

# Reinforcement Learning-based Wavefront Sensorless Adaptive Optics Approaches for Satellite-to-Ground Laser Communication

**Payam Parvizi<sup>1</sup>, Runnan Zou<sup>1</sup>, Colin Bellinger<sup>2,3</sup>, Ross Cheriton<sup>2</sup> and Davide Spinello<sup>1</sup>**

<sup>1</sup>Department of Mechanical Engineering, University of Ottawa, Ottawa, Ontario, Canada

<sup>2</sup>National Research Council of Canada, Ottawa, Ontario, Canada

<sup>3</sup>Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

{pparv056, rzou043}@uottawa.ca, {colin.bellinger, ross.cheriton}@nrc-cnrc.gc.ca, dspinell@uottawa.ca

## Abstract

Optical satellite-to-ground communication (OSGC) has the potential to improve access to fast and affordable Internet in remote regions. Atmospheric turbulence, however, distorts the optical beam, eroding the data rate potential when coupling into single-mode fibers. Traditional adaptive optics (AO) systems use a wavefront sensor to improve fiber coupling. This leads to higher system size, cost and complexity, consumes a fraction of the incident beam and introduces latency, making OSGC for internet service impractical. We propose the use of reinforcement learning (RL) to reduce the latency, size and cost of the system by up to 30 – 40% by learning a control policy through interactions with a low-cost quadrant photodiode rather than a wavefront phase profiling camera. We develop and share an AO RL environment that provides a standardized platform to develop and evaluate RL based on the Strehl ratio, which is correlated to fiber-coupling performance. Our empirical analysis finds that Proximal Policy Optimization (PPO) outperforms Soft-Actor-Critic and Deep Deterministic Policy Gradient. PPO converges to within 86% of the maximum reward obtained by an idealized Shack-Hartmann sensor after training of 250 episodes, indicating the potential of RL to enable efficient wavefront sensorless OSGC.

## 1 Introduction

The internet has become an essential tool for education and commerce, yet it remains inaccessible or costly in many remote regions of the globe. Radio frequency satellite constellations are an existing solution that provide internet service; however, they experience a bottleneck in bandwidth due to the carrier wavelength of the link. Optical satellite-to-ground communication operates at near-infrared, thus, enabling a much faster data transfer rate. However, optical beams in optical communication can become distorted as they propagate through atmospheric turbulence, as illustrated in Figure 1, reducing the channel’s potential bandwidth [Kaushal and Kaddoum, 2017; Ma *et al.*, 2015]

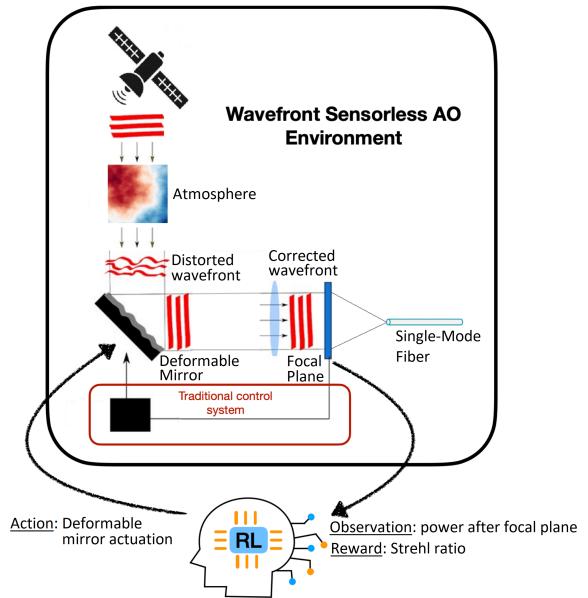


Figure 1: Schematic of the RL environment of wavefront sensorless AO system

Currently, most optical satellite communication ground stations have optical beam coupling into a single-mode fiber (SMF). This requires stronger optical beams from the satellite and a large telescope receiver, leading to high costs. Moreover, larger telescopes are more affected by atmospheric turbulence. This causes a diminishing return on a signal. The issue of atmospheric turbulence has been successfully mitigated in astronomy using adaptive optics (AO) methods [Wenhan, 2018; Roddier, 1999; Tyson and Frazier, 2022]. The traditional AO method dynamically corrects distorted wavefronts in a feedback loop by using measurements from a wavefront sensor and applying them to adjust the distribution of actuators on the deformable mirror (DM), see Figure 2.

Traditional AO systems are still costly and complex, with a significant portion of the cost arising from the wavefront sensor, especially when infrared beams are used for optical satellite-to-ground links. In addition, wavefront sensors in the infrared suffer from high read noise and require cooling and consume a fraction of the incident beam intensity. They

have a limited dynamic range and introduce latency between measurements and the actuation of the DM. This can result in outdated wavefront measurements as the satellite rapidly moves across the sky. This introduces significant errors at the characteristic space-time scales. Recently, research has demonstrated the potential of using reinforcement learning (RL) to solve complex control problems in other domains of AO, such as astronomy [Ren *et al.*, 2021; Tian *et al.*, 2019].

In this work, we propose to reduce cost and latency and improve accuracy using RL-based wavefront sensorless AO for optical satellite-to-ground links. This paper reports on the first phase of a 3-phase program that includes: *a)* Researching and developing RL algorithms for wavefront sensorless AO in a simulated atmosphere, *b)* Characterizing the RL algorithms through physical simulation, and *c)* Deploying the RL model in a real-world AO system. Under the proposed setup, the RL agent learns to control the DM directly from the Strehl ratio of the power after the focal plane. The proposed RL environment is illustrated in Figure 1.

