

PROCEEDINGS OF SPIE

SPIDigitalLibrary.org/conference-proceedings-of-spie

Self-optimizing adaptive optics control with reinforcement learning

Landman, R., Haffert, S., Radhakrishnan, V., Keller, C.

R. Landman, S. Y. Haffert, V. M. Radhakrishnan, C. U. Keller, "Self-optimizing adaptive optics control with reinforcement learning," Proc. SPIE 11448, Adaptive Optics Systems VII, 1144849 (13 December 2020); doi: 10.1117/12.2560053

SPIE.

Event: SPIE Astronomical Telescopes + Instrumentation, 2020, Online Only

Self-optimizing adaptive optics control with Reinforcement Learning

R. Landman^{a,*}, S.Y. Haffert^a, V.M. Radhakrishnan^a, and C.U. Keller^a

^aLeiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands

ABSTRACT

Current and future high-contrast imaging instruments require extreme Adaptive Optics (XAO) systems to reach contrasts necessary to directly image exoplanets. Telescope vibrations and the temporal error induced by the latency of the control loop limit the performance of these systems. Optimization of the (predictive) control algorithm is crucial in reducing these effects. We describe how model-free Reinforcement Learning can be used to optimize a Recurrent Neural Network controller for closed-loop adaptive optics control. We verify our proposed approach for tip-tilt control in simulations and a lab setup. The results show that this algorithm can effectively learn to suppress a combination of tip-tilt vibrations. Furthermore, we report decreased residuals for power-law input turbulence compared to an optimal gain integrator. Finally, we demonstrate that the controller can learn to identify the parameters of a varying vibration without requiring online updating of the control law. We conclude that Reinforcement Learning is a promising approach towards data-driven predictive control; future research will apply this approach to the control of high-order deformable mirrors.

Keywords: Adaptive Optics, Predictive Control, High Contrast Imaging, Reinforcement Learning

1. INTRODUCTION

One of the main limitations of current and future ground-based high contrast imaging (HCI) instruments is the wavefront error induced by the time lag between sensing and correcting the wavefront. This time lag results in a halo of speckles along the wind direction, which limits the contrast at small separations.¹ Furthermore, the finite control bandwidth results in low-order residuals after the correction. One of the main causes of these residuals are vibrations close to the maximum control frequency that originate from a variety of sources. All ground-based high-contrast imaging instruments suffer from vibrations including SCExAO,² GPI³ and SPHERE.⁴ These vibrations decrease the resolution of long-exposure images. Furthermore, some coronagraphs are highly sensitive to residual tip-tilt jitter,^{5,6} which will further reduce the contrast.

Both of these effects may be reduced by the use of better control algorithms. Early work on improved controllers involved optimizing the modal gains of a modal integrator based on the open-loop temporal Power Spectral Density (PSD).^{7,8} This optimal gain is a trade-off between turbulence rejection and the amplification of noise and disturbances outside the control bandwidth. However, to reduce the effect of the servo-lag we have to predict the disturbances before they are measured. A popular approach to optimal control is Linear Quadratic Gaussian (LQG) control,⁹ which is based on a Kalman filter. Variants of LQG-based controllers have been demonstrated to improve the performance in tip-tilt vibration mitigation in simulations,¹⁰ lab verifications¹¹ and on-sky.^{3,12,13} A major disadvantage of LQG-like controllers is their reliance on a state-space model of the turbulence and deformable mirror (DM) dynamics;¹⁴ this requires an accurate calibration of the model parameters. Hence, these controllers can also not correct for correlations in the turbulence that are not included in the model. A combination of an LQG controller for low-order vibrations and an optimal modal gain integrator for high order modes is currently being used in the AO systems of both SPHERE¹³ and GPI.¹⁵

Another group of data-driven algorithms construct a linear filter that minimizes the residual phase variance from (pseudo) open-loop data. This idea was originally proposed and demonstrated for modal control in ¹⁶. More recently, a spatio-temporal filter based on Empirical Orthogonal Functions (EOFs) was demonstrated to significantly increase the contrast in a simulated HCI system.¹⁷ Performance of such linear filters was tested

*R.Landman: E-mail: rlandman@strw.leidenuniv.nl

on open-loop AO telemetry for the Keck II AO system¹⁸ and SPHERE¹⁹ and showed a decrease in residual wavefront error. An issue with these data-driven predictive control algorithms is that they rely on the prediction of (pseudo) open-loop wavefront data, while XAO systems operate in closed-loop. To use these controllers, one has to reconstruct the pseudo open-loop residuals. Not only does this require knowledge of the servo-lag of the system and DM dynamics, it is also not trivial in the case where the response of the wavefront sensor is nonlinear, such as for the Pyramid Wavefront sensor. Furthermore, a relatively small error on the reconstruction or prediction of the pseudo open-loop wavefront may be detrimental to the closed-loop contrast as opposed to a relative error on the closed-loop wavefront.

