from the wavefront sensor, especially with infrared beams planned for optical satellite-to-ground links [3]. Moreover, wavefront sensors are limited in dynamic range, consume a fraction of the incident beam intensity, and introduce latency between the measurement and the actuation of the deformable mirror. This can result into outdated wavefront measurements as the satellite rapidly moves across the sky, introducing significant errors at the characteristic space-time scales defining this class of

systems. Consequently, researchers have started focusing on the development of wavefront sensorless adaptive optics systems that use cameras and reinforcement learning (RL) techniques [4]. Nevertheless, the high cost of cameras has hindered the development of small ground-based telescopes. As a solution, photodiode [5] and fibers [6, 7] have been introduced into the system for information acquisition.

While wavefront sensorless adaptive optics systems have been proposed for astronomy and microscopy that aim to maximize image quality, the optimization of an optical data link has significantly different requirements and has yet to be developed without a wavefront sensor. In the free-space optical communication, an adaptive optics system utilizes a controller to generate a control signal for the deformable mirror based on the observation signal received from the sensor on the focal plane. In order to design a controller that exhibits optimal performance, machine learning techniques are widely employed due to their ability to adapt to changes based on the measured effects on the control actions [8]. Reinforcement learning has emerged as a promising technique for addressing optical control challenges in the past decade [9]. Reinforcement learning techniques can be broadly classified into two categories: model-based and model-free methods. Model-based approaches are designed based on a known or learned model of the environment, aiming to approximate a global value or policy function [10]. Conversely, model-free reinforcement learning methods involve agents searching for an optimal policy through their interactions with the environment. This method's effectiveness stems from its capacity to make sequential decisions based on data sampled along system's trajectories and their effect on the environment. Through continuous interaction with the environment, model-free reinforcement learning can effectively learn and generate optimal actions for both simple and complex control problems. In order to develop an algorithm for a wavefront sensorless adaptive optics system controller that doesn't rely on the accurate modelling of the system, while also ensuring compatibility with systems featuring varying parameters, the integration of a model-free reinforcement learning method into the controller becomes imperative.

1.2 Objectives and Contribution

The primary objective of this thesis is to formulate an optimal online model-free controller for a wavefront sensorless adaptive optics system using reinforcement learning. The controller that is being proposed has been designed to function with

a photodiode, rather than utilizing either a wavefront sensor or a camera for the purpose of keeping response time and cost at a minimum. Despite the limited information provide by the photodiode, the adaptive optics system is still able to achieve a high Strehl ratio, which is the metric of how well an optical system is able to focus light to a point. To accomplish this objective, a model-free reinforcement learning controller will be developed to learn directly from the system's performance without a prior model of the system. The resulting reinforcement learning controller drives the adaptive optics system to achieve an optimal level of performance, while minimizing the response time and cost. The establishment of wavefront sensorless adaptive optics simulation environment supports the open source community and research.

The contributions are:

1. Developing an OpenAI gym package for the reinforcement learning training of wavefront sensorless adaptive optics environment based on HCIpy [11] in Python.

2. Adapting an online model-free off-policy reinforcement learning framework, the soft actor-critic, into the adaptive optics system for wavefront distortion correction with limited observations.

1.3 Outline

The structure of the thesis can be outlined as follows:

Chapter 2 provides an overview of research on adaptive optics systems, with a focus on the development of wavefront sensor-based and wavefront sensorless control methods.

Chapter 3 provides an overview of the foundational principles of reinforcement learning, with a specific focus on the soft actor-critic algorithm. The Soft actor-critic algorithm is integrated into the adaptive optics controller through the formulation of the wavefront sensorless adaptive optic optimal control problem. Additionally, we present a framework for hyperparameter optimization.

Chapter 4 provides a detailed discussion of the outcomes derived from hyperparameter optimization, along with an in-depth analysis of simulation results. These simulation results are presented within the contexts of both static and semi-dynamic atmospheres under a variety of conditions.

Chapter 5 summarizes the outcome of the project while suggesting possible future works to be done.

Chapter 2

Background and Literature Review

Section 2.1 provides an introduction to the optics system functions and their diverse applications. Section 2.2 delves into the context of free-space communication. Section 2.3 elucidates the fundamental components of adaptive optics, encompassing its working principle, atmospheric distortion, wavefront corrector, and wavefront sensor. Section 2.4 offers a comprehensive review of wavefront sensor-based methods, while section 2.5 presents wavefront sensorless control methods.

2.1 Optics System Functions and Applications

Optical systems have found widespread applications in various fields, including microscopy [12] , free-space communication [13] , and astronomy [3] . To achieve high-resolution optical imaging, it is imperative to focus light with high fidelity. However, the propagation of light through the optical system can perturb its fidelity. Consequently, the performance of the optical system is significantly affected by any disturbance along the light propagation path. In the field of biology, a non-uniform distribution of refractive index can cause blurring of the sample.

In the field of astronomy, telescopes deployed in space can observe celestial objects directly and without distortion. However, ground-based telescopes are more affordable and easier to maintain than their space-based counterparts. Unfortunately, even the best sites on earth, such as Paranal in Chile, suffer from atmospheric distortion that affects the quality of images produced by ground-based telescopes. This distortion occurs because the temperature and components of the atmosphere cause light to bend and become distorted, resulting in unclear images that lack detail.

To address this problem, scientists and engineers proposed the use of adaptive optics systems for ground-based telescopes as early as the last century [14, 15, 16]. These systems help mitigate the effects of atmospheric distortion by adjusting the telescope's optics in real-time to compensate for changes in the atmosphere. With adaptive optics systems, ground-based telescopes can produce images that are comparable in quality to those obtained from space-based telescopes.

Adaptive optics has a significant impact on system performance in the fields of astronomical imaging and free-space communication, with numerous applications. Since the successful deployment of diffraction-limited astronomical observations on ground-based telescopes in 1989, the adaptive optics system has become an essential component of large-aperture astronomical telescopes [17]. The high spatial resolution and sensitivity of adaptive optics make it ideal for precise scientific observations of dense multi-object and star fields. Thus, adaptive optics has evolved into an indispensable tool for ground-based astronomical research.

Adaptive optics technology has become an essential tool in correcting wavefront distortion induced by atmospheric turbulence in free-space optical communication systems. The use of adaptive optics can significantly enhance the efficiency of fiber coupling and minimize the bit error rate in free-space optical systems. An adaptive optics system consists of a wavefront sensor that measures the wavefront aberration on the upcoming wavefront and feeds it back to the controller as a feedback input. The controller employs an internal control algorithm to produce control signals for actuators to adjust a deformable mirror. This wavefront sensor-based adaptive optics system has already demonstrated success in fields such as astronomy observation and free-space communication.

However, a wavefront sensor-based adaptive optics system, while offering high performance, can come with complex system configurations, complicated calibration, and expensive costs. As a result, the adaptive optics field has called for a simpler structured system that can achieve optimal performance for observation purposes, leading to the demand for a wavefront sensorless system [18]. The wavefront sensorless adaptive optics system, as shown in Figure 2.1, has been developed to improve the performance of optical systems by correcting wavefront aberrations without relying on a wavefront sensor. Due to its simple structure, small size, and low cost, it has played a critical role in various fields, including astronomical observation [19, 20], microscopic imaging [21], and free-space communication [22].

Wavefront sensorless methods are used in the adaptive optics field instead of

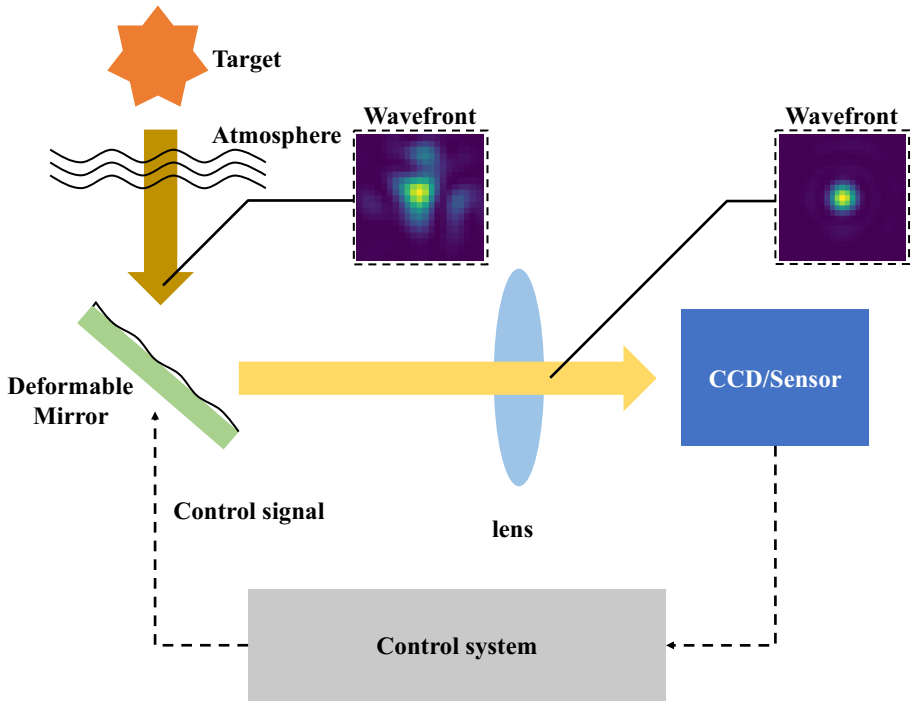


Figure 2.1: The Structure of Wavefront Sensorless Adaptive Optics System

measuring the wavefront distortion directly by a wavefront sensor. Instead, other information such as the point spread function and speckle pattern are used to speculate the wavefront aberration. The coming wavefront is then corrected by a deformable mirror based on the inferred wavefront distortion estimation. In comparison to wavefront sensor-based adaptive optics systems, wavefront sensorless adaptive optics systems offer several advantages, including a simple structure, budget-friendliness, and reduced light requirements. However, it is essential to acknowledge their limitations, such as decreased robustness in severe conditions and the necessity for a high-performance controller. Recently, deep learning and artificial neural networks have significantly advanced in the field of wavefront sensorless adaptive optics. Models such as convolutional neural networks, long short-term memory, deep reinforcement learning, and others have been used for wavefront aberration correction and deformable mirror control [23, 8, 9].

2.2 Background of Free-Space Optical Communication

The field of information and communication has seen significant growth and advancements in recent years. With the widespread use of high-speed internet, video conferencing, and live streaming, the demand for bandwidth and capacity has in-

creased dramatically. However, the traditional radio frequency spectrum is unable to meet this increased demand. To address this issue, a shift towards a method with higher bandwidth and capacity, such as optical carrier, is needed.

Free-space optical communication with optical carrier, is an attractive solution due to its high bandwidth and capacity and the lack of a requirement for spectrum licensing. A proposed structure, combining satellite-to-ground optical communication, satellite-to-satellite communication, and ground-to-satellite communication, is designed to meet the growing need for high-speed data transfer and extensive communication capabilities.

Compared to traditional radio frequency communication, free-space optical communication uses a different atmospheric transmission window in the near infrared wavelength range between 700 nm to 1600 nm under clear weather conditions, while the transmission window for radio frequency systems lies between 30 mm to 3 m. In terms of bandwidth, the usable bandwidth at an optical frequency of an optical carrier is almost 100,000 times that of a typical radio frequency carrier [24, 25]. Additionally, the beam divergence of optical carrier is narrower and smaller antennas can be used to achieve the same gain. The development of free-space optical communication systems is also free from the registration of spectrum, saving both cost and time. The high directivity of free-space optical communication also provides a high level of security, as signals are difficult to detect by spectrum analyzers and radio frequency meters [26]. With the increasing importance of security in the field of information technology, free-space optical communication systems provide an added advantage.

2.3 Adaptive Optics Function and Principle

The working principle and introduction of each parts in adaptive optics systems are delivered in this section.

2.3.1 Working Principle

An adaptive optics system based on wavefront sensors typically consists of three primary components: an instrument for measuring aberrations, a corrector to compensate for the aberrations, and a controller to manipulate the corrector based on the sensing signal received from the sensor. The Shack-Hartmann sensor is one of the most widely used sensors in adaptive optics systems due to its simplicity,

effectiveness, and rapid response.

There are also wavefront sensorless adaptive optics systems employ alternate methods to estimate wavefront distortions and make corrections, rather than directly measuring the distortion [27, 28, 29, 30]. By doing so, it eliminates the need for a wavefront sensor, which can be complex and expensive to implement. Instead, a point spread function capture device is placed on the focal plane to capture the point spread function, which is a measure of light spread from a point source and is utilized to estimate the wavefront aberration.

Adaptive optics is frequently employed in astronomy to correct for distortions in the wavefront of light originating from a celestial object. This technique relies on a guide star, which is a reference star situated in close proximity to the celestial body being observed and serves as a secondary light source. The guide star is selected based on its brightness, ensuring that it is easily detectable by the wavefront sensor. Additionally, since the guide star is in close proximity to the main objective, any wavefront distortions are assumed to affect both the guide star and the main objective in a similar manner.

By measuring the distortions in the wavefront of the guide star, it is possible to infer the distortions in the wavefront of the main objective. As a result, the wavefront sensor observes perturbations on the wavefront of the main objective with the assistance of the guide star. The feedback from the wavefront sensor is then utilized to correct the wavefront using a deformable mirror, producing a much clearer image of the main objective. A depiction of an astronomical adaptive optics system is presented in Figure 2.2. In a wavefront sensor-based adaptive optics system, the wavefront propagates through the various layers of the atmosphere, leading to the formation of an aberrated wavefront. To rectify this distortion, a deformable mirror is employed, which reflects the wavefront by adopting a deformed surface that is controlled by a control system. Subsequently, the wavefront sensor measures the diffraction present in the wavefront and transmits this information to the control system, enabling the generation of the surface configuration for the deformable mirror in the next time step.

In wavefront sensorless adaptive optics systems, the capture device records the point spread function and transmits the measurements to a controller. The controller utilizes this information to estimate the wavefront aberration, which in turn generates a control signal. This signal is utilized to regulate a deformable mirror, which corrects the estimated wavefront aberration. By using this method, the wavefront sensorless adaptive optics system is able to perform wavefront correc-

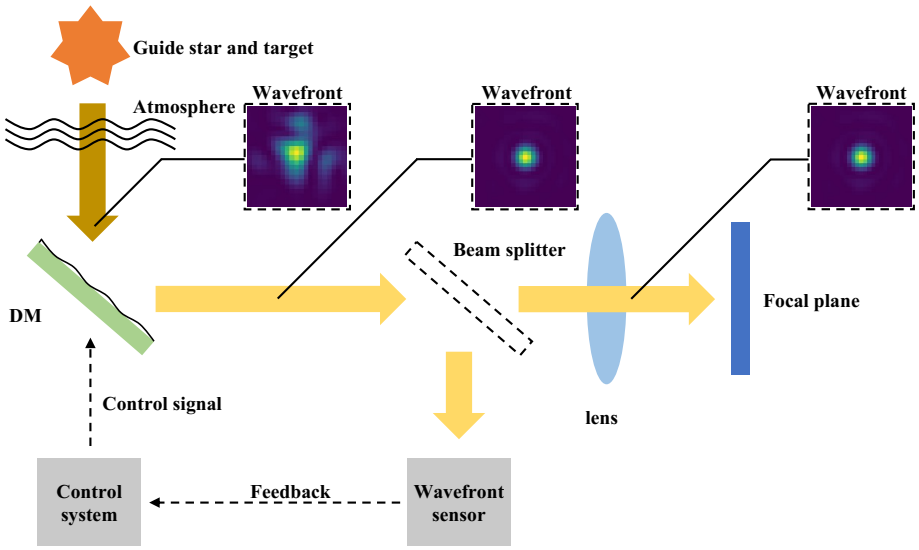


Figure 2.2: The Structure of Astronomy Adaptive Optics System With Wavefront Sensor

tion without the need for a wavefront sensor which reduces the complexity and cost of the system, while still allowing for high-quality imaging.

2.3.2 Distortion

Since the aim of adaptive optics system is to diminish the aberration on wavefront, it is important to study the properties of perturbations generated along the propagation path. Atmospheric turbulence, in the context of near ground communication, refers to the distortion that occurs in the wavefront of a signal as it passes through the Earth's atmosphere. This distortion is caused by several factors, including variations in atmospheric temperature and density distribution, as well as air flow. The uneven distribution and temperature of the air within the atmosphere act like a lens, which leads to bending and distorting of the light signal as it passes through. This phenomenon can result in a variety of visual distortions for astronomical objects, including blurring, twinkling, and shimmering. These distortions make it more difficult to obtain clear images and accurate measurements of such objects.

Atmospheric turbulence results from multiple phenomena, including convection, wind shear, and wind passing over objects. Convection arises due to the heating of the lower atmosphere, leading to convective gas bubbles rising and potentially forming cumulus clouds and lightning storms. Turbulence can also result from wind shear, which is a disparity in horizontal velocity between atmospheric

layers. Moreover, the flow of wind over objects like mountains or telescope domes can also induce turbulence. These various causes of turbulence can lead to "seeing" effects, causing wavefront to become distorted [31].

Atmospheric turbulence has two distinct effects on the wavefront, which are commonly referred to as "seeing" and "scintillation." Seeing pertains to random alterations in the direction of light entering a telescope, while scintillation is characterized by unpredictable changes in the intensity of the light. These effects are caused by fluctuations in the index of refraction, ultimately leading to a distorted wavefront.

Fried parameter is introduced [32] as a metric that quantifies the efficacy of optical transmission through the atmosphere due to variations in the atmosphere's refractive index which is introduced by the atmospheric turbulence. The Fried parameter is defined as the diameter of a circular area in which the root mean square wavefront aberration caused by passing through the atmosphere is equal to 1 radian. The units of length is expressed in centimeter, *cm*. The Fried parameter at wavelength of λ is expressed as [33]:

$$r_0 = [0.423k^2 \sec\zeta \int_{vertical} C_n^2(z)dz]^{-3/5} = (\cos\zeta)^{3/5} r_0^{\text{vertical}} \tag{2.1}$$

where the wavenumber is given by $k = 2\pi/\lambda$. The zenith angle, denoted by ζ, is defined as the angle between the local zenith and the position of interest, as illustrated in Figure 2.3. The variable z denotes the distance above the Earth's surface where atmospheric turbulence influences the propagation of light. It is worth noting that larger values of the zenith angle correspond to longer lines of sight through the atmosphere, resulting in increased turbulence and a reduction in the quality of the wavefront. The refractive index structure parameter of the atmosphere is represented by C_n^2 and is used as a statistical measure of the strength of turbulence in the atmosphere. Typical values of the coherence radius r_0 fall within the range of 5 cm to 20 cm.

Several models have been developed in the field of astronomy and free-space communication to express wavefront aberration. The Zernike modes parameters, which vary depending on the project's structure and instruments, have been collected based on numerous experiments [34, 35, 36, 37]. Typically, in order to conveniently predict aberration, turbulence is assumed to be a frozen flow on the millisecond time scale [38]. Consequently, turbulence is treated as a turbulent layer that flows across the sky at the speed of wind [39].

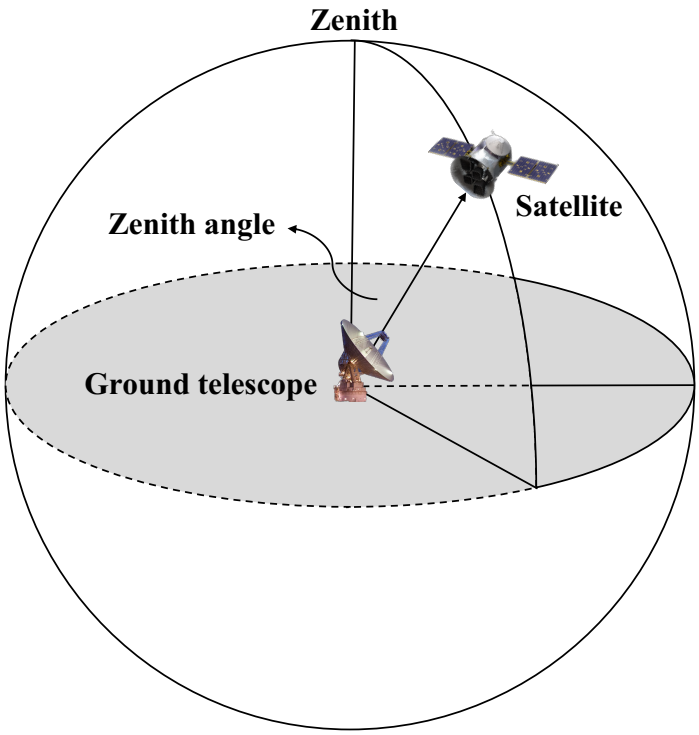


Figure 2.3: The Zenith Angle

Typically, scientists place their attention on identifying the type of aberration and observing its rate of change over time. To accurately illustrate the wavefront, Zernike polynomials are frequently employed as a mathematical tool [40]:

$$W(r,\theta) = \sum_{n,m} C_n^m Z_n^m(r,\theta) \tag{2.2}$$

the polar coordinates (r,θ) are utilized to denote the normalized pupil radius, which ranges between 0 and 1, and the angle measured in a clockwise direction from the y-axis, respectively. Moreover, C denotes the amplitudes of the Zernike polynomial $Z_n^m(r,\theta)$, where m and n represent the azimuthal frequency and radial order, respectively. Due to their orthogonality, each Zernike polynomial has an independent effect on the optical system.

2.3.3 Wavefront Corrector

In an adaptive optics system, the role of wavefront correctors is crucial as they are responsible for correcting aberrations on the wavefront and restoring it to its original form. There are three types of correctors commonly used in adaptive optics systems: deformable mirrors, liquid crystal spatial light modulators, and de-

formable phase plates. Among these, the deformable mirror is widely employed due to its effectiveness. Deformable mirrors come in two forms: continuous and segmented. Both types are capable of being employed in adaptive optics systems. However, the segmented deformable mirror is more commonly used, owing to its flexibility and ease of use. The mirror's deformable nature allows it to change its shape in response to the aberrations present in the incoming wavefront, thereby enabling it to compensate for the distortions in the wavefront.