In our empirical analysis, we compare Soft-Actor-Critic (SAC) [Haarnoja *et al.*, 2018a], Proximal Policy Optimization (PPO) [Schulman *et al.*, 2017], and Deep Deterministic Policy Gradient (DDPG) [Lillicrap *et al.*, 2015], to an idealized traditional AO system with a Shack-Hartmann wavefront sensor on the developed RL environment. Our results suggest that RL can significantly improve coupling light into an SMF without a wavefront sensor under static environmental conditions.

To summarize, the contributions of this work are:

- The development of simulated wavefront sensorless AO RL environment for training and testing RL algorithms<sup>1</sup>
- The first demonstration of the potential for RL in wavefront sensorless AO satellite-to-ground data links

The remainder of the paper is structured as follows: Section 2 discusses the relevance of this work with the United Nations Sustainable Development Goals. Section 3 includes background information on AO and related work on RL in the context of AO, and Section 4 details the RL environment developed as part of this work. The experimental setup and RL algorithms are described in Section 5, and Section 6 presents the results. Finally, in Section 7 we discuss the limitations of the RL environment and results, and Section 8 includes final remarks.

## 2 Relationship to the Sustainable Development Goals

This research is relevant to SDG 9 (*Building resilient infrastructure, promoting inclusive and sustainable industrialization and fostering innovation*), as it aims to facilitate faster, more reliable, and lower-cost satellite-to-ground communications that can improve access to the internet in remote and rural regions. RL offers a promising solution for addressing this issue through improved light coupling with less latency at a reduced cost. We estimate that for every

<sup>1</sup>[https://github.com/cbellinger27/adaptive\\_optics\\_gym](https://github.com/cbellinger27/adaptive_optics_gym)

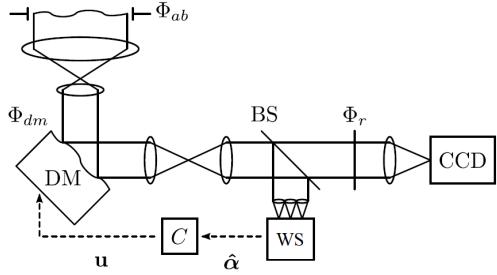


Figure 2: Example of a traditional AO system for a telescope (recreated from [Antonello *et al.*, 2014]).

1% increase in light coupling, the cost of the system is reduced by 2% [van Belle *et al.*, 2004; AstroSysteme, 2022; PlaneWave, 2022]. In traditional AO, improving light coupling in a 40 cm telescope rather than using a 60 cm telescope results in savings of at least 50000 USD for the mount and telescope system [PlaneWave, 2022], at least 30000 USD in the wavefront sensor, and at least 20000 USD in the dome enclosure. In contrast to traditional AO, adding an RL model to the system would only require a compact processor, such as a field-programmable gate array (FPGA), and a simpler detection unit, which would cost significantly less.

## 3 Background

### 3.1 Adaptive Optics

The objective of AO is to render and eliminate any phase distortions in the incoming light. The AO method was first proposed by [Babcock, 1953] to improve astronomical images by correcting for atmospheric distortions with a deformable optical element controlled by a wavefront sensor. Since then, new techniques and results have been consistently published, primarily focusing on advancements in wavefront sensors and DMs [Bifano, 2011; Corbett *et al.*, 2007; Nicolle *et al.*, 2004].

After propagating through the atmosphere to the aperture of a telescope, light is distorted by subtle changes in the temperature and pressure (and hence the index of refraction) of the air, which varies a function of time and space. The process of correcting the distortions with a traditional AO system is shown in Figure 2. At the entrance to the telescope, a phase distortion  $\Phi_{ab}$  is present. Within the telescope, the light is projected onto a DM with  $N_a$  actuators that create another phase aberration  $\Phi_{dm}$ . The phase aberration after the DM becomes  $\Phi_r = \Phi_{ab} + \Phi_{dm}$ . A beam splitter (BS) splits the light into two paths: one towards the exit pupil and the other towards the wavefront sensor (WS). On the path through the exit pupil, the image is created by focusing the light onto an exit pupil (detector) with the phase aberration of  $\Phi_r$ . The other path leads to the WS. This estimates the phase aberration  $\Phi_r$  in the form of a set of Zernike coefficients  $N_\alpha$  collected into a vector  $\hat{\alpha} \in \mathbb{R}^{N_\alpha}$ . Finally, a controller (C) receives the coefficients and computes a vector  $u \in \mathbb{R}^{N_a}$ , and creates a control signal of  $N_a$  actuators of the DM. The main purpose of the controller is to minimize the phase aberration  $\Phi_r$ .

The objective of this work is to reduce cost and increase power through the use of RL. We hypothesize that RL can learn a control policy for the DM, thereby eliminating the need for the BS and WS, and reducing cost and increasing power.

### 3.2 Satellite-To-Ground Communication

The optical beam containing the communication signal can come from any one of a constellation of low earth orbit (LEO) satellites equipped with laser transmitters. Line-of-sight is a key required transmission between the telescope and a satellite. This is only maintained for a few minutes for a particular satellite, at which point the telescope must point to another satellite.

The atmosphere has a characteristic turbulence timescale on the order of  $\sim 1$  ms, which varies with the satellite elevation angle. At the lowest elevations, the turbulence is the strongest, owing to the effective thickness of the atmosphere. The wavefront is also rapidly changing based on the change of optical path, which results in the turbulence profile appearing to translate across the aperture as the telescope is tracking the satellite. This means that the model must be able to determine the wavefront within 1 ms to maintain a high degree of correction as the wavefront changes while the satellite passes, and the adjustment of the DM must be made within tens of seconds. If an approximate solution can be found within a few milliseconds, the atmosphere can be considered to be in a quasi-static state and a static turbulence profile for the purposes of training can be considered valid.