Another issue for these approaches is non-stationary turbulence.²⁰ All these methods require online updating of the control parameters based on the current turbulence parameters. This updating of the control law is computationally very expensive and often involves inverting large matrices.^{17,21} Recently, it was shown that a single Artificial Neural Networks (ANN) can learn to predict the wavefront for varying wind speeds and directions for a simulated AO system.²² This eliminates the need for updating the control law to keep up with the changing turbulence.

Here, we propose to use Reinforcement Learning (RL) for the data-driven optimization of the closed-loop adaptive optics control law. Machine Learning has recently received more attention in the AO community for nonlinear wavefront reconstruction^{23,24} and wavefront prediction.^{22,25} Furthermore, Reinforcement Learning was suggested as a method to directly optimize the contrast in a high contrast imaging system²⁶ and for model-free focal-plane wavefront correction.²⁷ Our approach has some distinct advantages:

- **Model-free:** The algorithm is completely self-optimizing and data-driven. This means that it does not require a model of the disturbances or system dynamics, or knowledge of system parameters.
- **Flexible:** The algorithm allows for an arbitrary optimization objective. The user is free to choose the metric to optimize based on the science goals of the instrument. Furthermore, it can handle multiple inputs, such as in a system with multiple wavefront sensors.
- **Nonlinear:** The proposed method can optimize any parameterized control law. For example, one can optimize the parameters of a nonlinear Artificial Neural Network (ANN) controller. This allows the controller to model the nonlinear dynamics of the system, such as for systems with a Pyramid Wavefront Sensor.

This paper is structured as follows: Section 2 provides an introduction to Reinforcement Learning and describes the relevant algorithm. Section 3 simulates the algorithm's performance in tip-tilt control for vibrations and a power-law input, while section 4 shows the results from a lab experiment. Finally, section 5 draws conclusions and discusses future work.

2. REINFORCEMENT LEARNING CONTROL

Here, we describe the algorithm used in this paper. First, we give a basic introduction to the Reinforcement Learning framework and formulate the AO control problem in this framework. After that, we will discuss the Deterministic Policy Gradient algorithm demonstrate how it can be applied to closed-loop AO control.

Table 1. Overview and explanation of the various terms used in the RL framework.

Symbol	RL term	AO term/Explanation
s	State	Input to the controller.
o	Observation	Measurement of the residual wavefront.
a	Action	Incremental DM commands.
r	Reward	Measure of how good the current state is. This can be arbitrarily defined.
R	Return	Discounted sum of future rewards, the optimization objective.
π_θ	Actor/policy	The control law.
Q_ω	Critic	Estimation of the return/the cost function.

2.1 AO control as a Markov Decision Process

Reinforcement Learning (RL) algorithms deal with Markov Decision Processes (MDPs), discrete-time stochastic control processes. In the RL framework, an agent operates within an environment, which is formally defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P})$ and time index t . Here, $s_t \in \mathcal{S}$ is the state of the environment, $a_t \in \mathcal{A}$ the action taken by the agent, \mathcal{R} the reward function and \mathcal{P} the state transition probabilities defined by:

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

where $P(s_{t+1}|s_t, a_t)$ gives the probability of transitioning from state s_t to state s_{t+1} as a result of action a_t . Eq. 1 is also known as the Markov property.