A segmented mirror is a mirror composed of numerous small, stiff segments that can be adjusted individually to alter the mirror's shape. Each segment is capable of moving in one or more directions, and the movement of one segment does not have a significant impact on the others. However, the gaps between the segments can result in light and diffraction losses. Furthermore, when neighboring segments are not aligned, a step may appear on the mirror's surface, which can lead to additional diffraction losses. To address these concerns, some versions of segmented mirrors are equipped with three actuators per mirror, which enable each segment to be tilted [41].

Segmented mirror devices are commonly manufactured through micro electromechanical systems on a silicon platform, enabling a relatively low-cost production of a significant number of segments. However, it is essential to note that the production cost increases with the number of segments due to the need for a larger chip area. Furthermore, as the number of segments increases, the chances of faults within the device also increase, reducing the production yield. Consequently, this technology is most appropriate for mirrors that have a small size but a large number of segments. One of the problems encountered with segmented deformable mirrors is the need for protection from oxidation, which can cause parasitic reflections, even with window anti-reflection coatings. It is crucial to use an optical window to shield the segmented mirror from oxidation. Segmented mirrors are often utilized in applications where high image quality is not a strict requirement. Nevertheless, they can prove suitable for situations where small size, numerous degrees of freedom, and cost are significant factors to consider.

A continuous deformable mirror is a specialized type of mirror designed with a surface that can be continuously altered. This surface is typically composed of a thin sheet of glass coated with either metallic or dielectric materials. The surface deformation is facilitated by actuators located on either the back or the side of the glass surface. The continuous surface of a deformable mirror differs from a segmented one, as the deformation of a particular region can affect its neighbor-

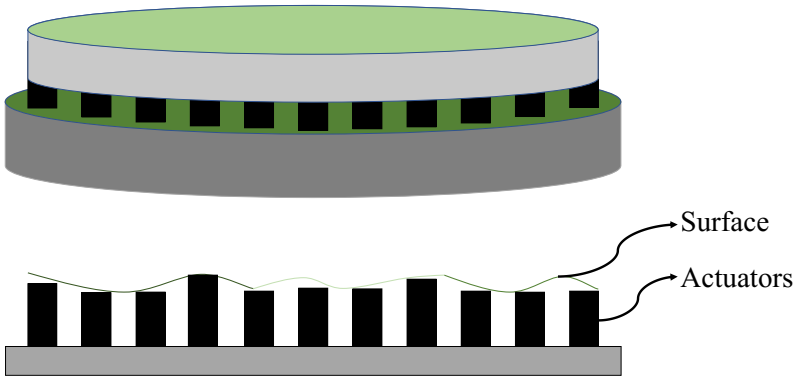


Figure 2.4: The Structure of a Continuous Deformable Mirror

ing areas, leading to some coupling effects. Not only does each actuator impact the shape of the immediate surrounding area, but it also affects a larger, overlapping region that encompasses the areas of neighboring actuators. While this phenomenon could be problematic, it is possible to accurately determine the voltage applied to the actuators by utilizing computer software that takes into account the coupling effects. Furthermore, feedback systems can be employed to address this issue. The structure of a continuous deformable mirror is shown in Figure 2.4.

A deformable mirror with a continuous surface and fine-tuned adjustment capabilities can achieve high optical performance with minimal power loss, making it ideal for applications that require precision and accuracy. The mirror's continuous surface also allows for a broad range of deformations, making it a versatile tool suitable for various uses.

2.3.4 Wavefront Sensor

The Shack-Hartmann wavefront sensor is the prevalent tool employed for measuring the shape of incoming light's wavefront in optical telescopes. Its nomenclature derives from the surnames of Johannes Franz Hartmann and Roland Shack, the pioneering researchers who developed the sensor. Specifically, this instrument is capable of gauging the wavefront of attenuated laser beams or starlight, making it a versatile device for a range of applications.

The Shack-Hartmann sensor is made up of a microlens array and an image sensor placed in the focal plane of the microlens array. The working principle of the sensor is simple. The incoming radiation is focused onto a spot on the sensorby the microlens. By analyzing the location of a specific spot, the overall direction of the wavefronts across the microlens entrance can be determined. A computer

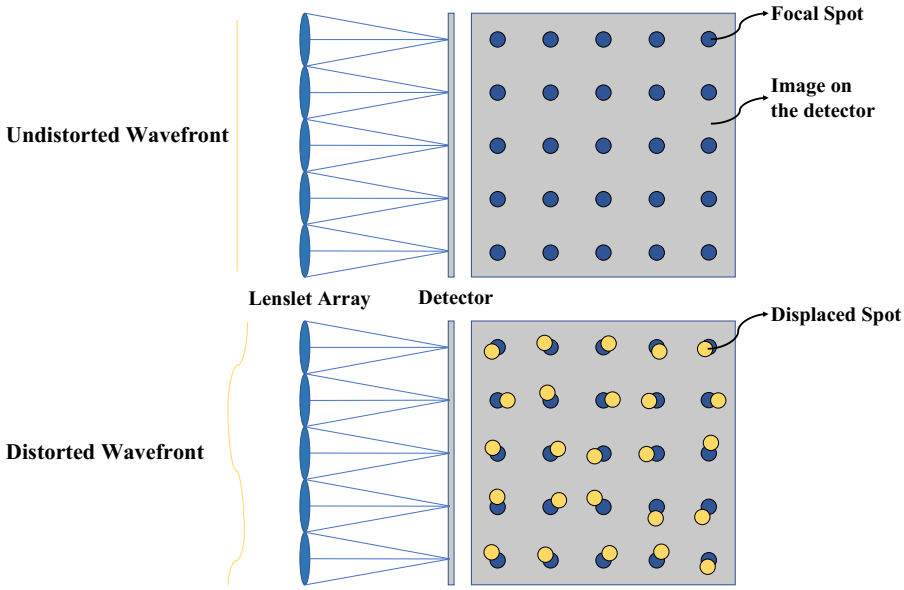


Figure 2.5: The Structure and Working Principle of a Shack-Hartmann Sensor

program is utilized to determine the positions of similar spots for all microlenses, using the obtained image. This process allows for the estimation of any distortions in the wavefronts across the entire entrance area of the sensor. Figure 2.5 displays the structure and working principle of the Shack-Hartmann sensor.

The Shack-Hartmann sensor comprises an arrangement of lenslets situated in a plane that is conjugate to the pupil, with a camera located on the focal plane of the lenslets. In the absence of aberration, the camera captures an evenly distributed array of spots. However, in the presence of aberration, the position of the spots is displaced in relation to the type of aberration. The Shack-Hartmann sensor detects the aberration by analyzing the location of the displaced spots on the camera. The spatial resolution of the image sensor is limited, and therefore determining the position of a spot based solely on the pixel with the highest optical intensity is not entirely accurate. To achieve higher accuracy, a more effective method involves computing the "center of gravity" by utilizing the first moments of the intensity distribution. This method can provide a higher position resolution than the pixel spacing. Advanced computational algorithms that are proficient in reducing noise can also provide more precise data. In challenging measurement situations such as large phase excursions or significant noise, the choice of numerical algorithm utilized becomes critical in determining the quality of the output data. Additionally, it is essential to make efforts to minimize or eliminate the effects of cross-talk, which is the impact of light from neighboring lenses.

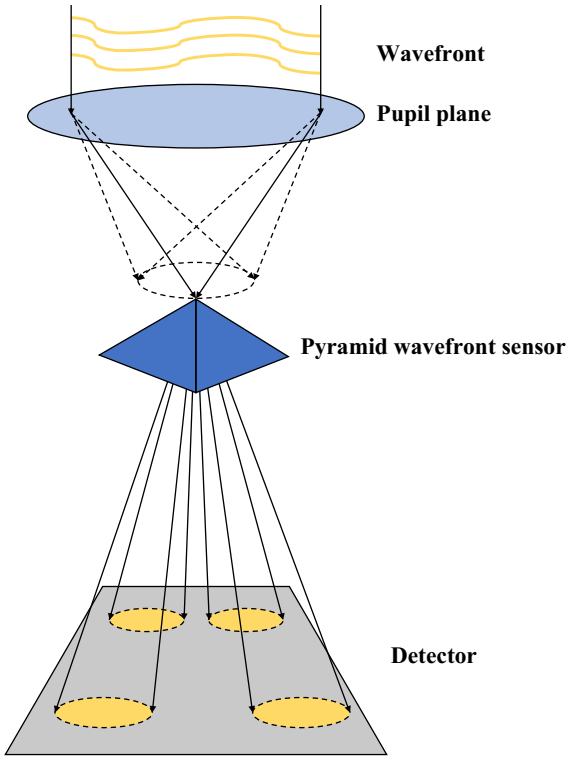


Figure 2.6: The Structure and Working Principle of a Pyramid Sensor

The Shack-Hartmann sensor has been commonly used in adaptive optics projects, but its popularity as a go-to sensor has been challenged by the emergence of a newer sensor, known as the pyramid sensor. The pyramid sensor boasts of independently adjustable dynamic range and sensitivity, providing more versatility than the Shack-Hartmann sensor. Due to this unique feature, the pyramid sensor can be utilized in a broader range of applications, making it a formidable alternative to the Shack-Hartmann sensor. A number of studies have supported this notion [42, 43, 44]. To provide a visual representation, the structure of the pyramid sensor is illustrated in Figure 2.6.

2.4 Wavefront Sensor-Based Control Methods

The adaptive optics system can be categorized into two types of control: wavefront sensor-based control and wavefront sensorless control. The primary distinction between these two control methods lies in the use of a wavefront sensor. The wavefront sensor-based control approach has been under development for several decades, due to its stability and high accuracy. However, given the cost and time latency factors associated with our free-space communication project, we prefer to

employ the wavefront sensorless control method, which is less expensive to deploy and operates at a higher speed.

The term wavefront sensor-based adaptive optics refers to the use of deformable mirrors that are driven by feedback wavefront sensor measurements in order to compensate for wavefront distortion caused by atmospheric turbulence [24, 25, 45, 46]. The phase profile or wavefront is estimated by the reconstruction algorithm based on the observation of wavefront sensor. The controller recovers slopes from each of the Shack-Hartmann sensor lenslets and applies them to the deformable mirror via the command matrix [47].

The proportional-integral-derivative (PID) controller has been extensively utilized in industrial control due to its stable performance and straightforwardness [26, 48]. Fundamentally, the PID algorithm is employed in static reference input tracking systems, where the control algorithm consists of three parameters: proportional (k_p), integral (k_i), and derivative (k_d) factors [49, 50, 51]. These three parameters are tuned and designed by designers or other optimization methods.

PID is seen widely used in the field of adaptive optics to control deformable mirrors by tracking wavefront shape. Wu et al. proposed a PID controller to control the surface shape of magnetic fluid deformable mirrors based on LMI-Based multivariable [52]. Ke et al. applied a fuzzy PID control algorithm was used to adjust the control parameters to complete the closed-loop control of the adaptive optics [53].

Although the PID control algorithm has proven to be a reliable method in various fields, its performance in complex and nonlinear systems may not always be optimal. This is especially true for systems with numerous parameters and intricate intrinsic relationships, making it difficult to derive an accurate mathematical model of the system [54, 55]. While optimization methods can be utilized to fine-tune the parameters k_p, k_i, and k_d [56, 57, 58], engineers typically rely on their prior experience to conventionally select and tune these parameters [59]. However, due to the complexity of the system, relying on these conventional methods may not always produce satisfactory results.

To control an adaptive optics system without requiring gain pre-tunning, adaptive control is applied with an adaptive law to enable the controller to continuously adapt to changes in the adaptive optics system.Chang and Gibson proposed a method based on adaptive control [60, 61, 62], where a filter and a controller were built for wavefront construction for linear time-invariant systems. To achieve performance in a more complex system, Liu achieved reconstruction and predic-

tion of wavefront by employing adaptive filtering in an adaptive optics system. Their adaptive control loop produce significant improvement in the point spread function of the adaptive optics system [63]. Despite of the great performance of adaptive control in adaptive optics, the aberration and irregularity of the system might cause the controller to be unable to guarantee a stable output while it also comes at a relatively large computational cost [64].

Considering the operation time of the optics system, with the aim of deriving and predicting the disturbance of atmosphere, predictive control is deployed to solve the optimization problem. To address the time lag in the control loop, Dessenne et al. designed model predictive control in adaptive optics system with closed-loop prediction and a synthesized linear model predictor [65]. By calculating postcoronagraph contrast, Males et al. presented a predictive controller based on linear prediction formalism for a ground-based adaptive optics system [66]. With the development of data-driven methods, a predictive control was applied on a single conjugate adaptive optics system based on a spatial-temporal dynamic model which was developed with data acquired from Shack-Hartmann wavefront sensor [67]. In aspect of vibration compensation on telescope of Extremely Large Telescope, Glück et al. designed a model predictive control to compensate the atmospheric turbulence and structural vibration on an Extremely Large Telescope [68].

Predictive control necessitates a precise model of system dynamics to ensure the effectiveness of the control algorithm. Additionally, a significant volume of data is essential for constructing accurate models, particularly in the case of nonlinear or high-dimensional systems. These challenges render the development of predictive control challenging for cost-effective adaptive optics systems, which face limitations in both data availability and the establishment of an accurate model.

Predictive control has also become a prominent method with the emergence and development of neural networks. Based on atmospheric models derived from history observation data, future turbulence was predicted and the control of deformable mirrors was therefore implemented [69]. As an improvement of the recursive neural network, long short-term memory is able to predict the future with sequential history data without the problem of vanishing gradients and gradient explosion during training [70, 71]. Compared with its predecessor, recursive neural network, long short-term memory passes historic data from the previous layer to the next layer. In long short-term memory, there are gates to regulate the flow of information. These gates learn which features or patterns are important or not to

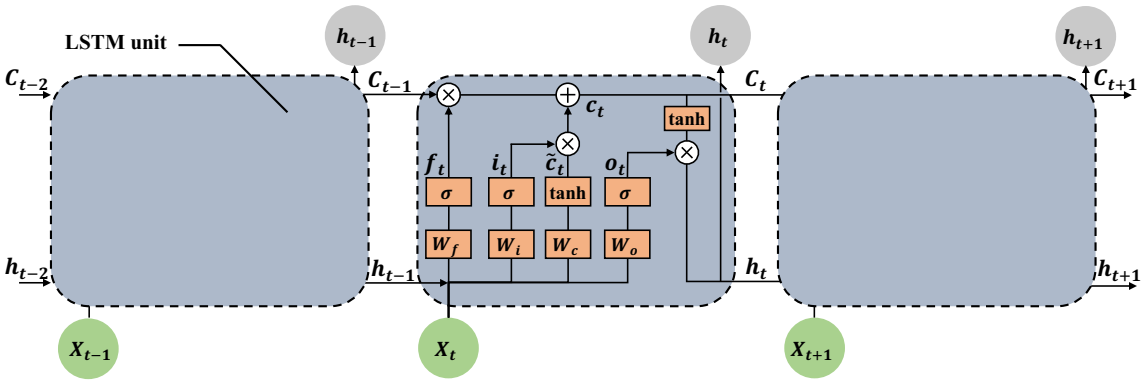


Figure 2.7: The Structure of Units in Long Short-Term Memory Algorithm

produce a good result. By updating the weights in the neural network, the more important data is kept for future prediction and useless information is forgotten. The structure of long short-term memory is shown in Figure 2.7. A long short-term memory unit is composed of an input gate, an output gate, and a forget gate.

Zernike parameters were used to depict wavefront aberration predicted by long short-term memory [72, 69, 73, 71]. The derived Zernike parameters were then applied to wavefront correction. With consideration of the correlation between historical data, since they were in form of an array and matrix, the convolutional neural network is commonly combined with long short-term memory in applications [47]. Swanson et al. first proposed and demonstrated the feasibility of the application of convolutional neural networks and long short-term memory in wavefront prediction and reconstruction. To further improve the performance of long short-term memory based control method, they trained long short-term memory and convolutional neural networks with generative adversarial networks [74]. With this novel supervised closed loop method, the robustness was increased in different seeing conditions.

Predictive control demands a precise model of system dynamics to ensure the effectiveness of the control algorithm. Additionally, a substantial volume of data is essential for constructing accurate models, particularly in the case of nonlinear or high-dimensional systems. Utilizing Long Short-Term Memory and Recurrent Neural Network, the predictive accuracy of the model surpasses that of traditional predictive control, as these advanced techniques leverage labeled sequential data for supervised learning. Nevertheless, the acquisition of labeled data poses a formidable challenge for cost-effective adaptive optics systems that lack accurate equipment for measuring wavefront aberrations and other environmental factors. These systems encounter limitations in both data availability and the establish-

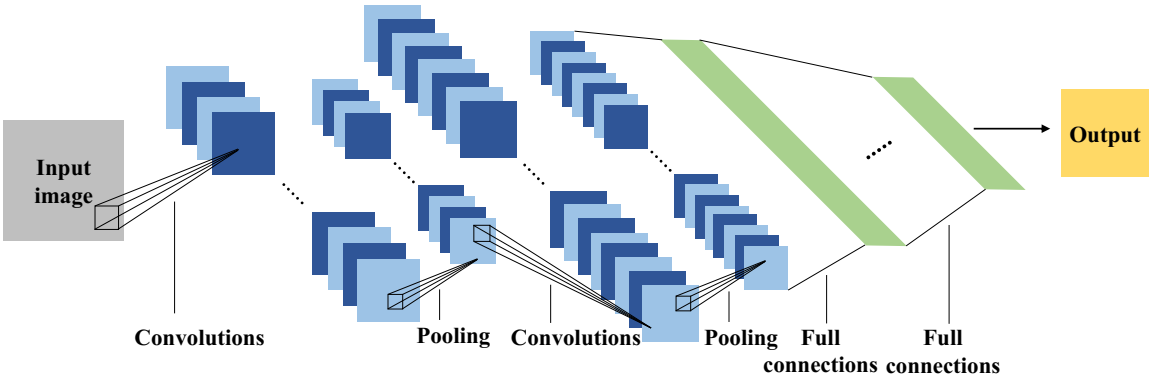


Figure 2.8: The Structure of Convolutional Neural Network

ment of an accurate model.

Convolutional neural networks and deep convolutional neural networks are developed the past few decades as effective and robust methods for image and video processing [75]. With a large volume of data that forms the training data set, convolutional neural networks is able to fit the functions which reflect the relationship between input data and output features. In many aspects of image and video processing problems, convolutional neural networks is adopted and proven to have a significant role [76, 77]. Neural networks have been successfully applied in adaptive optics before the emergence of convolutional neural networks [78]. Without the component of convolution, the temporal and spatial correlation in images was not capable to be derived from data. The structure of convolutional neural networks is shown in Figure 2.8. An input layer, hidden layers, and an output layer form the convolutional neural networks. The input of convolutional neural networks is a tensor of images with a shape of $(number\ of\ images) \times (input\ height) \times (input\ width) \times (input\ channels)$. By inputting into convolutional layers, the feature of images is convolved, derived, and passed to the next layer. To overcome overfitting with a large volume of data, pooling is used to reduce the size of features. The most common approach is max pooling. By connecting to several fully connected layers, the output is generated.

In order to deal with the convergence failure when facing large wavefront aberration, Paine et al. applied convolutional neural networks on initial wavefront phase estimation [79]. By inputting the point spread function, the initial wavefront phase is estimated by the generated Zernike parameters. The method of estimating Zernike parameters based on convolutional neural networks has been improved by obtaining from intensity image [80, 81, 82] and attention-based convolutional

neural networks [83]. Also, by combining with long short-term memory, the adaptive optics system is able to leverage the ability of long short-term memory and convolutional neural networks to obtain the aberration trend over time [47, 74].

Compared with the traditional control sequence with wavefront sensor, the image-based convolutional neural networks method is not limited by the quantity of wavefront sensor lenslet and the error brought by wavefront sensor like misregistration [82]. Yet, the imaged-based convolutional neural networks method suffers from the huge demand for computing resources and the requirement of the training data set. Also, the deployment of an image capture device increases the cost of the adaptive optics system while reducing its flexibility of the system. Therefore, in order to develop a system with low cost and complexity as well as relatively high effectiveness, a wavefront sensorless adaptive optics structure and control algorithm is selected.

2.5 Wavefront Sensorless Control Methods

The intrinsic nonlinearity and non-derivative properties of the optics system pose significant challenges in developing the wavefront sensorless adaptive optics technique. It is difficult to establish a direct solution to reflect the relationship between the surface of the deformable and the measurements on the focal plane. Consequently, the utilization of iterative optimization algorithms becomes imperative to address these issues.

Wavefront sensorless adaptive optics methods can be broadly classified into two principal categories: model-based methods and model-free methods. Typically, model-based approaches in wavefront sensorless techniques involve the initial establishment of a foundational basis function serving as the system model. Subsequently, the corresponding system control algorithm is derived from the established model. In the domain of wavefront sensorless adaptive optics systems, a phase-retrieval technique was developed to indirectly ascertain the unknown phase wavefront from focal-plane intensity measurements for the free-space communication adaptive optics system [84]. A self-calibration procedure was proposed aiming to obtain the Gram matrix without relying on a wavefront sensor in a model-based wavefront sensorless adaptive optics system [85]. The Gram matrix, thus obtained, can be directly employed for simultaneous correction.

In contrast to the model-based approaches, model-free wavefront sensorless adaptive optics methods present a more cost-effective implementation, provid-

ing more adaptability across diverse systems [86]. The model-free wavefront sensorless adaptive optics technique, widely acknowledged for its extensive adoption, concentrates on optimizing received signal performance without necessitating wavefront reconstructions. Consequently, it simplifies both system and control algorithm complexities, making it well-suited for a wide spectrum of wavefront correctors. The model-free methodology typically harnesses optimization algorithms to enhance system performance. Over the past decades, numerous optimization algorithms have been developed. Gradient descent optimization techniques such as stochastic parallel gradient descent (SPGD) [87, 88] and genetic algorithms [89] have been adopted to address wavefront sensorless adaptive optics aberration correction.