### 3.3 Reinforcement Learning (RL)

RL is a machine learning technique that has been successfully applied to various continuous control tasks, including adaptive optics. In a wavefront sensor-based system, [Pou *et al.*, 2022a] proposed a multi-agent RL method for compensating for bandwidth error with a Shack-Hartmann sensor. [Nousiainen *et al.*, 2021; Nousiainen *et al.*, 2022] applied model-based algorithms with a wavefront sensor in building an AO system to deal with the time delay error, and misregistration. [Pou *et al.*, 2022b], in another work, combined RL with a nonlinear reconstructor-based on neural networks in a U-net architecture for wavefront correction with a pyramid wavefront sensor. Within wavefront sensorless AO systems, DDPG and convolutional neural networks (CNN) were deployed to shift the performance of correction capacity, and speed for image sharpness [Ke *et al.*, 2019; Hu *et al.*, 2018]. In these applications, CNNs extract features from images captured by a detector on the exit pupil. The DDPG then generates a control signal of DM based on CNN output. DDPG was also applied on a microscope by Durech in a wavefront system in which there is no atmospheric turbulence but lower speed turbulence from aqueous solutions and optical aberrations [Durech *et al.*, 2021].

However, the existing implementations of RL on AO systems are insufficient for optical satellite communication due to their focus on optimizing image sharpness instead of optical data link reliability. Additionally, these implementations are optimized for different wavefront distortion conditions in environments such as microscopy and ophthalmology.

## 4 Wavefront Sensorless Adaptive Optics RL Environment

The RL environment is implemented according to the standards of the Open AI Gym framework [Brockman *et al.*, 2016]. The HCIPy: High Contrast Imaging for Python package [Por *et al.*, 2018] serves as the foundation of the RL environment. HCIPy offers a comprehensive set of libraries related to adaptive optics, including wavefront generation, atmospheric turbulence modeling, propagation simulation, fiber coupling, implementation of DMs and wavefront sensors. A simulated AO RL environment is a critical first step in the process of developing RL-based wavefront sensorless satellite-to-ground communication systems. It enables us and future researchers to assess and refine the RL to meet the strict requirements of this domain prior to costly evaluation in physical simulations and the real-world.

The adaptive optics system simulated in this environment couples 1550 nm light into an SMF under different static turbulence conditions represented by the parameter  $D/r_0$  (shown in Figure 3), which is the ratio of the telescope's diameter  $D$  to Fried's parameter  $r_0$ . This is a measure of the quality of optical transmission through the atmosphere. As Fried's parameter decreases, the transmission of light through the atmosphere becomes increasingly complex, resulting in more pronounced wavefront distortions. The  $D$ -value is fixed in this work to 0.5 m, and the  $r_0$ -value is varied to assess the performance under different levels of atmospheric turbulence. The term "static turbulence condition" refers to the assumption that the turbulence in the atmosphere is stationary and the satellite is at a fixed position.

A graphical presentation of the RL environment is shown in Figure 1. The simulation environment updates in discrete time steps for practical purposes and consistency with the standard RL framework. As illustrated in the figure, at each time step  $t$ , the RL agent receives an observation of the system's current state ( $o$ ) and a reward ( $r$ ). The observation encodes the power after the focal plane in the AO system, and the reward is computed as the Strehl ratio of the power after the focal plane. Based on the agent's policy and the current observation, that agent selects the next action  $\pi : o \rightarrow a$ . The agent's actions control the DM in the AO system. If controlled optimally, the incoming optical beam becomes concentrated and centered on the SMF.

### 4.1 Episodic Environment

Under real-world conditions, this RL problem can be characterized as a finite or infinite horizon problem. In the former,

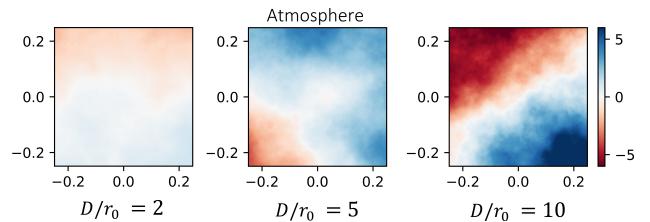


Figure 3: Atmospheric turbulence conditions with respect to  $D/r_0$

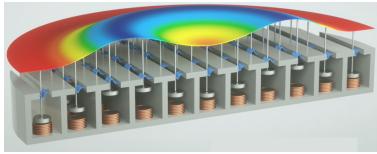


Figure 4: Example of ALPAO DM surface and actuators. The image is presented by the ALPAO company [ALPAO, 2021]

an episode lasts for the duration of the satellite’s communication with the receiver. Here, we assess how efficiently an RL policy can transform the DM from its neutral position (flat) to a formation that focuses the beam on the SMF. Therefore, we define short (30, 50, and 100 time steps), fixed-length episodes starting from a neutral mirror.

## 4.2 Action-Space

In the context of AO, the RL agent’s actions are movements of the actuators beneath the DM. The DM and actuators are illustrated in Figure 4. The number of actuators determines the degree of freedom of the mirror’s shape. The actuators are responsible for controlling the continuous reflective surface of the DM. The range of movement for these actuators is on the order of  $\pm 1 \mu\text{m}$ , providing a high degree of precision in the mirror’s surface shape and position.