In the case of AO control the agent is the DM controller. The environment consists of everything but the controller, so among others the evolution of the atmospheric turbulence and the DM dynamics. At each timestep the environment is in a state s_t . The representation of the state will be extensively discussed in section 2.4, but a first guess could be the most recent observation of the residual wavefront o_t . Based on this state, the controller takes an action a_t , following some control law π : $a_t = \pi(s_t)$. Since the controller operates in closed-loop, this action consists of the incremental voltages added to the DM, equivalent to an integral controller. The DM commands and the temporal evolution of the environment result in the transition to a new state s_{t+1} , according to the transition probabilities \mathcal{P} , and an accommodated reward $r_t = \mathcal{R}(s_t, a_t, s_{t+1})$. This reward is a measure of the performance of the controller and the reward function \mathcal{R} can be arbitrarily chosen, as long as it can be calculated at every time step. The goal is to determine the control law π that gives the highest cumulative future reward. To ensure this quantity is finite and to prioritize immediate rewards, future rewards are discounted at a rate $\gamma < 1$, the discount factor. The discounted future return R_t is then defined as:

$$R_t = \sum_{t=0}^{\infty} \gamma^t r_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots \quad (2)$$

The optimization objective can then be formally defined as:

$$J(\pi) = \mathbb{E}_{s \sim \rho^\pi, a \sim \pi}[R]. \quad (3)$$

Here, \mathbb{E} denotes the expectation value and ρ^π the state visitation distribution following control law π . This state visitation distribution describes how often we end up in state s if we follow control law π . In the rest of this work, when we denote an expectation value, it is always over $s \sim \rho^\pi$ and $a \sim \pi$ and we will leave this out for readability. There are a variety of algorithms to find an optimal control law in the above framework.²⁸ One of the most frequently used algorithms for continuous control problems is the (Deep) Deterministic Policy Gradient algorithm,^{29,30} which we will use here.

2.2 (Deep) Deterministic Policy Gradient

The (Deep) Deterministic Policy Gradient (DPG)²⁹ algorithm uses two parameterized models:

- An **Actor** $\pi_\theta(s)$ that models the control law with parameters θ . This model maps the state to the DM commands : $a_t = \pi_\theta(s_t)$.
- A **Critic** $Q_\omega(s, a)$ that models the expected return for a given state and DM command with parameters ω : $Q_\omega(s_t, a_t) = \mathbb{E}[R_t|s_t, a_t]$. This is an estimation of the cost function that has to be optimized.

Both the actor and critic may be any kind of parameterized, differentiable model. These may for example be function approximators, such as Artificial Neural Networks (ANN).³¹ The DPG algorithm with ANNs as function approximators is referred to as the Deep Deterministic Policy Gradient (DDPG) algorithm.³⁰ The algorithm finds the parameters θ of the controller that give the highest expected future return R . Using the actor and the critic we can calculate the gradients of the modelled expected return with respect to the control parameters using the chain rule:

$$\begin{aligned} \nabla_\theta J(\pi_\theta) &= \mathbb{E}[\nabla_\theta Q_\omega(s, a)] \\ &= \mathbb{E}[\nabla_a Q_\omega(s, a) \nabla_\theta \pi_\theta(s)] \end{aligned} \quad (4)$$

Using these gradients we can use a gradient-based optimizer to update the parameters θ of the control law in the direction in which the expected return increases. The quality of the controller thus inherently depends on the ability of the critic $Q_\omega(s, a)$ to successfully model the expected return. The critic can be taught using Temporal Difference (TD) learning, which uses the Bellman equation to bootstrap the expected future reward:

$$\begin{aligned} Q_\omega(s_t, a_t) &= \mathbb{E}[R_t | s_t, a_t] \\ &= \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots] \\ &= \mathbb{E}[r_t + \gamma(r_{t+1} + \gamma r_{t+2} + \dots)] \\ &= \mathbb{E}[r_t + \gamma Q_\omega(s_{t+1}, \pi(s_{t+1}))] \equiv \mathbb{E}[y_t], \end{aligned} \quad (5)$$

where y_t are referred to as the target values. Training the critic is then a supervised learning problem, where we have to find the parameters ω that minimize the mean-squared error between the output of the critic and the target values y_t . The critic loss L is thus given by:

$$L(\omega) = \frac{1}{2} \mathbb{E}[(y_t - Q_\omega(s_t, a_t))^2], \quad (6)$$

and its gradients with respect to the parameters ω by:

$$\nabla_\omega L(\omega) = \mathbb{E}[(y_t - Q_\omega(s_t, a_t)) \nabla_\omega Q_\omega(s_t, a_t)]. \quad (7)$$

We can then use a gradient-based optimizer to update the parameters in the direction that minimizes the loss. However, since the target values depend on the critic itself, this may quickly lead to instabilities in the training process. It is therefore common to use target models Q' and π' to calculate the target values y_t .³⁰ The parameters of these target models slightly lag behind the true models. For every update of the critic, the target models are updated as:

$$\begin{aligned} \theta' &= \tau \theta + (1 - \tau) \theta' \\ \omega' &= \tau \omega + (1 - \tau) \omega'. \end{aligned} \quad (8)$$

Here, $\tau \ll 1$ is a hyper-parameter that determines how quickly the target models are updated.