In wavefront sensor adaptive optics, reinforcement learning has already been applied as a promised approach for coping with temporal delay or calibration error. By driving the optics system model from data, model-based wavefront sensor adaptive optics method has been developed [90, 91, 92, 93]. Wavefront sensor adaptive optics model-free method like recursive deep policy gradient [94] and multi-agent reinforcement learning for adaptive optics [95] were developed. Recursive deep policy gradient provided meaningful insight into combining prediction and decision. The formulation of the recursive neural network and deep deterministic policy gradient algorithm compensates for the observation deficiency effectively. In the field of wavefront sensorless adaptive optics, without the assistance of wavefront measurement, the corrector is directly controlled by sensing results on the focal plane. By deploying the only charge-coupled device camera behind the focal lens, the deep deterministic policy gradient algorithm cooperated with convolutional neural networks for wavefront correction [96, 97]. The idea of building wavefront sensorless adaptive optics is attractive with its low cost and complexity of structure. However, without directly or indirectly sensing wavefront aberration, the controller needs to formulate a policy based on a long time of training and large data sets which brings a great challenge.

2.6 Summary

This chapter described into the function and principles of adaptive optics systems. It provided a comprehensive overview of the existing literature on control methods for adaptive optics systems, which included both wavefront sensor-based and wavefront sensorless control methods. The following chapter will introduce rein-

forcement learning algorithms to enhance the understanding of their principles. A state-of-the-art reinforcement learning method will be introduced and illustrated how it can be integrated into the controller of our wavefront sensorless adaptive optics system.

Chapter 3

Optimal Control Algorithm and Problem Formulation

Section 3.1 delves into the background of reinforcement learning. Section 3.2 introduces the deep reinforcement learning built based on the combination of reinforcement learning and neural networks. Section 3.3 presents a comprehensive explanation of the soft actor-critic algorithm as it is applied within the context of the wavefront sensorless adaptive optics system. The formulation of the controller, which is derived from the soft actor-critic algorithm, can be found in section 3.4. The hyperparameters tunning setup is detailed in section 3.5.

3.1 Reinforcement Learning

With the world-renowned success of AlphaGo [98], reinforcement learning has been applied in control in the field of astronomy [99], automobile [100, 101] and multi-agent control [102, 103, 104]. Reinforcement learning agent learns control tasks by interacting with the environment while updating its policy for better performance. Without prior knowledge, reinforcement learning is able to find the optimum control solution to the problem. The interaction between the reinforcement learning agent and environment is shown in Figure 3.1.

Reinforcement learning has gained significant attention in recent years due to its effectiveness in optimal control. Through its ability to interact with the environment and adjust actions based on received feedback rewards, a reinforcement learning controller can eventually acquire an optimal policy. In this regard, we propose the use of a reinforcement learning framework to formulate a wavefront

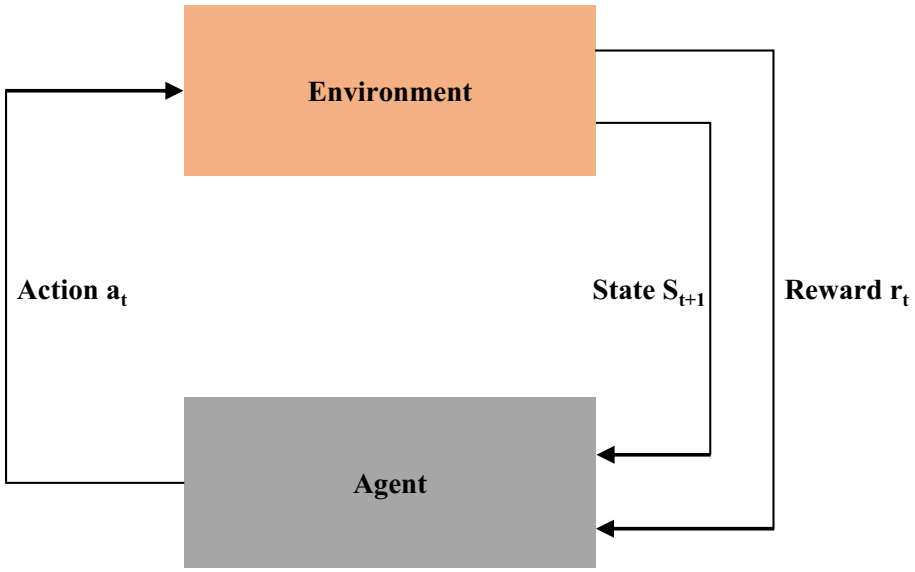


Figure 3.1: The Figure of Interaction Between Reinforcement Learning Agent and Environment

sensorless adaptive optics control system. In this chapter, we will first provide an overview of reinforcement learning and its introduction. We will then discuss its application in the wavefront sensorless adaptive optics system, highlighting the benefits it can offer in terms of improved performance and reduced complexity.

Different from supervised learning and unsupervised learning, reinforcement learning allows an agent to be trained online in an environment that changes over time. Notably, reinforcement learning has demonstrated superior performance to human players in games such as Atari and GO [105]. The reinforcement learning system consists of several components, including a controller (agent), environment, action, state, reward, and next state. At each time step t, the agent receives a state $s_t \in S$, where S is the set of possible states and s_t represents the state of the environment at time t. Based on this input and the agent's internal algorithm, an action $a_t \in A$ is generated, where a_t represents the action taken in the environment at time step t and A is the set of possible actions. As the environment transitions to the next time step $(t+1)$, the state of the environment changes from s_t to s_{t+1} in response to the action a_t and an immediate reward r_{t+1} is observed. The immediate reward $r_{t+1} \in R$ is a measure of the model's performance based on the optimization goal set prior to training. Specifically, r_{t+1} reflects the agent's performance at time step $t+1$ when applying a_t in state s_t. In the design of the wavefront sensorless adaptive optics, the reward is formulated based on the measurement of corrected wavefront quality.

3.1.1 Markov decision process

In general, the current and future state of a process can be influenced by information and states from past history. This phenomenon is known as a stochastic process in probability theory.

$$P(s_t, a_t, s_{t+1}) = P(s_{t+1}|s_t, ..., s_0, a_t, ..., a_0) \tag{3.1}$$

the notation $P(s_t, a_t, s_{t+1})$ represents the joint probability that the environment state transitions from s_t to s_{t+1} based on the action a_t. For each possible state $s_t \in S$ at time step t, the probability of transitioning from state s_t to state s_{t+1} depends on the previous states $s_0, ..., s_t \in S$ and the previous actions $a_0, ..., a_t \in A$. In the context of reinforcement learning, a memoryless process called the Markov Decision Process (MDP) is employed to model the environment. The Markov Decision Process is a discrete-time stochastic control process and serves as the foundation of reinforcement learning. The Markov Decision Process is defined by a tuple (S, A, R, P) that is generated when the reinforcement learning controller interacts with the environment. In this tuple, $s_t \in S$ represents the environmental state, $a_t \in A$ denotes the action taken by the controller, $r_t \in R$ is the reward that is earned based on a requirement acquired at each time step, and P represents the transition probability of state s at time t with action a resulting in a change to state s' at time $t + 1$. The transition probability is defined by the equation:

$$P(s_t, a_t, s_{t+1}) = P(s_{t+1}|s_t, a_t) \tag{3.2}$$

Equation 3.2 describes the memoryless structure of Markov Decision Process. In a stochastic process with the Markov property, the future state of the model solely depends on the present state.

3.1.2 Cumulative Reward

In the reinforcement learning, the term "reward" refers to a measurement of an agent's performance in a single time step, based on the model state and action taken. On the other hand, "return" represents the cumulative reward that an agent receives over a longer period of time. As the primary objective of optimization control in the reinforcement learning is to find an optimal policy that maximizes the long-term reward acquired from the environment, the reward received in each step influences action decision making and ultimately impacts the long-term cu-

mulative reward.

The design of the reward function is based on a description on the optimal problem with the objective of enabling the agent to adapt to the environment and learn its behavior, leading to an optimal long-term reward. The relationship between the reward received in a given time step, denoted by R_t, and the return at that same time step, denoted by G_t, is defined as follows:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{3.3}$$

the variable G_t represents the discounted sum of immediate rewards and future rewards that are discounted by a factor $0 < \gamma \leq 1$. Here, γ denotes the discount factor, a weighting factor that determines the relative significance of immediate reward and future reward in the calculation of the discounted return. Additionally, the discount factor γ is employed to balance the trade-off between exploration and exploitation when searching for the optimal action in action space. If the value of the discount factor is lower, the influence of future reward on the current reward is reduced, while a higher discount factor indicates that a long-term reward is being pursued.

3.1.3 Value Functions

When dealing with Markov Decision Process, it is important to determine which state or action is optimal among all possible choices. This is done by evaluating the cumulative reward and applying the Markov property. The state value function is a concept used to measure the value of a particular state. It is defined as the expected return of a given state. Mathematically, the state value function is

represented as follows:

$$\begin{aligned} V_\pi(s) &= E_\pi(R_t|s_t = s) \\ &= E_\pi(\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|s_t = s) \\ &= E_\pi(r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2}|s_t = s) \\ &= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma E_\pi(\sum_{k=0}^{\infty} \gamma^k R_{t+k+2}|s_t = s')] \\ &= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma V_\pi(s')] \end{aligned} \tag{3.4}$$

V_π represents the state value of a given state s under policy π. Here, policy π maps the current state $s_t \in S$ to an action $a_t \in A$. On the other hand, E_π refers to the expected value when following policy π. Additionally, $P^a_{ss'}$ and $R^a_{ss'}$ respectively denote the probability and immediate reward of transitioning from state s to state s' through action a.

To evaluate the quality of an action a in a given state s, the action-state value is introduced. This value represents the goodness or badness of applying the action a at a particular state s. Similarly, the action-state value function, denoted by $Q_\pi(s,a)$, can be defined as follows:

$$\begin{aligned} Q_\pi(s,a) &= E_\pi(R_t|s_{=}s, a = a_t) \\ &= E_\pi(\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|s = s_t, a = a_t) \\ &= E_\pi(r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2}|s_t = s, a_t = a) \\ &= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma E_\pi(\sum_{k=0}^{\infty} \gamma^k R_{t+k+2}|s_t = s', a_t = a')] \\ &= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma Q_\pi(s', a')] \end{aligned} \tag{3.5}$$

There exists a relationship between state value function $V_\pi(s)$ and action-state value function $Q_\pi(s,a)$ as shown in the Figure 3.2. For each a state s_t, there exists a value function $V(s_t)$ denotes the value of this state. Based on the current policy and the state s_t, the potential actions a^n_t ($a \in 1, 2, 3, ...$) are generated while their state-action value function is presented by $Q(s_t, a_t)$. For each action, the state

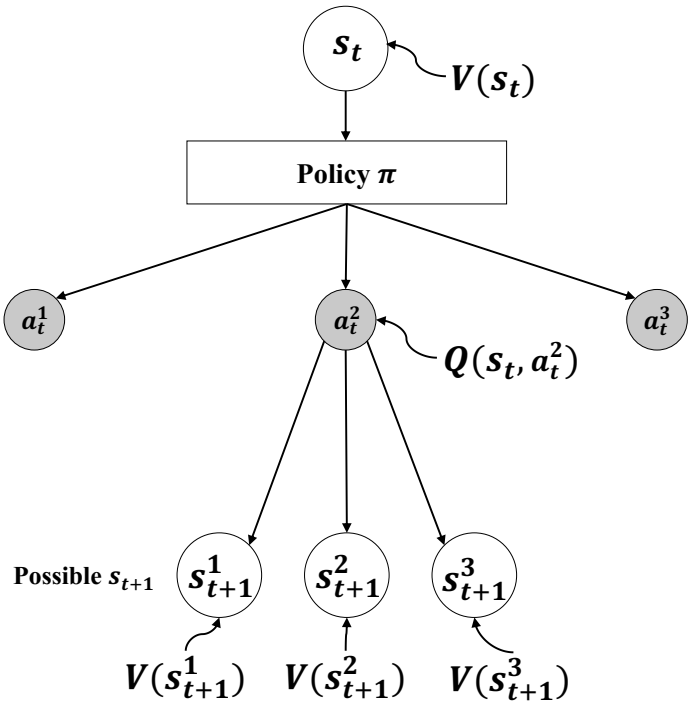


Figure 3.2: The Relationship Between the State Value Function and the Action-State Value Function [106]

has the possibility to transfer from s_t to s_{t+1}^n ($n \in 1, 2, 3, ...$) which comes with their state value function $V(s_{t+1}^n)$ ($n \in 1, 2, 3, ...$). Therefore, the relationship between state value function $V_\pi(s)$ and action-state value function $Q_\pi(s, a)$ is:

$$V_\pi(s) = \sum_a \pi(s, a) Q_\pi(s, a) \tag{3.6}$$

$$Q_\pi(s, a) = \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma V_\pi(s')] \tag{3.7}$$

In an optimal control problem, the ultimate goal is to find the optimal policy that generates the highest cumulative reward. If policy π' generates a better cumulative reward than policy π'', then it is considered to be the superior policy:

$$V^{\pi''}(s) \leq V^{\pi'}(s) \quad \forall s \in S \tag{3.8}$$

Among various state value functions, the optimal state value function is the one that yields the highest value and is associated with the optimal policy. Thus, the optimal state value function can be expressed as the maximum possible value that can be achieved by following the optimal policy:

$$V^*(s) = \max_{\pi} V_\pi(s) \tag{3.9}$$

where $V^*(s)$ is the optimal state value function. The optimal action-state value function therefore can be expressed as

$$Q^*(s,a) = \max_{\pi} Q_\pi(s,a) \tag{3.10}$$

where $Q^*(s,a)$ is the optimal action-state value function. Also, based on Bellman Equation:

$$V_\pi(s) = E_\pi(r_{t+1} + \gamma V_\pi(s_{t+1}) | s_t = s) \tag{3.11}$$

The state value function and the action state value function in their optimal states can be used to formulate the Bellman optimal equations in forms of:

$$V^*(s) = \max_{a} \sum_{s'} [R^a_{ss'} + \gamma V^*(s')] \tag{3.12}$$

$$Q^*(s,a) = \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma \max_{a} Q^*(s',a')] \tag{3.13}$$

The equation 3.12 and 3.13 are the Bellman optimal equation of state value function $V(s)$ and Bellman optimal equation of action state value function $Q(s,a)$.

3.1.4 Value Iteration and Policy Iteration

In reinforcement learning, the Bellman equation is used to calculate the state value and action state value, enabling it to find the optimal policy by iterating between $Q(s,a)$ and $V(s)$. To begin the value iteration, an arbitrary random value of state value $V(s)$ is set, and then the operations of $V(s)$ and $Q(s,a)$ are conducted until $V(s)$ converges. The update rule for value iteration is given below.

$$V'(s) = \max_{a} \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V(s')] \tag{3.14}$$

After the convergence of the iteration, the policy is acquired by

$$\pi(s) \leftarrow \operatorname*{argmax}_{a} \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V(s')] \tag{3.15}$$

where the notation $\operatorname*{argmax}_{a}$ denotes the set of values a that result in the maximum

state value function.

The value iteration method improves the state value function through iterative updates. On the other hand, the policy iteration method focuses on improving the policy itself by creating a strictly improved policy in each iteration. Initially, the policy iteration method begins with a random policy, denoted as π_0. Next, the evaluation of the initial policy, π_0, is conducted using an equation called policy evaluation, which can be expressed as follows:

$$\begin{aligned} V^{\pi_0}(s) &= E_{\pi_0}(r_{t+1} + \gamma V^{\pi_0} | s_t = s) \\ &= \sum_a \pi_0(s,a) \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma V^{\pi_0}(s')] \end{aligned} \tag{3.16}$$

After policy evaluation, policy is improved by considering the greedy policy. The policy improvement is given by

$$\pi'(s) \leftarrow \underset{a}{\operatorname{argmax}} \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma V^{\pi}(s')] \tag{3.17}$$

Once the policy evaluation process reaches convergence, the policy iteration is terminated, and the optimal policy is obtained.

3.2 Deep Reinforcement Learning

Deep reinforcement learning is an advanced technique that integrates reinforcement learning and deep learning. The incorporation of neural networks into the field of reinforcement learning was catalyzed by their remarkable success in image processing and natural language processing. Specifically, neural networks are utilized as, function approximations and value generators for value estimation and action formulation, respectively.

The actor-critic approach is a specialized version of reinforcement learning that integrates an actor and a critic to improve decision-making in a given environment. By integrating neural network with it, the actor-critic method is able to deal with continuous control problems. Unlike value-based methods, the actor is primarily responsible for generating decisions based on the input state, while the critic evaluates the performance of the actions generated by the actor. To recall the introduction of policy iteration in the section 3.1.4, the actor is responsible for policy improvement, while the critic focuses on policy evaluation. The structure of the actor-critic approach is depicted in Figure 3.3 [106].

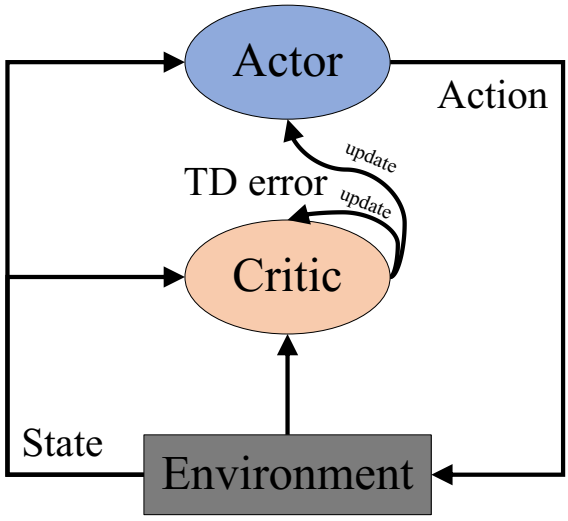


Figure 3.3: The Architecture of an Actor-Critic Method

The feedback provided by the critic enables the actor to improve the policy-generating algorithm and optimize its decision-making ability. During training, the critic itself also converges to a stable state, providing a reliable estimation of actions. Notably, several methods are variations of the actor-critic approach, including deep deterministic policy gradients, asynchronous advantage actor-critic, among others [107, 108].

The (s_t, a_t, r_t, s_{t+1}) is acquired from the environment based on the current actor policy $\pi_\theta(s, a)$. An action $\tilde{a}_{t+1}$, which will not be applied in the environment, is also sampled from the policy. V_t and V_{t+1} are calculated by

$$V_t = V(s_t, a_t; w_t) \tag{3.18}$$

$$V_{t+1} = V(s_{t+1}, \tilde{a}_{t+1}; w_t) \tag{3.19}$$

where w_t is the parameter in critic network. Then, the TD-error is calculated based on the immediate reward and the estimated value for s_{t+1} by:

$$\delta_t = r_t + \gamma V_{t+1} - V_t \tag{3.20}$$

The critic network is updated by conducting gradient:

$$w_{t+1} = w_t - \alpha \delta_t \frac{\partial V(s_t, a_t; w)}{\partial w}|_{w=w_t} \tag{3.21}$$

The actor network is updated by gradient ascent:

$$\theta_{t+1} = \theta_t + \beta V(s_t, a; w) \frac{\partial log \pi(a|s, \theta)}{\partial \theta}|_{\theta=\theta_t} \tag{3.22}$$

3.3 Soft Actor-Critic

The soft actor-critic algorithm is an off-policy actor-critic method that is based on the maximum entropy reinforcement learning framework. The main objective of the actor in this algorithm is to maximize both the expected return and the information entropy, which in turn helps to find the optimal policy for the given task while allowing for random actions. The concept of maximum entropy is incorporated into the algorithm by adding information entropy to the reward function formulation. This feature gives soft actor-critic its name, as it is considered a "soft" method in comparison to other algorithms that solely aim to achieve the best expected reward. Soft actor-critic goes beyond that by seeking to maximize the policy distribution entropy, which induces more exploration of the environment. Compared to just seeking the best cumulative reward, this approach is relatively soft, as it considers the exploration of the environment as well. The optimal policy for a finite process in discrete time can be expressed as [109]

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \sum_t E_{(s_t, a_t) \sim \rho(\pi)} [r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t))] \tag{3.23}$$

where the π^* is the optimal policy and π are all possible policies. (s_t, a_t) are the observations and actions generated based on policy distribution $\rho(\pi)$ and $r(s_t, a_t)$ is the corresponding reward. $\mathcal{H}(\pi(\cdot|s_t))$ denotes the information entropy of the generated action distribution. The variable α represents the temperature parameter, which is the weight of the information entropy in the reward function. Increasing the temperature parameter value causes the actor to become more inclined towards exploring the action space, resulting in a greater level of stochasticity for the optimal policy. During the process of policy training, the temperature parameter is automatically tuned.

In infinite horizon problems, a discount factor is introduced to guarantee that

the summation of future expected rewards remains finite.

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \sum_{t=0}^{\infty} E_{(s_t,a_t)\sim\rho(\pi)}[\sum_{l=t}^{\infty} \gamma^{l-t} E_{(s_l,a_l)\sim\rho(\pi)}[r(s_l,a_l) + \alpha\mathcal{H}(\pi(\cdot|s_l))|(s_t,a_t)]] \tag{3.24}$$

where $(s_l, a_l) \sim \rho(\pi)$ are the future observations and actions of time step t, $r(s_l, a_l)$ is the future reward of time step t and $\mathcal{H}(\pi(\cdot|s_l))$ is the corresponding information entropy.