The RL environment simulates a 64-actuator segmented DM, which has roughly 8 actuators across a linear dimension. Thus, the RL agent selects actions from a 64-dimensional continuous action-space where each actuator can be moved independently. This allows for smooth and precise control of the DM. The actuation has a limit corresponding to the maximum optical phase error that is possible under the atmospheric conditions used in training. The DM speed is assumed to be sufficiently fast to consider the atmosphere as quasi-static since most DMs are capable of correction speeds of up to a few kHz [Pengwang *et al.*, 2016]. Faster DMs are possible using smaller mirrors.

For a 50 cm telescope and this DM choice, the system can be expected to perform very well corrected for turbulence conditions of less than  $r_0 = 6.25$ . We expect  $r_0 = 0$  conditions to range from 5 cm to 15 cm for satellite elevation angles above  $15^\circ$ .

## 4.3 Observation-Space

We utilize the power of the wavefront propagated through the focal plane to form the observation of the state of the environment. The focal plane is shown on the left in Figure 5. We deal with observations rather than Markovian states because these can be directly and efficiently related to the light coupled into the fiber through a Strehl ratio calculation. Full state information would require access to information about the angle of the satellite and atmospheric conditions.

For the observations, we discretize the focal plane into a sub-aperture array of  $2 \times 2$  pixels that can be realized with a fast and relatively low-cost quadrant photodetector, as shown on the right in Figure 5. Using a low-pixel detector mitigates the use of slower and expensive read-out circuits used in infrared cameras, allows for more light per pixel for less noise, and improves the speed of the RL algorithm training.

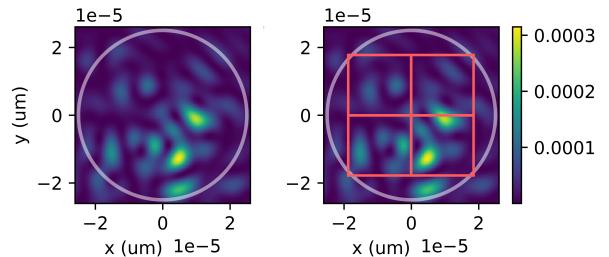


Figure 5: Focal plane profile (left) continuous, (right) discretized

## 4.4 Reward Function

The reward function is calculated as the Strehl ratio of the optical system. It is defined as the ratio of the normalized peak intensity of the wavefront’s point spread function (PSF) to the peak intensity of the ideal PSF without aberrations. A high Strehl ratio implies a high degree of wavefront correction, where a focused beam of light resembles an Airy disk and is approximately proportional to the amount of light that can be coupled into a fiber [Jovanovic, N. *et al.*, 2017]. It is considered an approximation since a focused beam should resemble a Gaussian profile for optimal coupling into an optical fiber. As proposed by Mahajan [Mahajan, 1983], the Strehl ratio of systems with a circular pupil is considered in terms of the variance of the phase aberration across the pupil. The expression is

$$\text{Strehl} = e^{-\sigma_\Phi^2} \quad (1)$$

where  $\sigma_\Phi^2$  is the variance of the phase aberration. This quantity is chosen as the reward function since it is directly correlated to the amount of light coupled into an SMF [Jovanovic, N. *et al.*, 2017].

## 5 Experimental Methodology

We quantify the RL performance as the mean and standard deviation of the Strehl ratio over 20 independent trials. Each RL algorithm has its hyperparameters tuned to the environment and the best-performing setup is compared to an idealized AO system with a Shack-Hartmann wavefront sensor.

We implement and compare three families of RL algorithms: Soft Actor-Critic (SAC), Deep Deterministic Policy Gradient (DDPG), and Proximal Policy Optimization (PPO). These cover on- and off-policy learning, stochastic and deterministic policies and entropy-based methods, which are expected to have strengths and weaknesses in the context of wavefront sensorless adaptive optics. In particular, the off-policy algorithms, SAC and DDPG, generally have better sample efficiency than the on-policy algorithms. Alternatively, the on-policy method, PPO, is often more stable and easier to train. Given sufficient time, on-policy methods can provide good performance. Finally, the entropy regularization in SAC enables rich exploration, which is beneficial in high-dimensional action spaces. Each algorithm is discussed in more detail below. Table 1 presents a comprehensive list of the hyperparameters selected for each method after tuning.

Hyperparameter	SAC	DDPG	PPO
Buffer size	128	256	-
Actor-lr	$5e^{-4}$	$5e^{-5}$	$1e^{-2}$
Critic-lr	$1e^{-2}$	$1e^{-2}$	$5e^{-6}$
Actor-Hidden dim.	150	250	150
Critic-Hidden dim.	80	65	50
Clipping $\epsilon$	-	-	0.35
Temp. $\alpha_t$ -lr	$1e^{-1}$	-	-
Temp. $\alpha_t$ -min limit	0.4	-	-
No episodes per iteration	1	2	2
No updates per iteration	20	20	20
Polyak ( $\rho$ )	0.99	0.99	-
Discount ( $\gamma$ )	0.95	0.95	0.95
Reward scaling	No	No	Mean-std
learned $\alpha_t$	Semi	-	-

Table 1: Hyperparameters and corresponding values

### 5.1 Soft Actor-Critic (SAC)

SAC is an off-policy actor-critic algorithm based on a maximum entropy RL framework. It is particularly useful in complex and stochastic environments, such as wavefront sensorless AO systems. SAC uses a deep neural network to approximate the actor and critics. The actor component of the algorithm is tasked with maximizing the expected return while promoting exploration through random actions rather than becoming trapped in suboptimal policies. On the other hand, the critic component is responsible for estimating the Q-function of a given state-action pair. The Q-function provides feedback for improving the policy by adjusting the actions that the agent takes in each state to maximize the expected sum of future rewards [Haarnoja *et al.*, 2018a].

SAC has shown promising results in various domains. However, one major drawback is its sensitivity to the choice of temperature ( $\alpha_t$ ) and intuitively selected target entropy parameters. These parameters play a crucial role in the algorithm’s performance, and their selection can significantly affect the outcome. To address this, [Haarnoja *et al.*, 2018b] proposed a method of automatic gradient-based temperature tuning by matching the expected entropy  $\log \pi_t^*(\mathbf{a}_t | \mathbf{s}_t; \alpha_t)$  to a target entropy value  $\bar{\mathcal{H}}$  at time  $t$

$$\alpha_t^* = \operatorname{argmin}_{\alpha_t} \mathbb{E}_{\mathbf{a}_t \sim \pi_t^*} [\alpha_t \log \pi_t^*(\mathbf{a}_t | \mathbf{s}_t; \alpha_t) - \alpha_t \bar{\mathcal{H}}], \quad (2)$$

where the temperature  $\alpha$  controls the stochasticity of the optimal policy.

In our preliminary assessment, we evaluate SAC with a fixed temperature, learned temperature, and semi-learned temperature. The fixed temperature condition was optimized by setting a constant value of  $\alpha_t = 0.4$ , resulting in the best reward with the mean value of 52.63% and standard deviation of 14.64 at the end of training. The learned temperature condition was optimized using the learning rate  $lr_{\alpha_t} = 1e^{-1}$ , with learning rates ranging from  $1e^{-6}$  to  $5e^{-1}$ . The semi-learned temperature condition was optimized using the same learning rate and a minimum alpha value of 0.4, resulting in improved performance compared to the total range of learning rates and minimum  $\alpha_t$  values between 0 and 1. The re-

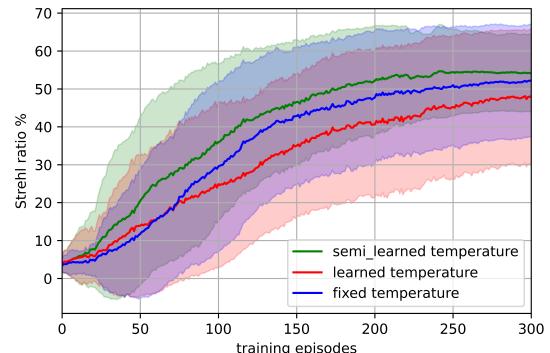


Figure 6: Comparison of the selection of the temperature ( $\alpha_t$ ) in SAC applied on 20 randomly selected static atmospheric turbulence of  $D/r_0 = 5$

sults show that the fixed and semi-learned learn at a faster rate than purely learned, and that semi-learned converges to the best policy. Thus, semi-learned is used in all subsequent experiments.

### 5.2 Deep Deterministic Policy Gradient (DDPG)

DDPG is an off-policy actor-critic algorithm for learning optimal control policies in continuous action spaces. It utilizes a deterministic policy, rather than a stochastic one, to map observations to actions.

DDPG utilizes a deep neural network to approximate the value function and policy, allowing to handle high-dimensional state spaces. This approach, as previously demonstrated by [Lillicrap *et al.*, 2015], can be effective in tackling complex environments. While DDPG has the advantage of ease of implementation, it can be sensitive to the choice of hyperparameters and can be prone to instability due to the choice of the reward function.

In our preliminary assessment, we compared the effect of reward normalization with the mean-standard deviation (mean-std norm) and the min-max norm method versus no normalization on the learned policy. The normalization process scales the rewards across episodes. This has been shown to help the model identify actions that lead to higher rewards, thus accelerating the algorithm’s convergence. In addition, as the model reaches convergence, the variance of the rewards tends to decrease, making it more challenging for the model to adjust itself. By normalizing the rewards, the model can more effectively recognize these rewards and continue to make adjustments.

Figure 7 demonstrates that omitting reward scaling in this domain leads to slow convergence to a lower reward. Minmax norm and mean-std norm learn at similar rates, however, mean-std norm converges to a higher reward. Thus, mean-std normalization is employed for the subsequent experiments in our Deep Deterministic Policy Gradient (DDPG) approach.

### 5.3 Proximal Policy Optimization (PPO)

PPO is an on-policy, policy gradient algorithm. It alternates between sampling data through interactions with the environment and optimizing a surrogate objective function via

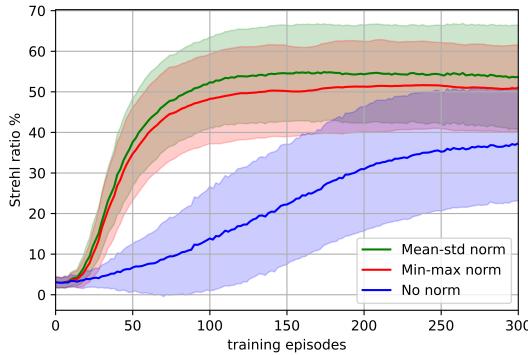


Figure 7: Comparison of the selection of normalization technique in DDPG applied on 20 randomly selected static atmospheric turbulence of  $D/r_0 = 5$ .

stochastic gradient ascent [Schulman *et al.*, 2017]. PPO utilizes a deep neural network to approximate the policy and value function. It uses a “clipped surrogate objective” to penalize significant changes to the policy. The policy regulation technique described in [Schulman *et al.*, 2017] limits the drastic changes of the policy during updates, but its effectiveness depends on the hyperparameter  $\epsilon$ , which controls the size of policy updates to prevent model collapse. Setting  $\epsilon$  too small may result in slow convergence while setting it too large increases the risk of model collapse.