Applying the maximum entropy algorithm in reinforcement learning encourages the policy to explore extensively in the action space while also prioritizing actions that promise a high future return. In cases where the policy generates multiple sets of actions with similar rewards, contrary to traditional reinforcement learning methods such as deep Q learning and deep deterministic policy gradient, these actions are given an equal probability of being selected, leading to comprehensive exploration [110]. This improved exploration enhances both policy performance and learning speed. With the addition of entropy, soft policy evaluation is generalized by [109]

$$\mathcal{T}^{\pi}Q(s_t,a_t) \triangleq r(s_t,a_t) + \gamma E_{s_{t+1}\sim p}[V(s_{t+1})] \tag{3.25}$$

$$V(s_t) = E_{a_t\sim\pi}[Q(s_t,a_t) - \alpha \log \pi(a_t|s_t)] \tag{3.26}$$

the operator $\mathcal{T}^{\pi}$ is known as the Bellman backup operator, which is typically utilized to update the value function with the provided value function. $Q(s_t, a_t)$ is the action-state value of (s_t, a_t). $V(s_t)$ is the state value of state s_t. $s_{t+1} \sim p$ is the state sampled from the state distribution p. The convergence of the Bellman backup operator has been guaranteed, ensuring that the sequence of Q will converge to the soft Q function.

During the process of policy improvement, it is necessary to develop policies that are manageable and easy to implement. To achieve this, the policy distribution Π is transformed into a Gaussian distribution using the Kullback-Leibler (KL) divergence. The KL divergence, also referred to as relative divergence, is utilized to measure the distinction between two probability distributions. Specifically, for continuous variables, the KL divergence is defined as [111]:

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx \tag{3.27}$$

where $D_{KL}(p(x)||q(x))$ is the Kullback-Leibler divergence between distribution $p(x)$ and $q(x)$.

Therefore, in the soft policy improvement process, the policy is improved by [109]

$$\pi_{new} = \underset{\pi' \in \Pi}{\operatorname{argmin}} \, \mathrm{D_{KL}}(\pi'(\cdot|s_t)||\frac{\exp(\frac{1}{\alpha}Q^{\pi_{old}}(s_t,\cdot))}{Z^{\pi_{old}}(s_t)}) \tag{3.28}$$

where $\pi'(\cdot|s_t)$ is the policy distribution which is generated by adding deviation on the old policy distribution. $\exp(\frac{1}{\alpha}Q^{\pi_{old}}(s_t,\cdot))$ is the exponential of the old policy distribution. The function $Z^{\pi_{old}}(s_t)$ is utilized to normalize the distribution. By finding the minimal Kullback-Leibler divergence between potential policies and the exponential of old policy distribution, the improved policy can be guaranteed in terms of its soft value [109]. The parameters of policy is improved by minimizing the Kullback-Leibler divergence of equation 3.27:

$$J_\pi(\phi) = \mathrm{E}_{s_t \sim \mathcal{D}}[D_{KL}(\pi_\phi(\cdot|s_t)||\frac{\exp(Q_\theta(s_t,\cdot))}{Z_\theta(s_t)})] \tag{3.29}$$

where D_{KL} is the Kullback-Leibler divergence, $\pi_\phi(\cdot|s_t)$ is the policy distribution with state s_t. ϕ is the parameter of policy network and θ is the parameter of soft Q value network.

Neural networks are commonly utilized as function approximators in actor-critic models, where they serve as policy and soft Q function estimators, respectively. In this framework, there exist five distinct neural networks, which include an actor network, a soft value network, a target value network, as well as two soft Q networks (namely, Q_1 and Q_2). The target value network is applied to lead the convergence of the soft value network. The two soft Q value networks are set for the purpose of stabilizing the learning process and preventing overfitting. While it is possible to derive the soft state value function using the soft Q function and policy, as shown in equation 3.26, introducing a separate neural network as a value approximator can significantly improve training stability [112]. The value and Q networks are responsible for the policy evaluation. The process of soft actor-critic algorithm is shown in Table 3.1. The soft value network is updated to minimize the square residual error [109]:

$$J_V(\psi) = \mathrm{E}_{s_t \sim \mathcal{D}}[\frac{1}{2}(V_\psi(s_t) - \mathrm{E}_{a_t \sim \pi_\theta}[Q_\theta(s_t,a_t) - \log \pi_\phi(a_t|s_t))^2]] \tag{3.30}$$

where the symbol ψ represents the parameter of the soft value network, while $\mathcal{D}$

refers to the distribution of the states in the replay buffer. The policy network is denoted by π_ϕ, where ϕ corresponds to the parameter of the actor network. Additionally, θ is the parameter of the soft Q network. To ensure stability during the training process, we use the smaller $Q(s_t, a_t)$ generated by two networks. We conduct gradient descent on $J_V(\psi)$ with the aim of minimizing the residual error by [109]:

$$\nabla J_V(\psi) = \nabla_\psi V_\psi(s_t)(V_\psi(s_t) - Q_\theta(s_t, a_t) + \log \pi_\phi(a_t|s_t)) \tag{3.31}$$

The two soft Q networks are updated in the same procedure by performing gradient descent on the Bellman residual [109]:

$$J_Q(\theta) = \mathrm{E}_{(s_t,a_t)\sim\mathcal{D}}[\frac{1}{2}(Q_\theta(s_t, a_t) - \mathcal{T}^{\pi_k} Q_\theta(s_t, a_t))^2] \tag{3.32}$$

$$= \mathrm{E}_{(s_t,a_t)\sim\mathcal{D}}[\frac{1}{2}(Q_\theta(s_t, a_t) - (r(s_t, a_t) + \gamma E_{s_{t+1}\sim p}[V_\psi(s_{t+1})))^2]$$

with

$$\nabla_\theta J_Q(\theta) = \nabla_\theta Q_\theta(s_t, a_t)(Q_\theta(s_t, a_t) - r(s_t, a_t) - \gamma V_{\bar{\psi}}(s_{t+1})) \tag{3.33}$$

where the output of target soft value network $V_{\bar{\psi}}$ is applied to generate an estimation of a target Q value. The update for the actor network is constructed using equation 3.36. To apply the gradient descent method for minimizing equation 3.35, a reparametrization technique is utilized to transform the actor [109].

$$a_t = f_\phi(\varepsilon_t; s_t) \tag{3.34}$$

where ε_t is the input noise which is applied here to generate a different mean and covariance of a policy π'. The equation 3.29 therefore written as:

$$J_\pi(\phi) = \mathrm{E}_{s_t\sim\mathcal{D},\varepsilon_t\sim\mathcal{N}}[\log \pi_\phi(f_\phi(\varepsilon_t; s_t)|s_t) - Q_\theta(s_t, f_\phi(\varepsilon_t; s_t))] \tag{3.35}$$

And gradient of the equation 3.35 is given by [109]:

$$\nabla_\phi J_\pi(\phi) = \nabla_\phi \log \pi_\phi(a_t|s_t) + \nabla_\phi f_\phi(\varepsilon_t; s_t)(\nabla_{a_t} \log \pi_\phi(a_t|s_t) - \nabla_{a_t} Q_\theta(s_t, a_t)) \tag{3.36}$$

The update rule of neural networks is shown in Figure 3.4. The output of policy network is designed as the covariance and mean of Gaussian distributions. The distributions are then applied to sample actions which will be applied in the

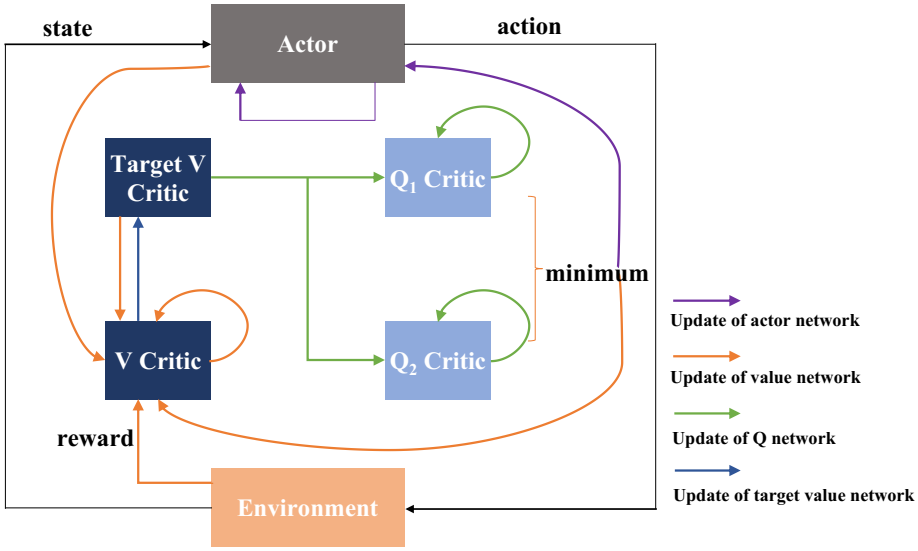


Figure 3.4: The Update of Soft Actor-Critic

environment.

3.4 Problem Formulation

When incorporating an optical control problem into reinforcement learning, it is important to indicate the state, action, and reward. Our wavefront sensorless adaptive optics system employs a photodiode on the focal plane to observe the power distribution on the focal plane, represented as $o(k)$. This photodiode helps reduce the requirement for expensive and sluggish read-out circuits in infrared cameras. An observation obtained from the photodiode is depicted in Figure 3.5.

The purpose of this study is to propose a cost-effective approach to the adaptive optics establishment by reducing the number of actuators beneath the surface of the deformable mirror to 16. To achieve a balance between the capability of correcting wavefront aberrations and the cost of the deformable mirror, we select a deformable mirror featuring actuators with dimensions of 4×4. By adopting this system configuration, the output of the reinforcement learning is a 16-dimensional continuous action space. It should be noted that the action amplitude is limited by the deformable mirror stroke.

The observation $o(k)$ is given by

$$o(k) = \begin{bmatrix} o^1(k) & o^2(k) \\ o^3(k) & o^4(k) \end{bmatrix} \tag{3.37}$$

Algorithm 1: Soft Actor-Critic

Initialize the parameter of networks ψ, $\bar{\psi}$, θ_1,θ_2, ϕ
for each episode do
  for each step do
    sample a_t from π_ϕ
    observe s_{t+1} and r_t by applying a_t into system
    store (s_t, a_t, r_t, s_{t+1}) into replay buffer $\mathcal{D}$
  end for
  for each gradient step do
    sample a batch of (s_t, a_t, r_t, s_{t+1}) from $\mathcal{D}$
    update soft Q_1 network: $\theta_1 \leftarrow \theta_1 - \lambda_Q \nabla_{\theta_1} J_Q(\theta_1)$
    update soft Q_2 network: $\theta_2 \leftarrow \theta_2 - \lambda_Q \nabla_{\theta_2} J_Q(\theta_2)$
    update soft value network: $\psi \leftarrow \psi - \lambda_V \nabla_\psi J_V(\psi)$
    update actor network: $\phi \leftarrow \phi - \lambda_\pi \nabla_\pi J_\pi(\phi)$
    update target value network: $\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}$
  end for
end for

Table 3.1: The Pseudocode of the Soft Actor-Critic Algorithm

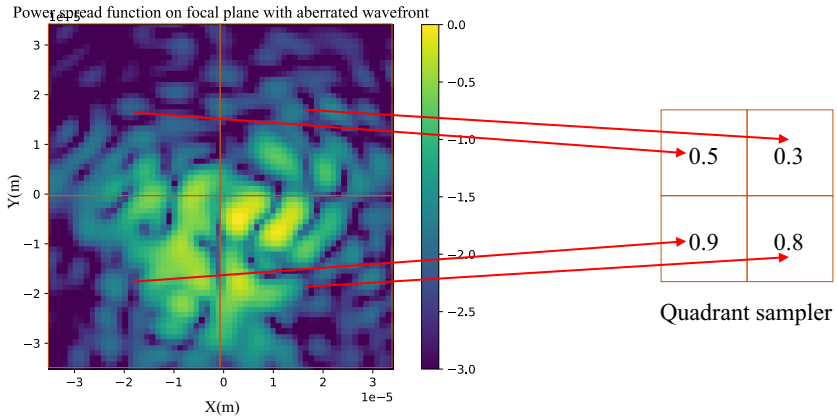


Figure 3.5: The Working Principle of Photodiode

where $o^n(k)$ denotes the power of the n-th pixel at time step k on the photodiode on focal plane, $n \in (1, 2, 3, 4)$. The photodiode subsamples the four corners of the power distribution on the focal plane by measuring the power intensity. In order to process the observation in neural networks, the matrix $o(k)$ is flattened and denoted as $x(k)$:

$$x(k) = [x^1(k), x^2(k), x^3(k), x^4(k)] \tag{3.38}$$

The amplitude of each small unit on the deformable mirror is defined as the action of the system, which is given by

$$a(k) = [a^1, a^2, ..., a^{16}] \tag{3.39}$$

where $a(k)$ stands for the control signal applied on the deformable mirror at the time step k, and $a^n(k), n \in (1, ..., 16)$ represents the amplitude of each unit on the surface of the deformable mirror at time step k. The output action is represented by a shape of 4×4 array:

$$a(k) = \begin{bmatrix} a^1(k) & a^2(k) & a^3(k) & a^4(k) \\ a^5(k) & a^6(k) & a^7(k) & a^8(k) \\ a^9(k) & a^{10}(k) & a^{11}(k) & a^{12}(k) \\ a^{13}(k) & a^{14}(k) & a^{15}(k) & a^{16}(k) \end{bmatrix} \tag{3.40}$$

The output of the reinforcement controller specifies of the displacements of actuators located beneath the surface of the deformable mirror. The surface of the deformable mirror is controlled by the action of these actuators, which manipulate the incoming wavefront according to an influence function. In our modeling, we apply the xinetics influence function. The influence function defines how local changes to the surface of the deformable mirror affect the wavefront passing through the mirror. Thus, it establishes a relationship between the surface shape of the deformable mirror and the phase of the incoming light. The actuators can move within a range of approximately $\pm 500nm$, thereby enabling precise control over the mirror's surface shape and position.

The reward function in reinforcement learning has a significant impact on the training result. A proper reward function can lead the training to an optimal policy. In this study, the reward function is formulated by the Strehl ratio. The Strehl ratio is an important metric used to assess the quality of corrections exerted by adaptive optics systems. Specifically, it is defined as the ratio of the peak intensity of an

aberrated image generated by an incoming wavefront point spread function to the peak intensity of an ideal, diffraction-limited optical system point spread function under the same instrument and observing conditions [113]. It should be noted that there exists an alternative definition of the Strehl ratio, which involves computing the ratio of the peak intensity at the aberrated image center to the unaberrated peak intensity at the center. Both definition of Strehl ratio work properly in our project since the peak intensity of the unaberrated wavefront locates exactly on the center of the focal plane. In either case, the Strehl ratio takes on values between 0 and 1, with higher values indicating a more focused point spread function and a better correction of the adaptive optics system. The definition of Strehl ratio is

$$S = e^{-\sigma^2} \tag{3.41}$$

where σ is the root mean square deviation over the aperture of the wavefront phase.

For the entropy of the generated action distribution, in each step, the policy network creates the mean and variance for the 16 actuators, which collectively constitute a 16-dimensional Gaussian distribution. The entropy of this Gaussian distribution measures the level of uncertainty and randomness within the action distribution. It represents the agent's exploration ability within the environment. A high entropy indicates that the action space is more uniformly distributed, and the agent has a greater chance of exploring unknown action spaces, ultimately resulting in the discovery of the optimal action distribution. In our soft actor-critic controller, the entropy of an action is calculated as

$$\mathcal{H} = \sum_{n=1}^{16} \mathcal{H}_n \tag{3.42}$$

where $\mathcal{H}_n, n \in (1, ..., 16)$ is the information entropy of each actuator on the deformable mirror. Therefore, the entropy of the action is denoted by the summation of each actuator's information entropy. The aim of training of the soft actor-critic is maximizing the wavefront correction and improving the exploration capabilities of the controller.

With the defined state, action, and reward, the reinforcement learning frame is embedded in our adaptive optics systems. At the time step t, based on the current state s_t, the agent's policy generates the action, $\pi : s \rightarrow a$ which then manipulates the shape of the deformable mirror surface. With the controlled deformable mirror, the focal plane photodiode observes the state s_{t+1} and the reward r_t. For an optimal

policy, the expected output of the system is a concentrated spot on the focal plane without scintillation.

3.5 Hyperparameters Tuning Setup

Hyperparameters of the wavefront sensorless adaptive optics controller are predetermined values that are set before the learning process commences, and optimizing the performance and generalization ability of an algorithm is heavily reliant on these values. Nevertheless, determining the optimal values for hyperparameters in advance can be a challenging task, which is why the need for optimization arises. It is important to carefully consider and adjust these parameters to ensure the algorithm performs efficiently and can generalize well. Specifically, the learning rate of the actor and critic, the discount factor, and the neural network structure are essential parameters that need to be investigated. To achieve this, the sweep function integrated in WandB, and the parallel training framework, META, in Compute Canada have been utilized to simulate hyperparameters training in an efficient manner.

Before conducting hyperparameter searching, it is important to establish the range of hyperparameters based on prior knowledge of the simulation. The range of hyperparameters is presented in Table 3.2. Bayesian hyperparameter search is used to conduct the hyperparameter optimization process. This technique utilizes Bayesian statistics to identify the optimal hyperparameters for machine learning algorithms effectively. The method constructs a probabilistic model based on Bayesian theory to estimate the performance of each hyperparameter combination. It selects the next parameter combination that is most likely to optimize the model performance for testing, and updates the model each time a new combination of parameters is tested. This iterative process leads to the identification of the optimal hyperparameter combination with limited computational resources.

Compared to traditional grid search or random search methods, Bayesian hyperparameter search offers several advantages. For example, it can usually find the optimal hyperparameter combination faster and with the same or fewer computational resources, while achieving better performance. Thus, Bayesian hyperparameter search is a valuable tool for optimizing machine learning algorithms and improving their predictive power.

To conduct a comprehensive comparison and analysis of the impact of each hyperparameter, we assigned discrete values to the actor learning rate, ranging

Hyperparameters	**Value**
actor learning rate	1×10^{-5} - 5×10^{-5}
critic learning rate	1×10^{-3} - 5×10^{-3}
discount factor	0.95 - 0.99
batch size	32 - 512
layer size	32 - 512

Table 3.2: The Range of Hyperparameters

from 1×10^{-5} to 5×10^{-5}, and to the critic learning rate, ranging from 1×10^{-3} to 5×10^{-3}. It is common practice to set the critic learning rate higher than that of the actor since the critic needs to update its value function estimates quickly and accurately to provide meaningful feedback to the actor. A higher critic learning rate allows for faster learning from feedback, leading to reduced variance in value estimates and more precise and stable guidance for the actor. On the other hand, the actor's learning rate can be set lower as its role is to gradually update its policy using the critic's value function estimates. A lower actor learning rate can prevent large policy changes that may cause instability and oscillations in the learning process. Additionally, we selected the discount factor from the values of 0.95 to 0.99, the batch size from 32 to 512, and the layer size from 32 to 512 to ensure a comprehensive evaluation of the hyperparameters. To gain a deeper comprehension of the impact of hyperparameters on simulation results, which can significantly aid in the formulation of real-time policies, an investigation has been conducted on the effectiveness of various hyperparameters. Specifically, the actor learning rate, critic learning rate, discount factor, batch size, and layer size are analyzed individually to evaluate their influence on the simulation outcome.

3.6 Summary

This chapter aims to introduce the principle of the soft actor-critic algorithm. By formulating the wavefront sensorless adaptive optics into the optimal control problem, the soft actor-critic algorithm is embedded into the adaptive optics controller. Initially, the structure and principles of the soft actor-critic algorithm are discussed, highlighting the role of entropy in enabling the agent to search more extensively for the optimal policy and avoid local optima. Due to its ability to search widely and converge efficiently, the soft actor-critic algorithm is incorporated into an adaptive optics controller with the defined state, action, and reward. The hyperparam-

eters optimization setup of the controller is discussed for the subsequent optimization discussions in the next chapter.

Chapter 4

Simulations and Results

This chapter presents a comprehensive examination of the simulation setup and the outcomes of the conducted simulations. The simulation setup is elaborated in section 4.1. The modeling of the atmosphere is addressed in section 4.2. The section 4.3 delves into the discussion of hyperparameters. To initially validate the effectiveness of the proposed controller, a simulation with a static atmosphere is performed, as outlined in section 4.4. Furthermore, to obtain a more profound understanding of the performance of the proposed controller, it is implemented in a semi-dynamic atmosphere, elucidated in section 4.5.

4.1 Simulation Set Up

This section describes the simulation setup that has been established to assess the potential of the proposed reinforcement learning approach and the adaptive optics system, which is designed to achieve high responsiveness at minimal costs. The adaptive optics system serves as the environment for the reinforcement learning controller, which is implemented using the Python packages HCIPy and PyTorch. The simulation environments generated for the reinforcement learning of an adaptive optics system provide static and semi-dynamic simulations. The static environment assumes a stable atmospheric condition, allowing the use of a fixed turbulence profile for training purposes. In the semi-dynamic environment, the static atmosphere's configuration changes in each episode, accounting for the influence of wind in the sky. In this context, the atmospheric movement is affected by a velocity vector, ensuring continuity between the final timestep of episode i and the initial timestep of episode $i+1$. Implementing a reinforcement learning controller

within a semi-dynamic environment proves beneficial for the initial deployment of the adaptive optics system in a genuinely dynamic environment.

HCIPy is an object-oriented, open-source framework for Python that enables simulations of high-contrast astronomical imaging instruments. It provides various libraries related to adaptive optics, including wavefront generation, atmospheric turbulence modeling, propagation simulation, fiber coupling, and the implementation of deformable mirrors and wavefront sensors. By integrating wavefronts and optical elements, HCIPy simulates the entire wavefront propagation process from space to the telescope. The reinforcement learning controller is implemented using PyTorch, a package that supports several tasks in the fields of machine learning and deep learning.