Our preliminary analysis found that PPO is robust to  $\epsilon \in [0.05, 0.4]$  in this environment. Settings within this range show very subtle differences in variance, convergence rate and convergence level. Generally, smaller  $\epsilon$  values resulted in slightly slower convergence, whereas larger values converge faster but to marginally lower levels. Based on this analysis, all subsequent experiments have  $\epsilon = 0.35$ .

## 6 Results

### 6.1 Comparison of RL algorithms

We used the light coupling performance obtained from Shack-Hartmann wavefront sensor data as the reference for comparison with the refined RL algorithms outlined in Section 5. This comparison was conducted under static turbulence condition of  $D/r_0 = 5$ . The results are presented in Figure 8. Using the Shack-Hartmann wavefront sensor with 12 lenslets as benchmark allowed for a thorough evaluation of the effectiveness of the proposed RL algorithms in improving light coupling performance in the presence of atmospheric turbulence.

The results presented in Figure 8 indicate a better performance of the PPO algorithm in comparison to SAC and DDPG algorithms. Specifically, when considering the randomly selected static atmospheric turbulence of  $D/r_0 = 5$ , the PPO algorithm achieved a maximum reward of 67%, which is close to the maximum reward obtained by the Shack-Hartmann sensor, around 73%. While SAC and DDPG still demonstrated acceptable performance with a maximum reward of 53% in the early training episodes, the PPO algorithm consistently demonstrated better results. The improved performance of PPO compared to SAC and DDPG can be

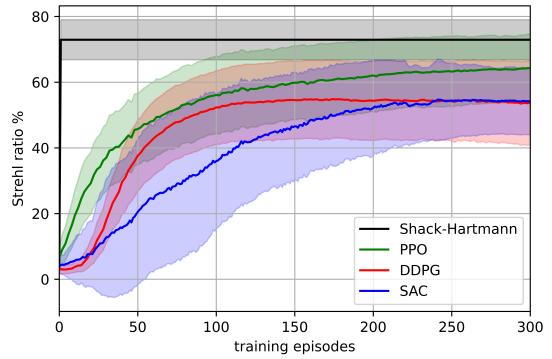


Figure 8: Comparison of models applied on 20 randomly selected static atmospheric turbulence of  $D/r_0 = 5$ .

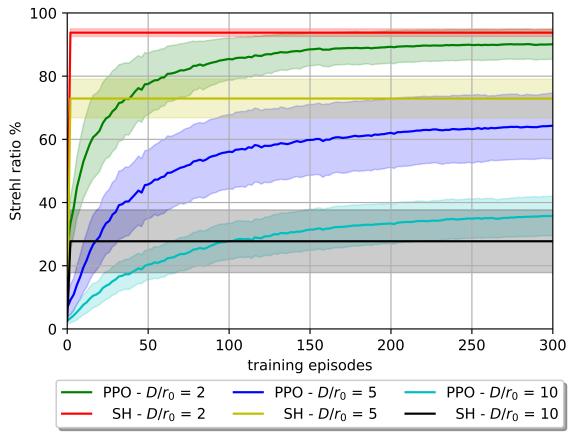


Figure 9: Average reward of 20 randomly selected static atmospheres of different  $D/r_0$  ratios with 64 actuators and 4 observers on PPO algorithm and Shack-Hartmann wavefront sensor

attributed to its ability to sample from the action distribution, which helps avoid suboptimal local minima in high-dimensional state spaces. However, in SAC and DDPG, the actions that look nearly optimal can have an equal likelihood of being tried as those that look highly unoptimal.

### Performance with Turbulence severity

Until this point, all experiments presented have been conducted under the turbulent condition of  $D/r_0 = 5$ . As previously stated, as the value of Fried’s parameter decreases, there is a corresponding increase in the difficulty of transmitting light through the atmosphere. In this section, we used the PPO algorithm to assess its performance under mild and severe static turbulent conditions. The results are presented in Figure 9.

As anticipated and illustrated in Figure 9, a decrease in the value of Fried’s parameter (or an increase in the ratio of  $D/r_0$ ) results in a decline in the model’s ability to attain a higher reward. If the agent performance cannot significantly improve beyond the uncorrected 2 to 10% Strehl ratio (depending on  $D/r_0$ ), it can be considered impractical.

The impact of using a PPO algorithm within an RL environment and using a Shack-Hartmann wavefront sensor on

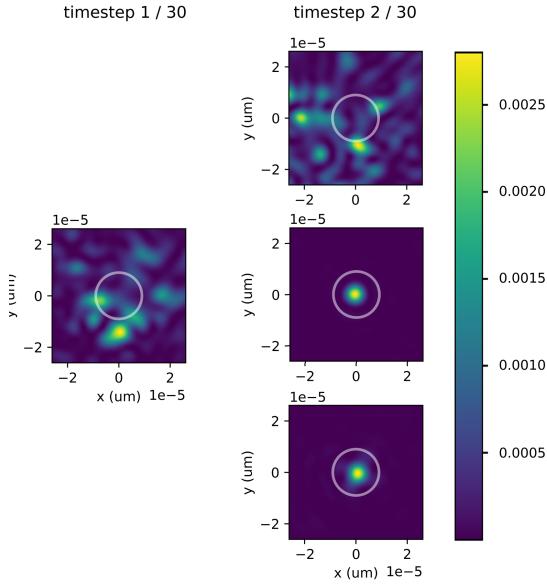


Figure 10: Power distribution on a focal plane (left) before proposed AO, (right-up) after the implementation of random actions through the PPO algorithm in the initial episodes, (right-middle) after the utilization of the Shack-Hartmann wavefront sensor, (right-down) after the application of the PPO algorithm following a sequence of episodes.

the power distribution of a wavefront in a static turbulence condition of  $D/r_0 = 5$  have been analyzed and illustrated in Figure 10. On the left of the figure, the power distribution at the focal plane at the start of each episode when the wavefront is reflected through a flat DM is illustrated. The right of the figure displays the results after (i) the implementation of random actions through the PPO algorithm in the initial episodes (upper right), (ii) the utilization of the Shack-Hartmann wavefront sensor (middle right), and (iii) the application of the PPO algorithm following a sequence of episodes (lower right).

The utilization of the Shack-Hartmann wavefront sensor, as illustrated in Figure 10 (right-middle), has resulted in a significant concentration of power at the center of the focal plane, with a Strehl ratio of approximately 70%. Similarly, the application of the PPO algorithm, as illustrated in Figure 10 (right-down), has also produced a significant concentration of power at the center of the focal plane; however, with a slightly lower Strehl ratio of approximately 60%.

## 7 Discussion

The results indicate that while it is possible for the PPO, SAC, and DDPG RL models to determine an accurate set of DM actions, the results generally underperform the output of the Shack-Hartmann wavefront sensor. PPO outperforms the SAC and DDPG models, converging on a Strehl ratio of over 60% after hundreds of training episodes. One of the main limitations of these results is the applicability to a dynamic atmosphere with limited deformable mirror speeds. The 3 dB bandwidth of the fastest mechanical deformable mirrors is limited to  $<10$  kHz, which implies that less than 10 action-

measurement loop iterations are required for the quasi-static turbulence condition to hold. For most DMs, this requirement cannot be met. However, novel photonic chip-based phase corrector arrays are capable of speeds well in excess of 20 kHz [Diab *et al.*, 2022] and can make use of a model requiring tens of action steps to converge.

Despite this limitation, we expect such simplifications to be reasonable as an RL-based system may still outperform a Shack-Hartmann wavefront sensor with its corresponding and photon count requirements which are not considered in our comparisons. For existing and planned optical ground stations, adaptive optics may not be considered at all due to the cost and unreliability of wavefront sensing. Therefore, any improvement to the wavefront beyond an uncorrected case is still of value. Furthermore, the application of models trained on quasi-static turbulence profiles may be of value to RL environments with dynamically changing turbulence profiles for improvement to signal re-acquisition times.

The choice in using a  $2 \times 2$  quadrant photodetector as a source of feedback arises from its simplicity, cost and a high degree of correlation to improved SMF coupling. Using PPO, the Strehl ratio improves rapidly, but may be insufficient for achieving a reasonable degree of wavefront correction under high turbulence conditions because of the presence of a high-dimensional continuous action space in contrast to a low-dimensional observation space.

## 8 Conclusion

We present a reinforcement learning-based approach for wavefront sensorless Adaptive Optics in the context of optical satellite-to-ground communication. Specifically, we used off-policy algorithms like SAC and DDPG, as well as an on-policy algorithm, PPO, to achieve optimal coupling of 1550 nm light into a single-mode fiber under various static turbulence conditions. The results show that the PPO algorithm is particularly effective in achieving a high average Strehl ratio in a low number of training episodes. Furthermore, our approach eliminates the requirement of wavefront sensor measurements, thereby reducing the cost and latency of optical satellite-to-ground communication, making it a promising solution for fast and affordable internet access in remote and low-resources areas. Future work could include investigating the performance of Reinforcement Learning algorithms in various dynamic turbulence conditions at different times of the day.

## Acknowledgements

This research was supported by the National Science and Engineering Research Council (NSERC) of Canada through Discovery grant RGPIN-2022-03921, and by the National Research Council (NRC) of Canada through the AI4D grant AI4D-135-2.

## References

- [ALPAO, 2021] ALPAO. ALPAO adaptive optics deformable mirrors. <https://www.alpao.com/products-and-services/deformable-mirrors>, 2021.