The simulations are conducted on the high-performance computing (HPC) platform Compute Canada. We use the Cedar node, which integrates a central processing unit (CPU), graphics processing unit (GPU), and random access memory (RAM), to execute the simulation code. The central processing unit is used for the generation of the atmospheric model, while the graphics processing unit is harnessed to the training of neural networks within the soft actor-critic controller.

To validate the efficacy of the proposed soft actor-critic controller, our initial focus is on evaluating its performance in static atmospheric environments with varying levels of atmospheric turbulence, represented by the parameter D/r_0, where D represents the diameter of the telescope, and r_0 represents the Fried parameter. A higher D/r_0 value indicates more severe atmospheric turbulence.

4.2 Modeling of the Atmosphere

The atmospheric model is composed of four crucial elements: a signal source, a deformable mirror, a controller, and a focal plane receiver. The signal source emits a wavefront, as shown in Figure 4.1, which represents the absence of aberrations. This results in a power distribution that resembles an airy disk.

In the model, the deformable mirror consists of 16 actuators beneath its surface. The primary function of these actuators is to modify the mirror's shape, thereby correcting the wavefront distortions caused by the atmosphere. The deformable mirror permits the transmission of light, which is then captured by the focal plane receiver. When subjected to harsh atmospheric conditions, Figure 4.2 illustrates the atmospheric phase and the power distribution in the focal plane, resulting in a clearer view of the intended target.

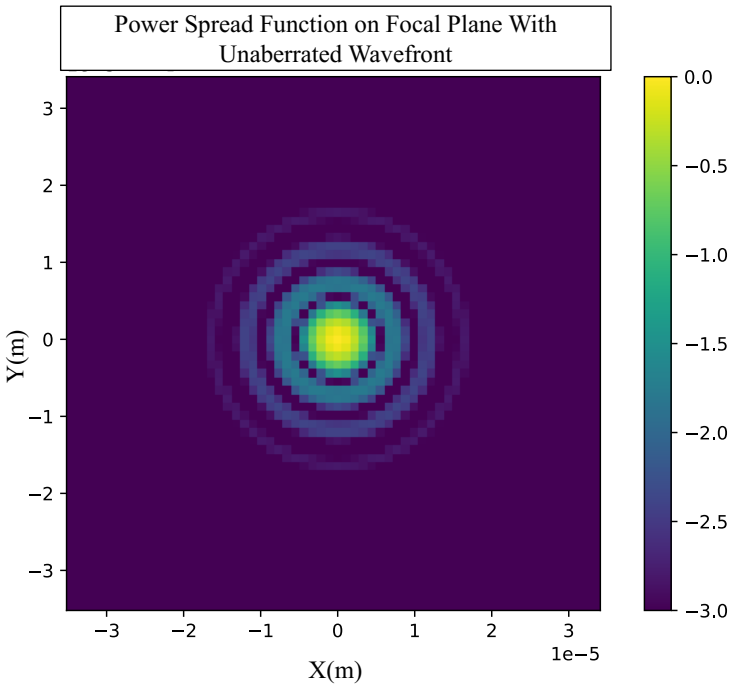


Figure 4.1: The Power Distribution of Unaberrated Wavefront

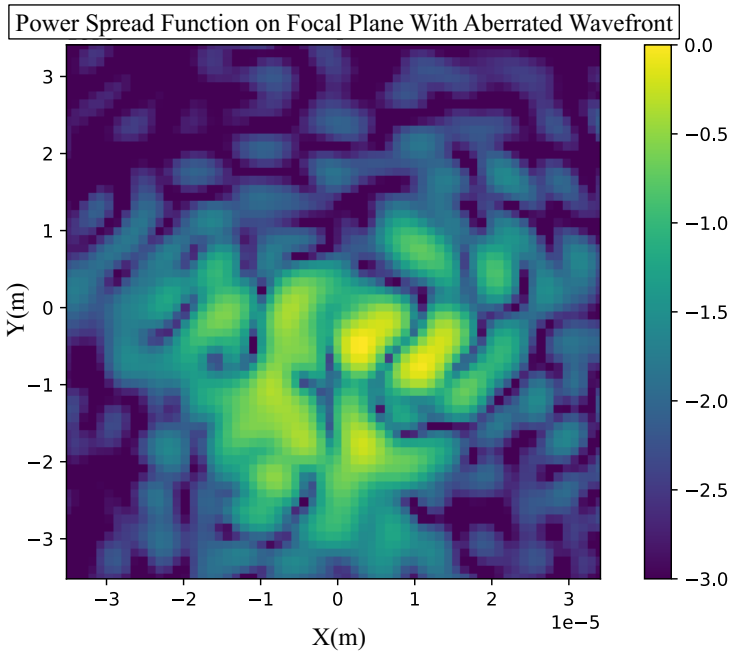


Figure 4.2: The Power Distribution on Focal Plane With Aberrated Wavefront

Hyperparameters	Set 1	Set 2	Set 3	Set 4	Set 5
actor learning rate	1×10^{-5}	2×10^{-5}	3×10^{-5}	4×10^{-5}	5×10^{-5}
critic learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
discount factor	0.95	0.95	0.95	0.95	0.95
layer size	256	256	256	256	256
batch size	128	128	128	128	128

Table 4.1: The Hyperparameters Set for Investigation of Actor Learning Rates

This atmospheric model is crucial in mitigating the effects of atmospheric turbulence on the performance of optical systems. The deformable mirror corrects the distortions in the wavefront to maintain a high-quality image, and the focal plane receiver captures the corrected image. This process is critical in fields such as astronomy, where atmospheric turbulence can significantly impact image quality, and in laser communication, where distortions in the wavefront can lead to signal degradation.

4.3 Hyperparameters Tunning Result

Using the established adaptive optics system model, the tuning of hyperparameters for the proposed soft actor-critic controller takes place within a static atmosphere simulation. The impact of actor learning rate on convergence is illustrated in the results and parameters shown in Figure 4.3 and Table 4.1, respectively. For instance, when the actor learning rate was set to a smaller value, such as 1×10^{-5} and 2×10^{-5}, the curve failed to converge even after 300 epochs. This is because the smaller learning rate slows down the actor network optimizer, which can be beneficial for avoiding suboptimal policies, but it also results in slower convergence rates. Hence, selecting an appropriate actor learning rate that corresponds well with other hyperparameters is crucial to achieving optimal results.

The figure illustrated in Figure 4.4 illustrates the effect of the learning rate of the critic on the adaptive optics system. Additionally, Table 4.2 displays the corresponding hyperparameters. The plot presented in Figure 4.4 demonstrates that policies obtained are suboptimal and inferior to the optimal solution when the learning rate of the critic is not 1×10^{-3}.

The Figure 4.5 and Table 4.3 demonstrate that the variance in discount factor has a notable impact on both the rate and extent of convergence, with the optimal value being 0.95. In reinforcement learning, the discount factor is a crucial

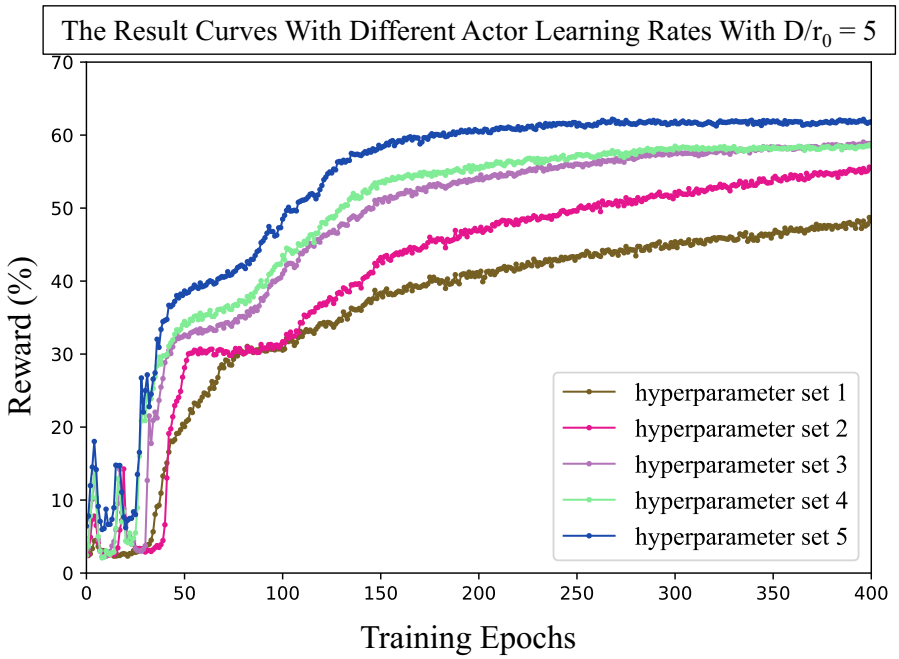


Figure 4.3: The Result of Soft Actor-Critic Controller With Different Actor Learning Rates at $D/r_0 = 5$

Hyperparameters	**Set 6**	**Set 7**	**Set 8**	**Set 9**	**Set 10**
actor learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
critic learning rate	1×10^{-3}	2×10^{-3}	3×10^{-3}	4×10^{-3}	5×10^{-3}
discount factor	0.95	0.95	0.95	0.95	0.95
layer size	256	256	256	256	256
batch size	128	128	128	128	128

Table 4.2: The Hyperparameters Set for Investigation of Critic Learning Rates

parameter used to balance the relative importance of immediate and future rewards. Specifically, it determines the degree to which an agent values a reward it will receive in the future compared to one it will receive immediately. A value of 0 implies that the agent solely cares about immediate rewards, whereas a value of 1 indicates that all future rewards are valued equally. By employing a discount factor, the agent is able to balance the trade-off between immediate and future rewards. If the discount factor is high, the agent will prioritize long-term rewards over short-term ones. Conversely, if the discount factor is low, the agent will prioritize immediate rewards over long-term ones. The choice of a discount factor of 0.95 results in the actor converging to an optimal outcome, whereas other values lead to local optima.

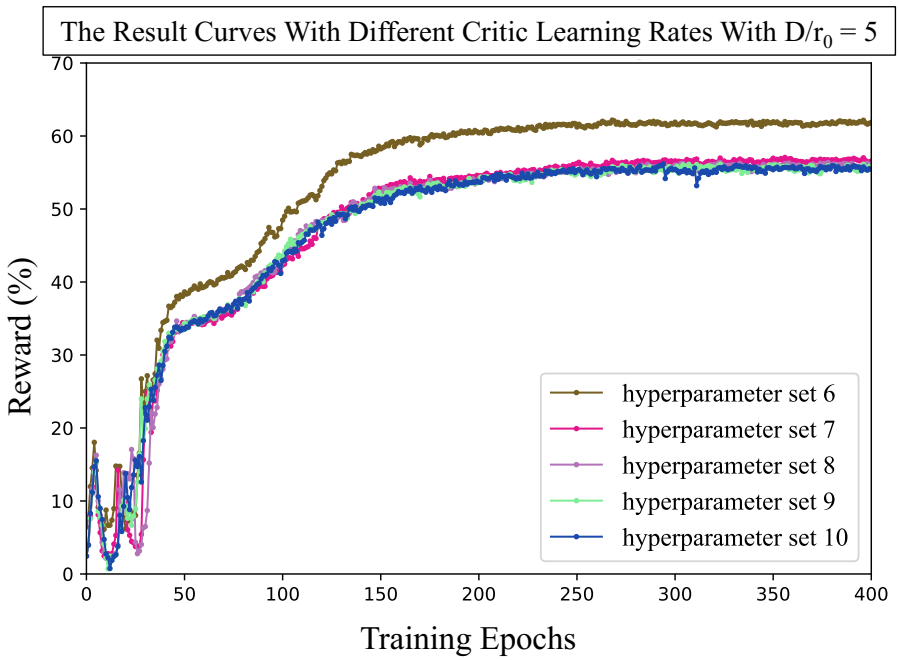


Figure 4.4: The Result of Soft Actor-Critic Controller With Different Critic Learning Rates at $D/r_0 = 5$

Hyperparameters	**Set 11**	**Set 12**	**Set 13**	**Set 14**	**Set 15**
actor learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
critic learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
discount factor	0.95	0.96	0.97	0.98	0.99
layer size	256	256	256	256	256
batch size	128	128	128	128	128

Table 4.3: The Hyperparameters Set for Investigation of Discount Factors

Figure 4.6 displays the study of layer size, while Table 4.4 showcases the hyperparameters. To account for the complexity of the adaptive optics system and the data volume, both the actor and critic have been designed with three hidden layers. For hyperparameter set 16, with a layer size of 32, the result curve exhibits oscillations from the start and fails to settle down until 400 epochs. A smaller layer size leads to an unstable and suboptimal solution. As the layer size increases, the result curve stabilizes, leading to a higher final result, faster convergence speed, and a smoother curve. The increased layer size strengthens the neural network's representational capacity, thus enabling the soft actor-critic controller to model complex and nuanced relationships between the adaptive optics system's observation and the deformable mirror's action.

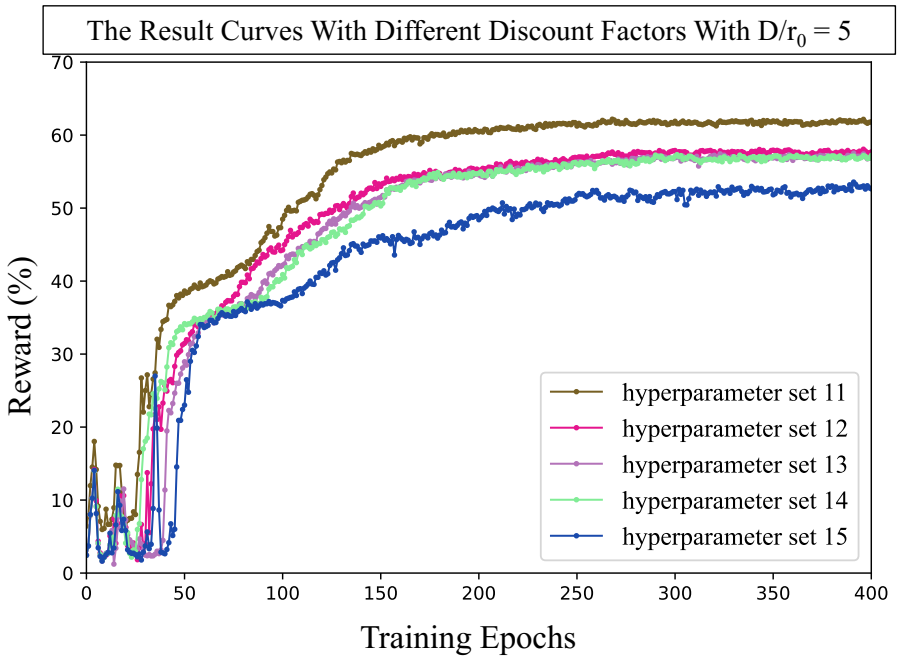


Figure 4.5: The Result of Soft Actor-Critic Controller With Different Discount Factors Rates at $D/r_0 = 5$

Hyperparameters	**Set 16**	**Set 17**	**Set 18**	**Set 19**	**Set 20**
actor learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
critic learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
discount factor	0.95	0.95	0.95	0.95	0.95
layer size	32	64	128	256	512
batch size	128	128	128	128	128

Table 4.4: The Hyperparameters Set for Investigation of Layer Sizes

A study on batch size has been conducted to examine the impact of its value on convergence speed and results. Soft actor-critic is an off-policy reinforcement learning algorithm that requires the controller to sample a batch of transitions from the replay buffer for neural network update and optimization. Batch size refers to the volume of transitions sampled and utilized for training in a single step. Appropriate batch size can reduce variance of the gradient estimate and enhance the efficiency of the learning algorithm. Moreover, larger batch sizes can facilitate more stable updates and contribute to the algorithm's ability to converge to a superior policy. Based on the simulation results, the batch size of 128 yields the best outcome, whereas smaller or larger batch sizes result in unstable training and slow convergence.

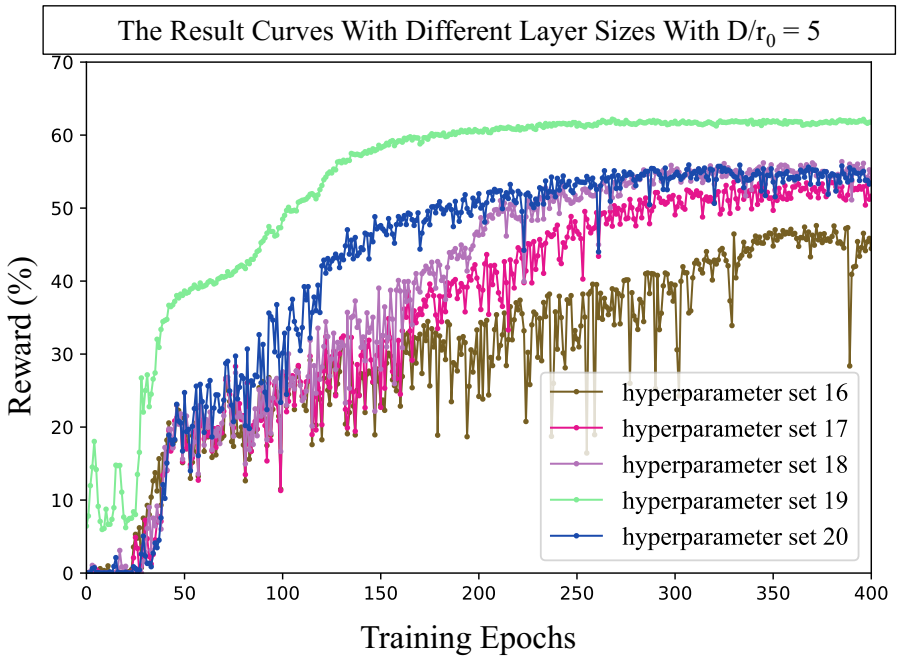


Figure 4.6: The Result of Soft Actor-Critic Controller With Different Layer Sizes Rate at $D/r_0 = 5$

Hyperparameters	**Set 21**	**Set 22**	**Set 23**	**Set 24**	**Set 25**
actor learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
critic learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
discount factor	0.95	0.95	0.95	0.95	0.95
layer size	256	256	256	256	256
batch size	32	64	128	256	512

Table 4.5: The Hyperparameters Set for Investigation of Batch Sizes

The plot in Figure 4.8 illustrates the optimal set of hyperparameters under atmospheric conditions where $D/r_0 = 5$. Through a Bayesian hyperparameter search, Table 4.6 displays the optimal set of hyperparameters. Notably, Figure 4.8 indicates a reduced initial vibration of the curve compared to Figure 4.11 (d), with the resulting curve converging to an optimal Strehl ratio of 62% within 150 steps. This outcome represents a notable improvement in both speed and performance.

4.4 Static Atmosphere Simulation

The severity of atmospheric turbulence varies and increases with D/r_0 values ranging from 2 to 5. In the case of an atmosphere with $D/r_0 = 2$, the level of

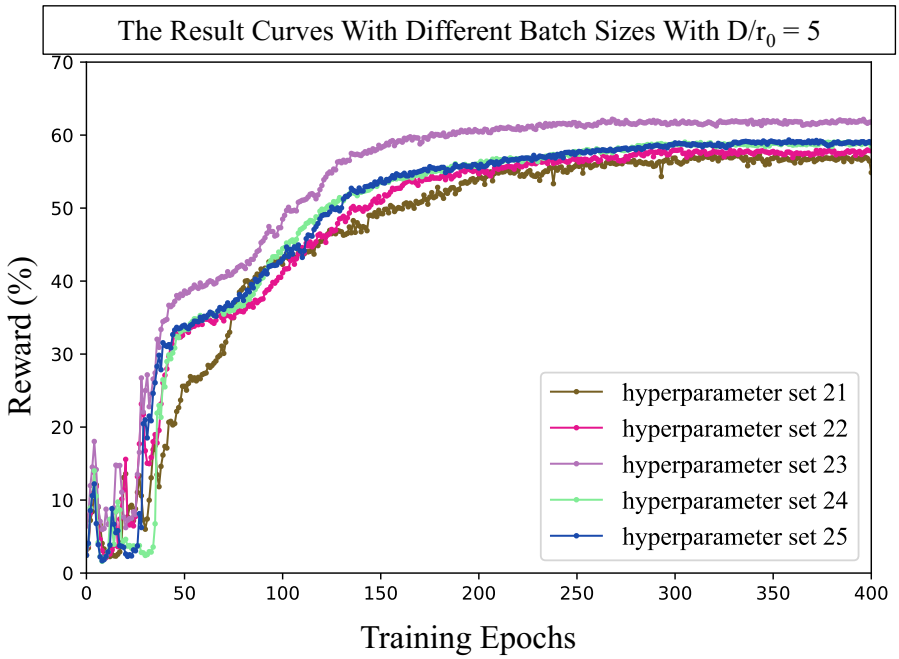


Figure 4.7: The Result of Soft Actor-Critic Controller With Different Batch Sizes Rate at $D/r_0 = 5$

Hyperparameters	Optimal hyperparameter value
actor learning rate	5×10^{-5}
critic learning rate	1×10^{-3}
discount factor	0.95
layer size	256
batch size	128

Table 4.6: The Optimal Hyperparameters Set

turbulence in the atmosphere layers is relatively low. This condition facilitates the operation of the adaptive optics controller. On the other hand, when the atmosphere has a $D/r_0 = 5$, a greater amount of aberration is imposed on the wavefront as the signal propagates through the sky. To test the effectiveness of the proposed controller, a simulation is conducted under the $D/r_0 = 2$ condition. This simulation is conducted in a clear atmosphere with a clear sky and low turbulence. Figure 4.9 (a) shows the initial condition of the atmosphere, which indicates an even and flat distribution with a even distribution of atmosphere phase.

Figure 4.9 depicts the various atmospheric phases at a wavelength of 2.2×10^{-6} for D/r_0 values ranging from 2 to 5. The term "atmospheric phase" refers to the distortion that light undergoes as it passes through the Earth's atmosphere. Such

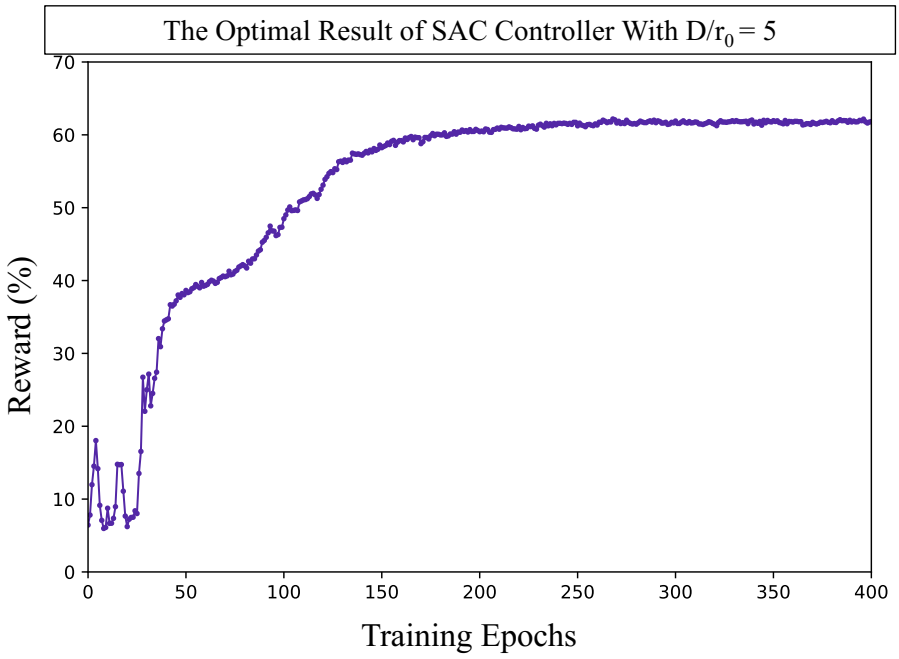


Figure 4.8: The Optimal Result Curve of Soft Actor-Critic Controller With $D/r_0 = 5$

distortion is attributed to the changes in the refractive index of air that arise from variations in temperature, pressure, and humidity. As illustrated in Figure 4.9, the atmospheric phase varies across the entire field of view. The right-hand side color bar of the figure indicates the correlation between color and phase amplitude. Notably, Figure 4.9 highlights that an increase in D/r_0 values corresponds to a surge in the amplitude of the phase. This rise in amplitude indicates a higher level of turbulence and greater distortion of the wavefront.

Figure 4.10 shows the power distribution at the focal plane for a static atmosphere with D/r_0 values of 2, 3, 4, and 5 before the implementation of any wavefront correction. The plots reveal that the distorted wavefront that reaches the focal plane is scattered, causing the peak power of the wavefront to be insufficiently concentrated to form a clear spot. As the D/r_0 value increases, the peak power diminishes and the power distribution becomes increasingly fragmented. The red circles in the figure demarcate the boundaries of the fiber, which is used to capture the signal wavefront. The initial power distribution of the wavefront makes it challenging for the fiber to capture all of the signal power.

The effectiveness of the soft actor-critic controller is assessed across four distinct atmospheric conditions to evaluate its performance relative to a benchmark controller based on a wavefront sensor-based approach. The evaluation is based

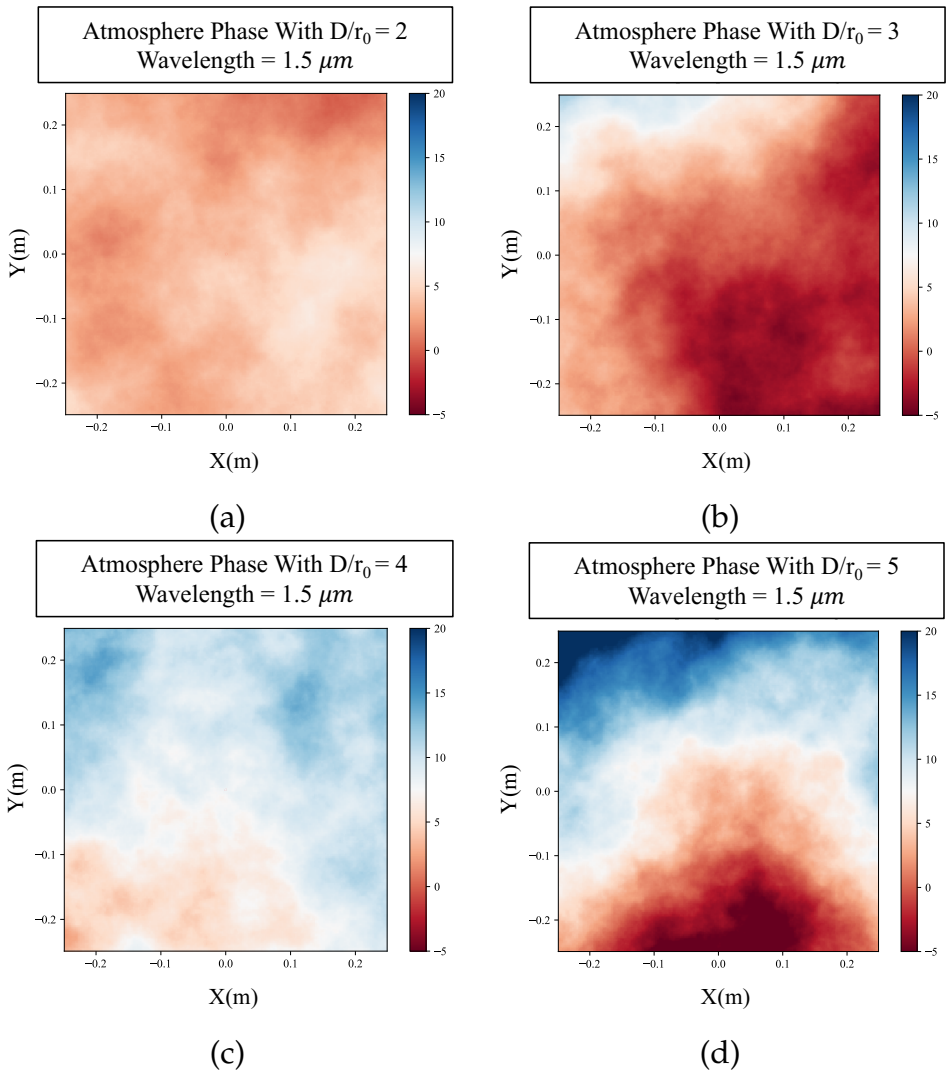


Figure 4.9: The Atmosphere Phase for Wavelength of 1.5×10^{-6}m With (a) $D/r_0 = 2$, (b) $D/r_0 = 3$, (c) $D/r_0 = 4$ and (d) $D/r_0 = 5$

on the utilization of hyperparameters outlined in Table 4.6. The training is conducted with a preset static atmosphere, and each training episode lasts for 100 steps. The average rewards of each episode are recorded and presented in Figure 4.11. The training process is conducted for 400 epochs. In our simulation, each epoch contains 200 training of all neural networks.

Figure 4.11 illustrates the Strehl ratio achieved by two different controllers under varying atmospheric conditions. The soft actor-critic controller obtained a Strehl ratio of 72%, 66%, 66%, and 57% for D/r_0 values of 2, 3, 4, and 5, respectively. Notably, when $D/r_0 = 5$, the reward curve exhibited fluctuations before the initial 75 epochs due to the heightened challenge in searching for the optimal policy, aris-

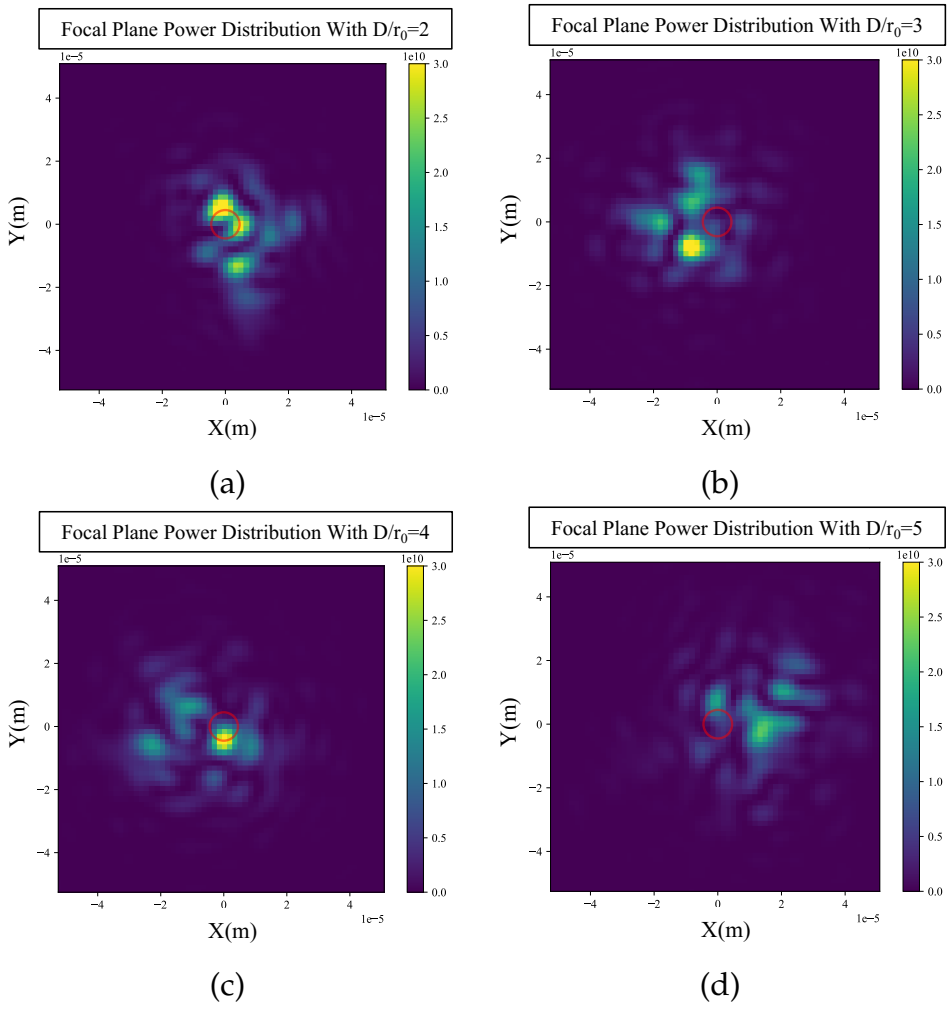


Figure 4.10: The Power Distribution on Focal Plane Before Training With (a) $D/r_0 = 2$, (b) $D/r_0 = 3$, (c) $D/r_0 = 4$ and (d) $D/r_0 = 5$

ing from increased atmospheric turbulence. Additionally, the soft actor-critic controller was capable of converging within 200 epochs of training, indicating its ability to find the optimal solution efficiently. Meanwhile, the wavefront sensor-based controller achieved a higher Strehl ratio of 95%, 87%, 80%, and 74% under the same conditions. Notably, the soft actor-critic controller performed impressively, achieving over 75% of the wavefront sensor-based controller's performance. However, it doesn't surpass the performance of the wavefront sensor-based method due to the limited information on wavefront aberrations acquired from the focal plane, resulting in less accurate manipulation of the surface of the deformable mirror. Nonetheless, achieving more than 75 % of the wavefront sensor-based method's performance proves the effectiveness of the soft actor-critic controller in the wavefront sensorless adaptive optics system, striking a balance between cost and per-

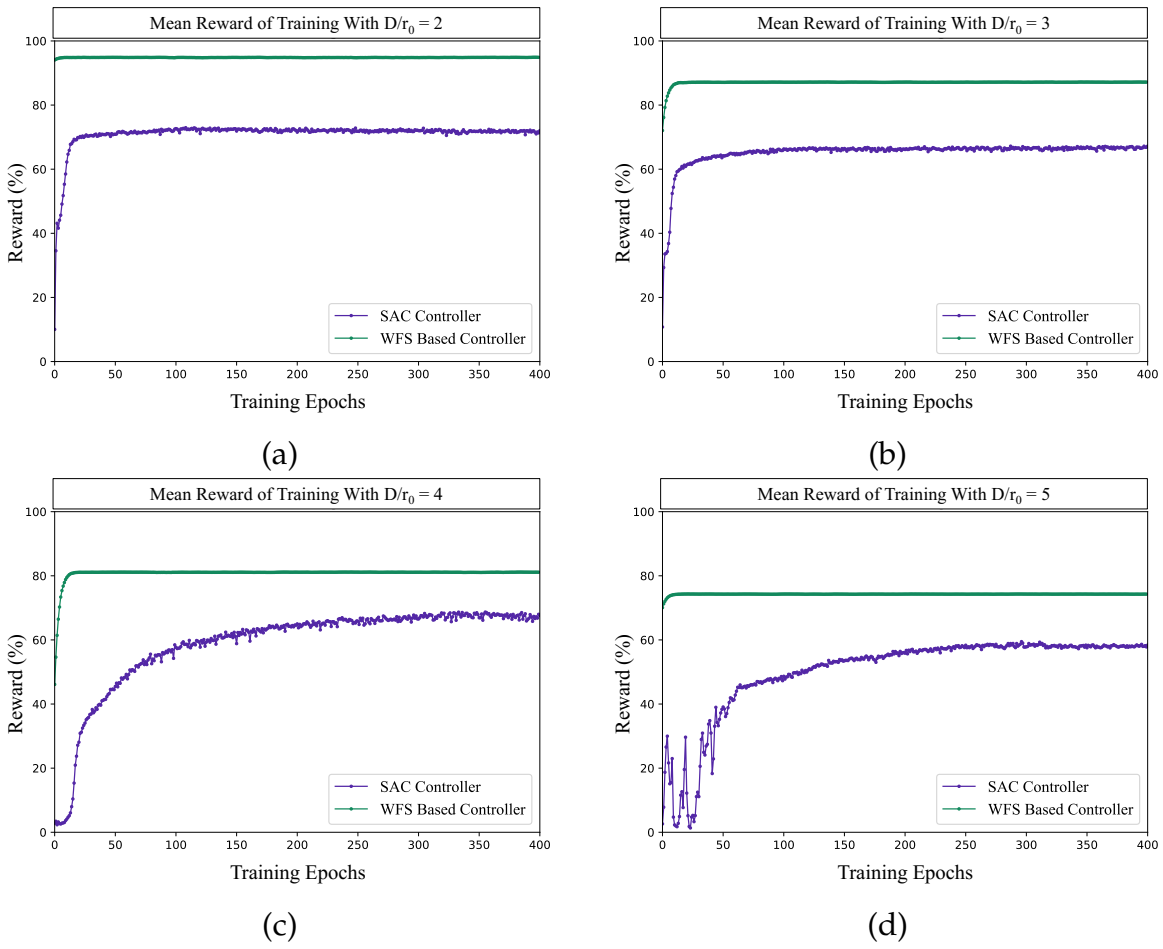


Figure 4.11: The Strehl Ratio of Soft Actor-Critic and Wavefront Sensor-Based Controller With (a) $D/r_0 = 2$, (b) $D/r_0 = 3$, (c) $D/r_0 = 4$ and (d) $D/r_0 = 5$

formance.

Figures 4.12 and 4.13 illustrate the power distribution on the focal plane resulting from two different controllers: the wavefront correction generated by the soft actor-critic controller and the wavefront sensor-based controller, which serves as benchmarks for the study. The power distribution of the focal plane exhibits a concentrated bright spot within the fiber boundaries. Compared to the wavefront sensor-based controllers, the soft actor-critic controllers effectively concentrate the power of the distorted wavefront to the center of the focal plane. This method achieves a performance of more than 75%, as mentioned previously.

To conduct a more comprehensive investigation into the effectiveness of the soft actor-critic controller in terms of converging speed and finding the optimal solution, additional simulations have been carried out. These simulations involve varying values of D/r_0, specifically 2, 3, 4, and 5, and are conducted under differ-

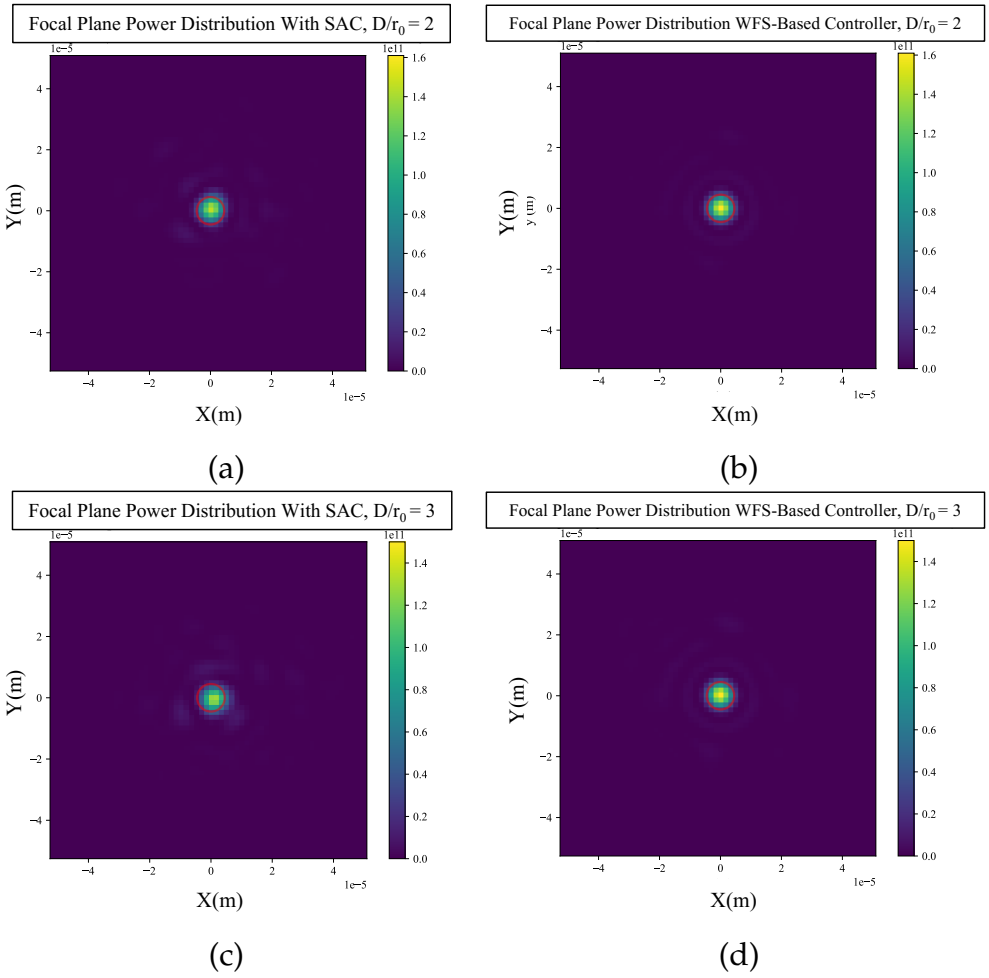


Figure 4.12: The Focal Plane Power Distribution of Soft Actor-Critic Controller Result of (a) $D/r_0 = 2$, Soft Actor-Critic Controller, (b) $D/r_0 = 2$, Shack-Hartmann Sensor-Based Controller, (c) $D/r_0 = 3$, Soft Actor-Critic Controller and (d) $D/r_0 = 3$, Shack-Hartmann Sensor-Based Controller

ent random seeds for the generation of atmosphere layers which will result in a range of atmospheric conditions. To ensure the reliability of the results, a total of 20 groups of random seeds have been employed. The mean and variance values obtained from these simulations are presented graphically in Figure 4.14.

With a computer equipped with an Intel 10700F CPU, Nvidia 3080 GPU, and 32 GB of RAM, Table 4.7 demonstrates that the average epoch requires 10 seconds. As the value of D/r_0 increases, the converging time also increases. However, the modeling and simulation of atmosphere layers require a considerable amount of time, as evidenced by the data in the table. On average, each epoch takes 2 seconds for training and 8 seconds for generating the model of atmosphere. Consequently, when implementing this soft actor-critic controller in real-time with interaction in

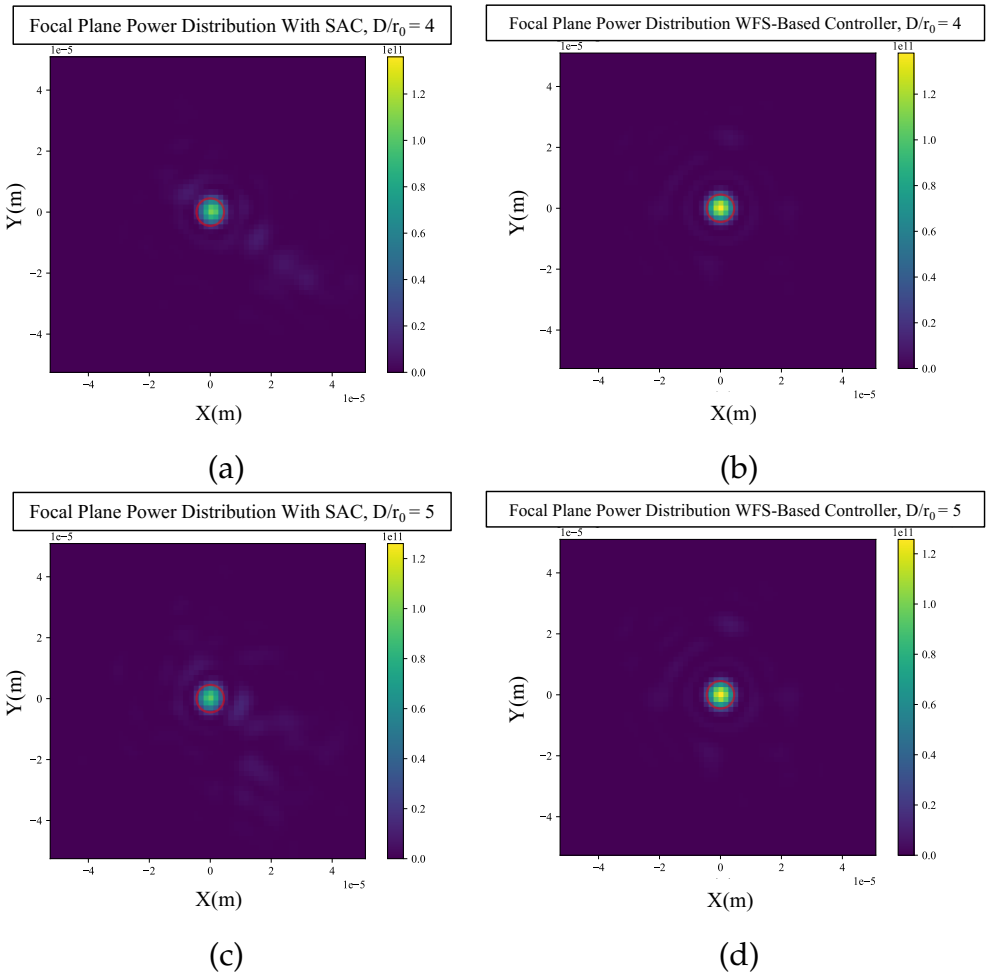


Figure 4.13: The Focal Plane Power Distribution of Soft Actor-Critic Controller Result of (a) $D/r_0 = 4$, Soft Actor-Critic Controller, (b) $D/r_0 = 4$, Shack-Hartmann Sensor-Based Controller, (c) $D/r_0 = 5$, Soft Actor-Critic Controller and (d) $D/r_0 = 5$, Shack-Hartmann Sensor-Based Controller

the real environment, it is worth considering these timing factors.