- [Antonello *et al.*, 2014] Jacopo Antonello, Tim van Werkhoven, Michel Verhaegen, Hoa H. Truong, Christoph U. Keller, and Hans C. Gerritsen. Optimization-based wavefront sensorless adaptive optics for multiphoton microscopy. *J. Opt. Soc. Am. A*, 31(6):1337–1347, Jun 2014.
- [AstroSysteme, 2022] AstroSysteme. AstroSysteme Austria products. <https://www.astrosysteme.com/products>, Feb 2022.
- [Babcock, 1953] Horace W Babcock. The possibility of compensating astronomical seeing. *Publications of the Astronomical Society of the Pacific*, 65(386):229–236, 1953.
- [Bifano, 2011] Thomas Bifano. MEMS deformable mirrors. *Nature photonics*, 5(1):21–23, 2011.
- [Brockman *et al.*, 2016] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.
- [Corbett *et al.*, 2007] Alexander D Corbett, Timothy D Wilkinson, Jiang J Zhong, and Luis Diaz-Santana. Designing a holographic modal wavefront sensor for the detection of static ocular aberrations. *JOSA A*, 24(5):1266–1275, 2007.
- [Diab *et al.*, 2022] Momen Diab, Ross Cheriton, and Suresh Sivanandam. Photonic phase correctors based on grating couplers: proof of concept simulations and preliminary performance metrics. In Laura Schreiber, Dirk Schmidt, and Elise Vernet, editors, *Adaptive Optics Systems VIII*, volume 12185, page 12185Q. International Society for Optics and Photonics, SPIE, 2022.
- [Durech *et al.*, 2021] Eduard Durech, William Newberry, Jonas Franke, and Marinko V Sarunic. Wavefront sensorless adaptive optics using deep reinforcement learning. *Biomedical optics express*, 12(9):5423–5438, 2021.
- [Haarnoja *et al.*, 2018a] 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*, pages 1861–1870. PMLR, 2018.
- [Haarnoja *et al.*, 2018b] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.
- [Hu *et al.*, 2018] K Hu, ZX Xu, W Yang, and B Xu. Build the structure of wfsless ao system through deep reinforcement learning. *IEEE Photonics Technology Letters*, 30(23):2033–2036, 2018.
- [Jovanovic, N. *et al.*, 2017] Jovanovic, N., Schwab, C., Guyon, O., Lozi, J., Cvetojevic, N., Martinache, F., Leon-Saval, S., Norris, B., Gross, S., Doughty, D., Currie, T., and Takato, N. Efficient injection from large telescopes into single-mode fibres: Enabling the era of ultra-precision astronomy. *A&A*, 604:A122, 2017.
- [Kaushal and Kaddoum, 2017] Hemani Kaushal and Georges Kaddoum. Optical communication in space: Challenges and mitigation techniques. *IEEE Communications Surveys & Tutorials*, 19(1):57–96, 2017.
- [Ke *et al.*, 2019] 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. *Optik*, 178:785–793, 2019.
- [Lillicrap *et al.*, 2015] 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. *arXiv preprint arXiv:1509.02971*, 2015.
- [Ma *et al.*, 2015] Jing Ma, Kangning Li, Liying Tan, Siyuan Yu, and Yubin Cao. Performance analysis of satellite-to-ground downlink coherent optical communications with spatial diversity over gamma-gamma atmospheric turbulence. *Appl. Opt.*, 54(25):7575–7585, Sep 2015.
- [Mahajan, 1983] Virendra N Mahajan. Strehl ratio for primary aberrations in terms of their aberration variance. *JOSA*, 73(6):860–861, 1983.
- [Nicolle *et al.*, 2004] M Nicolle, T Fusco, G Rousset, and V Michau. Improvement of Shack-Hartmann wave-front sensor measurement for extreme adaptive optics. *Optics letters*, 29(23):2743–2745, 2004.
- [Nousiainen *et al.*, 2021] Jalo Nousiainen, Chang Rajani, Markus Kasper, and Tapio Helin. Adaptive optics control using model-based reinforcement learning. *Optics Express*, 29(10):15327–15344, 2021.
- [Nousiainen *et al.*, 2022] Jalo Nousiainen, C Rajani, M Kasper, T Helin, SY Haffert, C Vérinaud, JR Males, K Van Gorkom, LM Close, JD Long, et al. Towards on-sky adaptive optics control using reinforcement learning. *arXiv preprint arXiv:2205.07554*, 2022.
- [Pengwang *et al.*, 2016] Eakkachai Pengwang, Kantz Rabenorosoa, Micky Rakotondrabe, and Nicolas Andreff. Scanning micromirror platform based on MEMS technology for medical application. *Micromachines (Basel)*, 7(2):24, February 2016.
- [PlaneWave, 2022] PlaneWave. Planewave instruments observatory systems. <https://planewave.com/observatory-systems>, Feb 2022.
- [Por *et al.*, 2018] 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*, volume 10703, pages 1112–1125. SPIE, 2018.
- [Pou *et al.*, 2022a] B Pou, Florian Ferreira, Eduardo Quinones, Damien Gratadour, and Mario Martin. Adaptive optics control with multi-agent model-free reinforcement learning. *Optics express*, 30(2):2991–3015, 2022.

- [Pou *et al.*, 2022b] B Pou, J Smith, E Quinones, M Martin, and D Gratadour. Model-free reinforcement learning with a non-linear reconstructor for closed-loop adaptive optics control with a pyramid wavefront sensor. In *Adaptive Optics Systems VIII*, volume 12185, pages 945–958. SPIE, 2022.
- [Ren *et al.*, 2021] Hongxi Ren, Bing Dong, and Yan Li. Alignment of the active secondary mirror of a space telescope using model-based wavefront sensorless adaptive optics. *Applied Optics*, 60(8):2228–2234, 2021.
- [Roddier, 1999] François Roddier. *Adaptive Optics in Astronomy*. Cambridge University Press, 1999.
- [Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [Tian *et al.*, 2019] 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. *Optics express*, 27(8):10765–10776, 2019.
- [Tyson and Frazier, 2022] Robert K Tyson and Benjamin West Frazier. *Principles of adaptive optics*. CRC press, 2022.
- [van Belle *et al.*, 2004] Gerard Theodore van Belle, Aden Baker Meinel, and Marjorie Pettit Meinel. The scaling relationship between telescope cost and aperture size for very large telescopes. In Jacobus M. Oschmann Jr., editor, *Ground-based Telescopes*, volume 5489, pages 563 – 570. International Society for Optics and Photonics, SPIE, 2004.
- [Wenhan, 2018] Jiang Wenhan. Overview of adaptive optics development. *Opto-Electronic Engineering*, 45(3):170489–1–170489–15, 2018.