Once the neural network has been trained, it is crucial to test whether the policy can generate the optimal action when applied to the same environment. The average time required for the acquired policy is presented in Table 4.8. From the table, it is shown that the average processing time for the policy to generate actions of deformable mirror which leads to an optimum Strehl ratio is between 0.015 to 0.02 second based on the current platform.

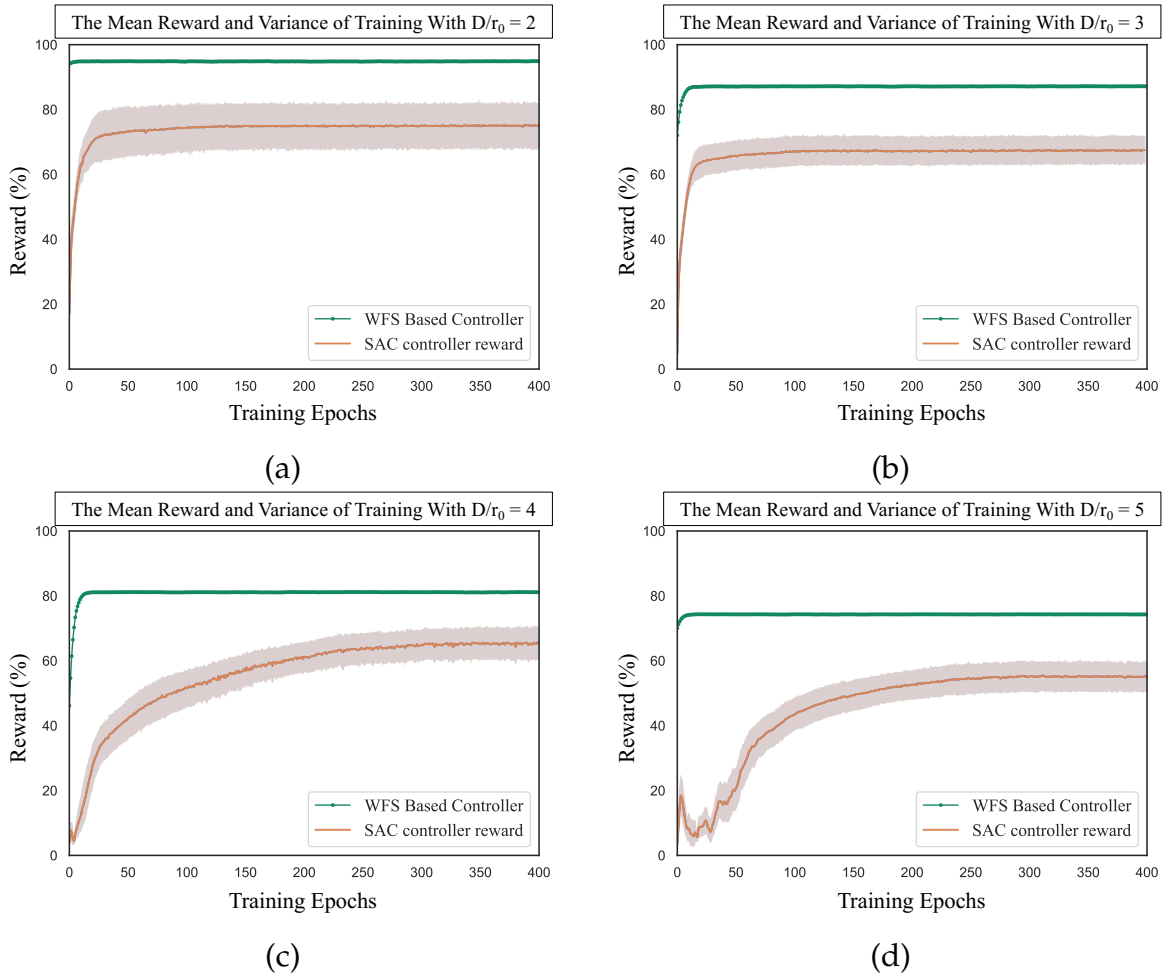


Figure 4.14: The Plot of Soft Actor-Critic Controller With Mean and Variance of (a) $D/r_0 = 2$, (b) $D/r_0 = 3$, (c) $D/r_0 = 4$ and (d) $D/r_0 = 5$

4.5 Semi-Dynamic Atmosphere Simulation

Based on the previous results and discussions from the static atmosphere simulation, the proposed soft actor-critic controller has demonstrated its effectiveness in handling stable atmospheric conditions. Specifically, in severe static atmosphere situations, the proposed controller achieves a Strehl ratio of over 60%, which is more than 75% of the performance achieved by the wavefront sensor-based method.

To further assess the effectiveness of the controller and facilitate its real-time implementation, we performed simulations under semi-dynamic atmospheric conditions, wherein the layers were in motion by the effection of wind while the atmospheric Fried parameter remained constant. These simulations were initiated with an optimal atmospheric state, characterized by a D/r_0 ratio of 3. The input was set

D/r_0	Epochs	Time (seconds)	Modeling Time	Training Time
2	30	299	239	60
3	34	345	273	72
4	221	2218	1768	450
5	250	2490	2010	480

Table 4.7: Epochs and Time for Soft Actor-Critic to Reach Optimum in Training

D/r_0	Average Time deployed (seconds)
2	0.018
3	0.017
4	0.020
5	0.015

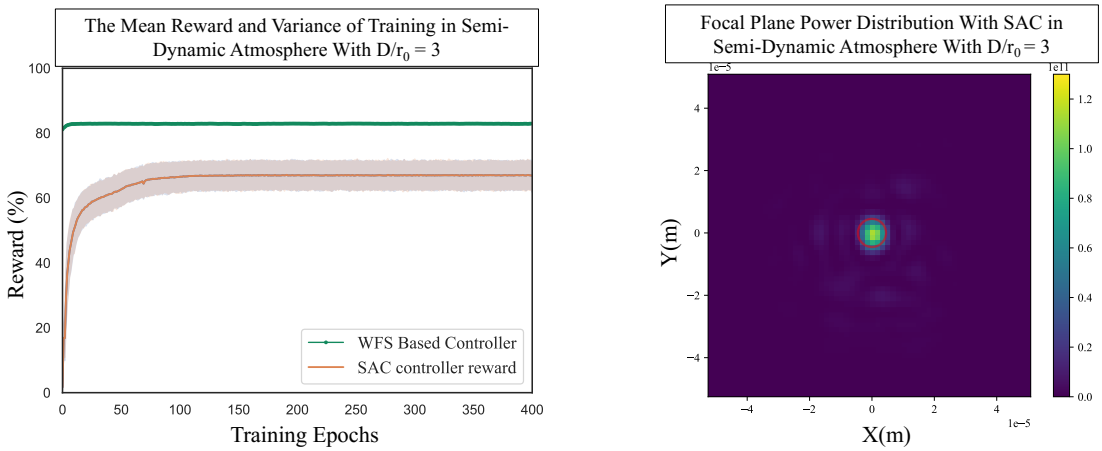
Table 4.8: Average Time for Soft Actor-Critic Controller to Generate Optimal Result in Tests

as observations on the focal plane, with dimensions of 2×2, and the action was applied to a deformable mirror measuring 4×4. The applied atmosphere moved across the telescope's field of view at a velocity of 10 meters per second during the simulation. To ensure the stability of the proposed soft actor-critic controller in a moving atmosphere, we conducted 30 simulations of the moving atmosphere with D/r_0 ratios of 3, 4, and 5, each with different random seeds. We also simulated the Shack Hartmann sensor-based controller as a benchmark for comparison with the soft actor-critic controller.

Figure 4.15 (a) shows that the proposed controller attains an average Strehl ratio of 66% with a low variance in a moving atmosphere. This suggests that, in a relatively stable and favorable moving atmosphere, the soft actor-critic controller is capable of achieving a relatively high Strehl ratio. In comparison with the Shack Hartmann sensor-based controller, the soft actor-critic controller outperforms it by achieving more than 80% performance. Regarding the power on the focal plane, Figure 4.15 (b) demonstrates that the power of the wavefront is concentrated in the middle of the focal plane and inside the boundary of the fiber. Thus, it is confirmed that the proposed controller effectively concentrates the aberrated wavefront power into the fiber.

In addition to the semi-dynamic atmosphere with a D/r_0 value of 3, this study also examines the performance of the soft actor-critic controller under more severe atmospheric conditions, specifically for D/r_0 values of 4 and 5, as illustrated

in Figure 4.16 and Figure 4.17. These conditions are characterized by increased wavefront distortion and greater challenges in searching for the optimal policy. Despite these challenges, the soft actor-critic controller achieves an average Strehl ratio of 45%, and its performance is compared to that of a Shack Hartmann sensor-based controller, which serves as the benchmark. The proposed controller achieves 60% of the benchmark's performance, which corresponds to a Strehl ratio of 74%. In the most severe atmospheric conditions, the soft actor-critic controller can still achieve a Strehl ratio of up to 40%, with a mean Strehl ratio of 32%. By analyzing the power distribution on the focal plane, it is evident that the power of the wavefront can still be focused on the center of the fiber and the focal plane, further indicating the effectiveness of the soft actor-critic controller in dealing with severe atmospheric conditions.

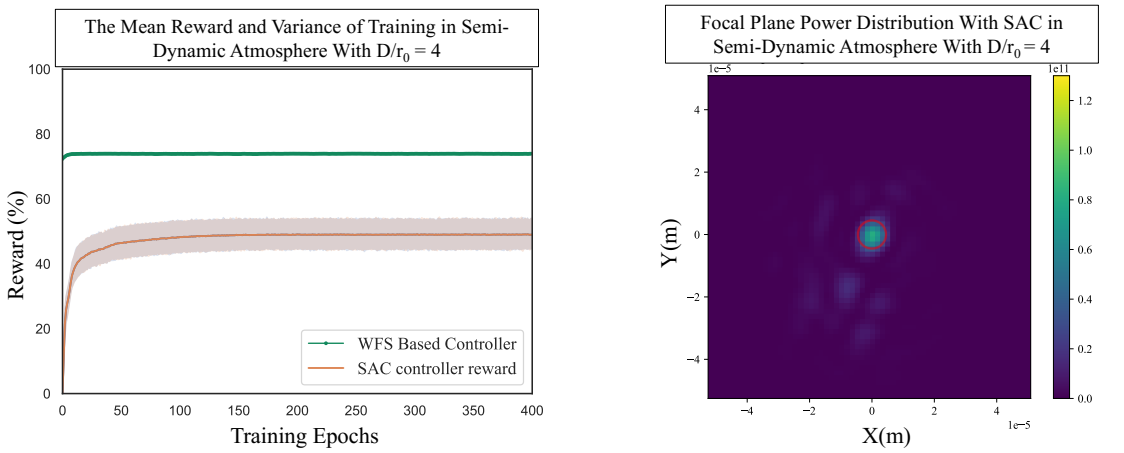


(a) (b)

Figure 4.15: The Result Curves and Final Focal Power Distribution of Soft Actor-Critic Controller With $D/r_0 = 3$ (a) Result Curves With Mean and Variance , (b) Final Focal Plane Power Distribution

4.6 Summary

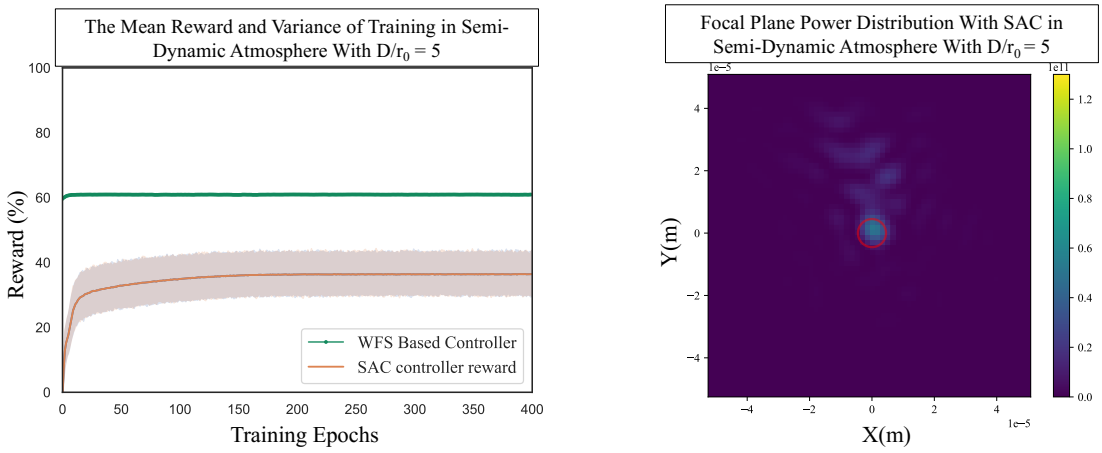
In this chapter, simulations are conducted in a static atmosphere to determine the optimal set of hyperparameters for performance improvement. A detailed exploration of the impact of each hyperparameter set is presented. The effectiveness of this controller is evaluated through simulations in both static and semi-dynamic atmospheres. The proposed controller is able to achieve high Strehl ratios, which demonstrate its efficacy in a static atmosphere. To further assess the effectiveness



(a) (b)

Figure 4.16: The Result Curves and Final Focal Power Distribution of Soft Actor-Critic Controller With $D/r_0 = 4$ (a) Result Curves With Mean and Variance , (b) Final Focal Plane Power Distribution

of the soft actor-critic controller, semi-dynamic atmospheres are introduced into the simulation. Despite high turbulence from the semi-dynamic atmosphere, the soft actor-critic controller can still concentrate the power of the disturbed wavefront into the center of the focal plane and the fiber, indicating that the proposed controller is functional within preliminary semi-dynamic atmospheres. Therefore, the proposed soft actor-critic controller has been verified to correct distorted wavefronts in a cost and time-efficient manner, even with limited observation and measurement devices. These findings suggest that the soft actor-critic controller can be a valuable tool in the field of adaptive optics.



(a) (b)

Figure 4.17: The Result Curves and Final Focal Power Distribution of Soft Actor-Critic Controller With $D/r_0 = 5$ (a) Result Curves With Mean and Variance , (b) Final Focal Plane Power Distribution

Chapter 5

Conclusions and Future Work

Wireless communication has significantly contributed to the convenience and prosperity of modern society. As a vital component of information technology, the communication between near-ground satellites and the Earth's surface plays a critical role in delivering high-quality signals.

The purpose of this study is to develop an adaptive optics controller for a cost-effective wavefront sensorless adaptive optics system. To achieve this, an online off-policy reinforcement learning controller is designed and tested in both static and semi-dynamic atmospheres. The optimal policy is obtained through training based on the interaction with the adaptive optics systems. The soft actor-critic controller only generates control signals based on the limited information provided by a photodiode, eliminating the need for precise and expensive equipment. Furthermore, this method does not require any prior knowledge of the adaptive optics system, as it updates itself through the interaction feedback from the environment during the iterations. The proposed controller demonstrates good performance when tested in different atmospheric conditions in both static and semi-dynamic simulations. Therefore, this model-free online off-policy controller is a viable solution for budget-limited free-space optical communication systems.

However, there remain several issues that require further consideration in the implementation of the soft actor-critic controller. Firstly, it is worth considering the variance of Fried parameters when conducting dynamic atmosphere simulations. This requires gathering more data for training purposes and employing more complex reinforcement learning models for control. Secondly, it is crucial to take into account various factors that may affect the quality of the wavefront signal and cause distortion when the satellite traverses the sky in a real-time environment. These factors may include temperature, humidity, and the zenith angle. Thirdly,

the computation capability of the controller may present an obstacle for real-time applications, particularly when considering the application of artificial intelligence code. Additionally, it is worth to conduct experiments in a physical system for further verification of the proposed controller. There may be other industrial issues that arise when implementing the soft actor-critic controller in a real-time project. Further exploration and analysis of these issues are necessary for successful implementation.

Bibliography

[1] Jalo Nousiainen, Chang Rajani, Markus Kasper, and Tapio Helin. "Adaptive Optics Control Using Model-Based Reinforcement Learning". In: *Optics Express* 29.10 (2021), pp. 15327–15344.

[2] Huimin Ma, Haiqiu Liu, Yan Qiao, Xiaohong Li, and Wu Zhang. "Numerical Study of Adaptive Optics Compensation Based on Convolutional Neural Networks". In: *Optics Communications* 433 (2019), pp. 283–289.

[3] Karen M. Hampson, Raphaël Turcotte, Donald T. Miller, Kazuhiro Kurokawa, Jared R. Males, Na Ji, and Martin J. Booth. "Adaptive Optics for High-Resolution Imaging". In: *Nature Reviews Methods Primers* 1.1 (2021), pp. 1–26.

[4] Hu Ke, Bing Xu, Zhenxing Xu, Lianghua Wen, Ping Yang, Shuai Wang, and Lizhi Dong. "Self-Learning Control for Wavefront Sensorless Adaptive Optics System through Deep Reinforcement Learning". In: *Optik* 178 (2019), pp. 785–793.

[5] Gert Finger, Ian Baker, Reinhold Dorn, Siegfried Eschbaumer, Derek Ives, Leander Mehrgan, Manfred Meyer, and Jörg Stegmeier. "Development of high-speed, low-noise NIR HgCdTe avalanche photodiode arrays for adaptive optics and interferometry". In: *High Energy, Optical, and Infrared Detectors for Astronomy IV*. Vol. 7742. SPIE. 2010, pp. 471–484.

[6] Xiling Shen, Joseph M. Kahn, and Mark A. Horowitz. "Compensation for Multimode Fiber Dispersion by Adaptive Optics". In: *Optics letters* 30.22 (2005), pp. 2985–2987.

[7] Thomas Weyrauch, Mikhail A. Vorontsov, Jay Gowens, and Thomas G. Bifano. "Fiber Coupling with Adaptive Optics for Free-Space Optical Communication". In: *Free-Space Laser Communication and Laser Imaging*. Vol. 4489. SPIE, 2002, pp. 177–184.

[8] Rico Landman, Sebastiaan Y. Haffert, Vikram Mark Radhakrishnan, and Christoph U. Keller. "Self-Optimizing Adaptive Optics Control with Reinforcement Learning for High-Contrast Imaging". In: *Journal of Astronomical Telescopes, Instruments, and Systems* 7.3 (2021), p. 039002.

[9] Rico Landman, Sebastiaan Y. Haffert, Vikram M. Radhakrishnan, and Christoph U. Keller. "Self-Optimizing Adaptive Optics Control with Reinforcement Learning". In: *Adaptive Optics Systems VII*. Vol. 11448. SPIE, 2020, pp. 842–856.

[10] Thomas M Moerland, Joost Broekens, Aske Plaat, and Catholijn M Jonker. "Model-based reinforcement learning: A survey". In: *Foundations and Trends® in Machine Learning* 16.1 (2023), pp. 1–118.

[11] Emiel H Por, Sebastiaan Y Haffert, Vikram M Radhakrishnan, David S Doelman, Maaike van Kooten, and Steven P Bos. "High Contrast Imaging for Python (HCIPy): an open-source adaptive optics and coronagraph simulator". In: *Adaptive Optics Systems VI*. Vol. 10703. SPIE. 2018, pp. 1112–1125.

[12] Na Ji. "Adaptive Optical Fluorescence Microscopy". In: *Nature methods* 14.4 (2017), pp. 374–380.

[13] Yukun Wang, Huanyu Xu, Dayu Li, Rui Wang, Chengbin Jin, Xianghui Yin, Shijie Gao, Quanquan Mu, Li Xuan, and Zhaoliang Cao. "Performance Analysis of an Adaptive Optics System for Free-Space Optics Communication through Atmospheric Turbulence". In: *Scientific reports* 8.1 (2018), pp. 1–11.

[14] François Roddier. "Adaptive Optics in Astronomy". In: (1999).

[15] Jason Porter, Hope Queener, Julianna Lin, Karen Thorn, and Abdul Awwal. "Adaptive Optics for Vision Science". In: (2006).

[16] Yosuke Minowa, Yutaka Hayano, Shin Oya, Makoto Watanabe, Masayuki Hattori, Olivier Guyon, Sebastian Egner, Yoshihiko Saito, Meguro Ito, and Hideki Takami. "Performance of Subaru Adaptive Optics System AO188". In: *Adaptive Optics Systems II*. Vol. 7736. SPIE, 2010, pp. 1302–1308.

[17] F. Merkle, P. Kern, P. Léna, F. Rigaut, J. C. Fontanella, G. Rousset, C. Boyer, J. P. Gaffard, and P. Jagourel. "Successful Tests of Adaptive Optics." In: *The Messenger* 58 (1989), pp. 1–4.

[18] Martin J. Booth. "Wavefront Sensorless Adaptive Optics for Large Aberrations". In: *Optics letters* 32.1 (2007), pp. 5–7.

[19] Huang Linhai and Changhui Rao. "Wavefront Sensorless Adaptive Optics: A General Model-Based Approach". In: *Optics express* 19.1 (2011), pp. 371–379.

[20] Wen Lianghua, Ping Yang, Yang Kangjian, Chen Shanqiu, Wang Shuai, Liu Wenjing, and Bing Xu. "Synchronous Model-Based Approach for Wavefront Sensorless Adaptive Optics System". In: *Optics express* 25.17 (2017), pp. 20584–20597.

[21] Syed Asad Hussain, Toshiki Kubo, Nicholas Hall, Dalia Gala, Karen Hampson, Richard Parton, Mick A Phillips, Matthew Wincott, Katsumasa Fujita, Ilan Davis, Ian Dobbie, and J. Martin Booth. "Wavefront-sensorless adaptive optics with a laser-free spinning disk confocal microscope". In: *Journal of Microscopy* (2020).

[22] Xu He, Xiaohui Zhao, Suying Cui, and Haijun Gu. "A Rapid Hybrid Wave Front Correction Algorithm for Sensor-Less Adaptive Optics in Free Space Optical Communication". In: *Optics Communications* 429 (2018), pp. 127–137.

[23] Eduard Durech, William Newberry, Jonas Franke, and Marinko V. Sarunic. "Wavefront Sensor-Less Adaptive Optics Using Deep Reinforcement Learning". In: *Biomedical optics express* 12.9 (2021), pp. 5423–5438.

[24] John W Hardy. "Adaptive optics". In: *Scientific American* 270.6 (1994), pp. 60–65.

[25] G Rousset, JC Fontanella, Pierre Kern, P Gigan, and Francois Rigaut. "First diffraction-limited astronomical images with adaptive optics". In: *Astronomy and Astrophysics* 230 (1990), pp. L29–L32.

[26] Kiam Heong Ang, Gregory Chong, and Yun Li. "PID control system analysis, design, and technology". In: *IEEE transactions on control systems technology* 13.4 (2005), pp. 559–576.

[27] Martin J Booth. "Adaptive optical microscopy: the ongoing quest for a perfect image". In: *Light: Science & Applications* 3.4 (2014), e165–e165.

[28] Aurélie Facomprez, Emmanuel Beaurepaire, and Delphine Débarre. "Accuracy of correction in modal sensorless adaptive optics". In: *Optics express* 20.3 (2012), pp. 2598–2612.

[29] Na Ji, Daniel E Milkie, and Eric Betzig. "Adaptive optics via pupil segmentation for high-resolution imaging in biological tissues". In: *Nature methods* 7.2 (2010), pp. 141–147.

[30] Daniel E Milkie, Eric Betzig, and Na Ji. "Pupil-segmentation-based adaptive optical microscopy with full-pupil illumination". In: *Optics letters* 36.21 (2011), pp. 4206–4208.

[31] N Pourré, JB Le Bouquin, J Milli, JF Sauvage, T Fusco, C Correia, and S Oberti. "Low-wind-effect impact on Shack-Hartmann-based adaptive optics". In: (2022).

[32] David L Fried. "Optical resolution through a randomly inhomogeneous medium for very long and very short exposures". In: *JOSA* 56.10 (1966), pp. 1372–1379.

[33] John W Hardy. *Adaptive optics for astronomical telescopes*. Vol. 16. Oxford University Press on Demand, 1998.

[34] Andrei Nikolaevich Kolmogorov. "Dissipation of energy in the locally isotropic turbulence". In: *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences* 434.1890 (1991), pp. 15–17.

[35] Andrey Nikolaevich Kolmogorov. "The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers". In: *Cr Acad. Sci. URSS* 30 (1941), pp. 301–305.

[36] Larry N Thibos, Xin Hong, Arthur Bradley, and Xu Cheng. "Statistical variation of aberration structure and image quality in a normal population of healthy eyes". In: *JOSA A* 19.12 (2002), pp. 2329–2348.

[37] Nicholas Devaney, Eugenie Dalimier, Thomas Farrell, Derek Coburn, Ruth Mackey, David Mackey, Francois Laurent, Elizabeth Daly, and Chris Dainty. "Correction of ocular and atmospheric wavefronts: a comparison of the performance of various deformable mirrors". In: *Applied optics* 47.35 (2008), pp. 6550–6562.

[38] Lisa Poyneer, Marcos van Dam, and Jean-Pierre Véran. "Experimental verification of the frozen flow atmospheric turbulence assumption with use of astronomical adaptive optics telemetry". In: *JOSA A* 26.4 (2009), pp. 833–846.

[39] Olivier Guyon and Jared Males. "Adaptive optics predictive control with empirical orthogonal functions (EOFs)". In: *arXiv preprint arXiv:1707.00570* (2017).

[40] Vasudevan Lakshminarayanan and Andre Fleck. "Zernike polynomials: a guide". In: *Journal of Modern Optics* 58.7 (2011), pp. 545–561.

[41] Robert K Tyson and Benjamin West Frazier. *Principles of adaptive optics*. CRC press, 2022.

[42] Byron Engler, Steve Weddell, and Richard Clare. "Wavefront Sensing with Prisms for Astronomical Imaging with Adaptive Optics". In: *2017 International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, 2017, pp. 1–7.

[43] Stéphane R. Chamot, Chris Dainty, and Simone Esposito. "Adaptive Optics for Ophthalmic Applications Using a Pyramid Wavefront Sensor". In: *Optics Express* 14.2 (2006), pp. 518–526.

[44] Ignacio Iglesias. "Pyramid Phase Microscopy". In: *Optics letters* 36.18 (2011), pp. 3636–3638.

[45] Brent L Ellerbroek, Charles Van Loan, Nikos P Pitsianis, and Robert J Plemmons. "Optimizing closed-loop adaptive-optics performance with use of multiple control bandwidths". In: *JOSA A* 11.11 (1994), pp. 2871–2886.

[46] Michael Lloyd-Hart and Patrick McGuire. "Spatio-temporal prediction for adaptive optics wavefront reconstructors". In: *European Southern Observatory Conference and Workshop Proceedings*. Vol. 54. 1996, p. 95.

[47] Robin Swanson, Masen Lamb, Carlos Correia, Suresh Sivanandam, and Kiriakos Kutulakos. "Wavefront reconstruction and prediction with convolutional neural networks". In: *Adaptive Optics Systems VI*. Vol. 10703. SPIE. 2018, pp. 481–490.

[48] Rakesh P Borase, DK Maghade, SY Sondkar, and SN Pawar. "A review of PID control, tuning methods and applications". In: *International Journal of Dynamics and Control* 9.2 (2021), pp. 818–827.

[49] Xin Wu, Yanhe Xu, Jie Liu, Cong Lv, Jianzhong Zhou, and Qing Zhang. "Characteristics analysis and fuzzy fractional-order PID parameter optimization for primary frequency modulation of a pumped storage unit based on a multi-objective gravitational search algorithm". In: *Energies* 13.1 (2019), p. 137.

[50] T Herlambang, D Rahmalia, and T Yulianto. "Particle swarm optimization (pso) and ant colony optimization (aco) for optimizing pid parameters on autonomous underwater vehicle (auv) control system". In: *Journal of Physics: Conference Series*. Vol. 1211. 1. IOP Publishing. 2019, p. 012039.

[51] Fulu Cao. "PID controller optimized by genetic algorithm for direct-drive servo system". In: *Neural Computing and Applications* 32.1 (2020), pp. 23–30.

[52] Zhizheng Wu, Azhar Iqbal, and Foued Ben Amara. "LMI-based multivariable PID controller design and its application to the control of the surface shape of magnetic fluid deformable mirrors". In: *IEEE Transactions on Control Systems Technology* 19.4 (2010), pp. 717–729.

[53] Xizheng Ke, Shangjun Yang, and Jiali Wu. "Hadamard matrix calibration and fuzzy proportional integral differential closed-loop control of adaptive optics system". In: *Optical Engineering* 61.2 (2022), p. 026103.

[54] Mohammed Abouheaf and Wail Gueaieb. "Model-free adaptive control approach using integral reinforcement learning". In: *2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE)*. IEEE. 2019, pp. 1–7.

[55] Ning Wang, Mohammed Abouheaf, Wail Gueaieb, and Nabil Nahas. "Model-free optimized tracking control heuristic". In: *Robotics* 9.3 (2020), p. 49.

[56] Mahmud Iwan Solihin, Lee Fook Tack, and Moey Leap Kean. "Tuning of PID controller using particle swarm optimization (PSO)". In: *Proceeding of the international conference on advanced science, engineering and information technology*. Vol. 1. 2011, pp. 458–461.

[57] Zhi Qi, Qian Shi, and Hui Zhang. "Tuning of digital PID controllers using particle swarm optimization algorithm for a CAN-based DC motor subject to stochastic delays". In: *IEEE Transactions on Industrial Electronics* 67.7 (2019), pp. 5637–5646.

[58] Mojgan Misaghi and Mahdi Yaghoobi. "Improved invasive weed optimization algorithm (IWO) based on chaos theory for optimal design of PID controller". In: *Journal of Computational Design and Engineering* 6.3 (2019), pp. 284–295.

[59] BY Zhao, ZG Zhao, Y Li, RZ Wang, and RA Taylor. "An adaptive PID control method to improve the power tracking performance of solar photovoltaic air-conditioning systems". In: *Renewable and Sustainable Energy Reviews* 113 (2019), p. 109250.

[60] Yu-Tai Liu, Neil Chen, and Steve Gibson. "Adaptive filtering and control for wavefront reconstruction and jitter control in adaptive optics". In: *Proceedings of the 2005, American Control Conference, 2005.* IEEE. 2005, pp. 2608–2612.

[61] JS Gibson, C-C Chang, and BL Ellerbroek. "Adaptive optics: wavefront reconstruction by adaptive filtering and control". In: *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No. 99CH36304)*. Vol. 1. IEEE. 1999, pp. 761–766.

[62] C-C Chang and JS Gibson. "Parallel control loops based on spatial subband processing for adaptive optics". In: *Proceedings of the 2000 American Control Conference. ACC (IEEE Cat. No. 00CH36334)*. Vol. 3. IEEE. 2000, pp. 2113–2117.

[63] Gerardo Escobar, Paolo Mattavelli, Alex M Stankovic, Andrs A Valdez, and Jesus Leyva-Ramos. "An adaptive control for UPS to compensate unbalance and harmonic distortion using a combined capacitor/load current sensing". In: *IEEE Transactions on Industrial Electronics* 54.2 (2007), pp. 839–847.

[64] Jimmie J Perez, Gregory J Toussaint, and Jason D Schmidt. "Adaptive control of woofer-tweeter adaptive optics". In: *Advanced Wavefront Control: Methods, Devices, and Applications VII*. Vol. 7466. SPIE. 2009, pp. 108–119.

[65] C Dessenne, P-Y Madec, D Rabaud, B Fleury, and G Rousset. "First sky tests of adaptive optics predictive control". In: *European Southern Observatory Conference and Workshop Proceedings*. Vol. 56. 1999, p. 165.

[66] Jared R Males and Olivier Guyon. "Ground-based adaptive optics coronagraphic performance under closed-loop predictive control". In: *Journal of Astronomical Telescopes, Instruments, and Systems* 4.1 (2018), p. 019001.

[67] Baptiste Sinquin and Michel Verhaegen. "Tensor-based predictive control for extremely large-scale single conjugate adaptive optics". In: *JOSA A* 35.9 (2018), pp. 1612–1626.

[68] Martin Glück, Jörg-Uwe Pott, and Oliver Sawodny. "Model predictive control of multi-mirror adaptive optics systems". In: *2018 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE. 2018, pp. 909–914.

[69] Ying Chen. "LSTM recurrent neural network prediction algorithm based on Zernike modal coefficients". In: *Optik* 203 (2020), p. 163796.

[70] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. "A review of recurrent neural networks: LSTM cells and network architectures". In: *Neural computation* 31.7 (2019), pp. 1235–1270.

[71] Xuewen Liu, Tim Morris, and Chris Saunter. "Using long short-term memory for wavefront prediction in adaptive optics". In: *International Conference on Artificial Neural Networks*. Springer. 2019, pp. 537–542.

[72] Qi Xin, Guohao Ju, Chunyue Zhang, and Shuyan Xu. "Object-independent image-based wavefront sensing approach using phase diversity images and deep learning". In: *Optics Express* 27.18 (2019), pp. 26102–26119.

[73] Ying Chen. "Voltages prediction algorithm based on LSTM recurrent neural network". In: *Optik* 220 (2020), p. 164869.

[74] Robin Swanson, Masen Lamb, Carlos M Correia, Suresh Sivanandam, and Kiriakos Kutulakos. "Closed loop predictive control of adaptive optics systems with convolutional neural networks". In: *Monthly Notices of the Royal Astronomical Society* 503.2 (2021), pp. 2944–2954.

[75] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[76] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[77] Sander Dieleman, Kyle W Willett, and Joni Dambre. "Rotation-invariant convolutional neural networks for galaxy morphology prediction". In: *Monthly notices of the royal astronomical society* 450.2 (2015), pp. 1441–1459.

[78] E Gendron, F Vidal, M Brangier, T Morris, Z Hubert, A Basden, G Rousset, R Myers, F Chemla, A Longmore, Butterley T, N Dipper, Dunlop C, D Geng, Gratadour D, Henry D, Laporte P, Looker N, Perret D, Sevin A, Talbot G, and Younger E. "MOAO first on-sky demonstration with CANARY". In: *Astronomy & Astrophysics* 529 (2011), p. L2.

[79] Scott W Paine and James R Fienup. "Machine learning for improved image-based wavefront sensing". In: *Optics letters* 43.6 (2018), pp. 1235–1238.

[80] Huimin Ma, Haiqiu Liu, Yan Qiao, Xiaohong Li, and Wu Zhang. "Numerical study of adaptive optics compensation based on convolutional neural networks". In: *Optics Communications* 433 (2019), pp. 283–289.

[81] Marko Noppen. "Focal plane phase retrieval using deep convolutional neural networks: A study on the feasibility of phase retrieval in free space optical communications from a single out of focus intensity measurement using a deep convolutional neural network". In: (2019).

[82] Yangjie Xu, Hongyang Guo, Qiang Wang, Dong He, and Yongmei Huang. "A method of measuring wavefront aberration with CNN". In: *AOPC 2019: AI in Optics and Photonics*. Vol. 11342. SPIE. 2019, pp. 27–32.

[83] Chenda Lu, Qinghua Tian, Lei Zhu, Ran Gao, Haipeng Yao, Feng Tian, Qi Zhang, and Xiangjun Xin. "Mitigating the ambiguity problem in the CNN-based wavefront correction". In: *Optics Letters* 47.13 (2022), pp. 3251–3254.

[84] Carlos E Carrizo, Ramon Mata Calvo, and Aniceto Belmonte. "Intensity-based adaptive optics with sequential optimization for laser communications". In: *Optics express* 26.13 (2018), pp. 16044–16053.

[85] Hongxi Ren and Bing Dong. "Self-calibrated general model-based wavefront sensorless adaptive optics for both point-like and extended objects". In: *Optics Express* 30.6 (2022), pp. 9562–9577.

[86] Qinghua Tian, Chenda Lu, Bo Liu, Lei Zhu, Xiaolong Pan, Qi Zhang, Leijing Yang, Feng Tian, and Xiangjun Xin. "DNN-based aberration correction in a wavefront sensorless adaptive optics system". In: *Optics express* 27.8 (2019), pp. 10765–10776.

[87] Jiaxun Li, Lianghua Wen, Hankui Liu, Guiming Wei, Xiang Cheng, Qing Li, and Bing Ran. "A Novel SPGD Algorithm for Wavefront Sensorless Adaptive Optics System". In: *IEEE Photonics Journal* (2023).

[88] Yan Li, Tairan Peng, Wenlai Li, Hongming Han, and Jianqiang Ma. "Laser beam shaping based on wavefront sensorless adaptive optics with stochastic parallel gradient descent algorithm". In: *14th National Conference on Laser Technology and Optoelectronics (LTO 2019)*. Vol. 11170. SPIE. 2019, pp. 846–851.

[89] Liangliang Han, Yinkang Dai, and Yang Qiu. "Compensation for aberrant wavefront in UOWC based on adaptive optics technique employing genetic algorithm". In: *Optik* 281 (2023), p. 170832.

[90] Jalo Nousiainen, Chang Rajani, Markus Kasper, and Tapio Helin. "Adaptive optics control using model-based reinforcement learning". In: *Optics Express* 29.10 (2021), pp. 15327–15344.

[91] Jalo Nousiainen, C. Rajani, M. Kasper, T. Helin, S. Y. Haffert, C. Vérinaud, J. R. Males, K. Van Gorkom, L. M. Close, J. D. Long, A. D. Hedglen, O. Guyon, L. Schatz, M. Kautz, J. Lumbres, A. Rodack, J. M. Knight, and K. Miller. "Towards on-sky adaptive optics control using reinforcement learning". In: *arXiv preprint arXiv:2205.07554* (2022).

[92] Jalo Nousiainen, Byron Engler, Markus Kasper, Tapio Helin, Cédric T Heritier, and Chang Rajani. "Advances in model-based reinforcement learning for adaptive optics control". In: *Adaptive Optics Systems VIII*. Vol. 12185. SPIE. 2022, pp. 882–891.

[93] Tomi Krokberg. "Reinforcement learning in multi-mirror adaptive optics". In: (2022).

[94] Rico Landman, Sebastiaan Y Haffert, Vikram Mark Radhakrishnan, and Christoph U Keller. "Self-optimizing adaptive optics control with reinforcement learning for high-contrast imaging". In: *Journal of Astronomical Telescopes, Instruments, and Systems* 7.3 (2021), p. 039002.

[95] B Pou, Florian Ferreira, Eduardo Quinones, Damien Gratadour, and Mario Martin. "Adaptive optics control with multi-agent model-free reinforcement learning". In: *Optics express* 30.2 (2022), pp. 2991–3015.

[96] Hu Ke, Bing Xu, Zhenxing Xu, Lianghua Wen, Ping Yang, Shuai Wang, and Lizhi Dong. "Self-learning control for wavefront sensorless adaptive optics system through deep reinforcement learning". In: *Optik* 178 (2019), pp. 785–793.

[97] K Hu, ZX Xu, W Yang, and B Xu. "Build the structure of wfsless ao system through deep reinforcement learning". In: *IEEE Photonics Technology Letters* 30.23 (2018), pp. 2033–2036.

[98] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587 (2016), pp. 484–489.

[99] Benjamin P Moster, Thorsten Naab, Magnus Lindström, and Joseph A O'Leary. "GalaxyNet: connecting galaxies and dark matter haloes with deep neural networks and reinforcement learning in large volumes". In: *Monthly Notices of the Royal Astronomical Society* 507.2 (2021), pp. 2115–2136.

[100] Runnan Zou, Likang Fan, Yanrui Dong, Siyu Zheng, and Chenxing Hu. "DQL energy management: An online-updated algorithm and its application in fix-line hybrid electric vehicle". In: *Energy* 225 (2021), p. 120174.

[101] Runnan Zou, Yuan Zou, Yanrui Dong, and Likang Fan. "A self-adaptive energy management strategy for plug-in hybrid electric vehicle based on deep Q learning". In: *Journal of Physics: Conference Series*. Vol. 1576. 1. IOP Publishing. 2020, p. 012037.

[102] Adekunle A Adepegba, Suruz Miah, and Davide Spinello. "Multi-agent area coverage control using reinforcement learning". In: *The Twenty-Ninth International Flairs Conference*. 2016.

[103] Mohammed Abouheaf, Wail Gueaieb, Davide Spinello, and Salah Al-Sharhan. "A Data-Driven Model-Reference Adaptive Control Approach Based on Reinforcement Learning". In: *2021 IEEE International Symposium on Robotic and Sensors Environments (ROSE)*. IEEE. 2021, pp. 1–7.

[104] Shuzheng Qu, Mohammed Abouheaf, Wail Gueaieb, and Davide Spinello. "An adaptive fuzzy reinforcement learning cooperative approach for the autonomous control of flock systems". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 8927–8933.

[105] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. "Playing atari with deep reinforcement learning". In: *arXiv preprint arXiv:1312.5602* (2013).

[106] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[107] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. "Continuous control with deep reinforcement learning". In: *arXiv preprint arXiv:1509.02971* (2015).

[108] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. "Asynchronous methods for deep reinforcement learning". In: *International conference on machine learning*. PMLR. 2016, pp. 1928–1937.

[109] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor". In: *International conference on machine learning*. PMLR. 2018, pp. 1861–1870.

[110] Petros Christodoulou. "Soft actor-critic for discrete action settings". In: *arXiv preprint arXiv:1910.07207* (2019).

[111] Imre Csiszár. "I-divergence geometry of probability distributions and minimization problems". In: *The annals of probability* (1975), pp. 146–158.

[112] Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. "Improving the robustness of deep neural networks via stability training". In: *Proceedings of the ieee conference on computer vision and pattern recognition*. 2016, pp. 4480–4488.

[113] Virendra N Mahajan. "Strehl ratio for primary aberrations in terms of their aberration variance". In: *JOSA* 73.6 (1983), pp. 860–861.