2.3 Data collection

Updating the actor and critic requires evaluating expectation values over the state visitation distribution. This can naturally be done by collecting tuples (s_t, a_t, r_t, s_{t+1}) while following the control law π_θ . This data collection process is often separated into episodes of finite length, after which the sequence of tuples is saved in a so-called replay buffer, which is our training dataset. For computational reasons, the expectation values are often estimated over a limited number of these tuples, a batch, instead of over the full dataset. To reduce the correlation in a batch the tuples are randomly sampled from the replay buffer.

The training data can also be generated using another controller than the current best controller. This can for example be done by randomly perturbing the DM commands during the data collection, which is called exploration in the RL framework. For example, we can add some zero mean, normally distributed exploration noise with standard deviation σ :

$$a_t = \pi_\theta(s_t) + \mathcal{N}(0, \sigma) \quad (9)$$

This results in the controller not always giving the same DM commands for a given state, such that we can observe the results of these different commands. A common trick is to use large initial exploration noise and slowly decay it. This allows the algorithm to first observe the results of large changes in DM commands and finally observe the result of small changes in these commands. We empirically choose the standard deviation of the exploration noise during episode k as:

$$\sigma_k = \frac{\sigma_0}{1 + \zeta k} \text{ rad.} \quad (10)$$

Here, σ_0 specifies the initial standard deviation of the exploration noise and ζ how quickly it decays. Something that we do not employ here, but is also possible, is to use a completely different controller for the data collection, such as an integrator. This will introduce a slight bias in the models, as the data is then not exactly distributed according to state visitation distribution ρ^π , but this can often be neglected.

2.4 State representation

It is crucial to have a good representation of the state s_t . A first guess for the state might be the most recent wavefront observation, o_t . This may, for example, be retrieved from a dedicated wavefront sensor (WFS). However, the DDPG algorithm assumes that the environment obeys the Markov property (Eq. 1), which requires that the state-transition probabilities are fully described by the current state and the action taken. This is not trivial in closed-loop AO control. First of all, we do not have access to the full state of the environment but only observe a noisy representation with the WFS measurements. Furthermore, the most recent wavefront measurement does not contain all the required information; for example, it lacks information about possible vibrations and wind flow. When the AO loop is closed, we also have to distinguish changes in the turbulence profile from changes due to previous commands, as we only observe the closed-loop residuals. These issues may be solved using state augmentation, where sequences of previous wavefront measurements and applied DM commands are used in the state. However, it may require a large number of previous measurements to have a full representation of the state. For example, Guyon & Males¹⁷ used a vector of the previous 800 measurements to correct tip-tilt vibrations. A disadvantage of this is that it increases the computational demand. Furthermore, all previous measurements are in principle weighted equally, which may lead to additional noise if there are spurious correlations in the training data.

To circumvent the problem of partial observability, it was shown that Recurrent Neural Networks (RNNs) are able to solve control problems even for partially observable MDPs (POMDPs), where the Markov property is not obeyed.^{32,33} RNNs are a type of neural network often used in processing sequential data. Instead of only mapping an input to an output, RNNs also update a memory state m_t based on the previous memory state and the input. This allows it to summarize previous information in this memory state and then use it to calculate the output. This way, RNNs can learn an internal representation of the true state of the system based on sequences of noisy measurements. An additional advantage is that calculating the control commands only requires the most recent state to be propagated through the model. This is computationally less demanding, which may be important for high-order AO systems operating at kHz frequencies. Furthermore, RNNs also effectively use the temporal structure of the data by incorporating basic priors, such as that the most recent measurement is the most relevant instead of using one large vector of components with equal weights. In our case, we will use a Long Short-Term Memory (LSTM) cell³⁴ as recurrent architecture. Training is now done on sequences of state-action-reward tuples using Truncated Backpropagation Through Time (TBTT) over a length l . This version of the algorithm is known as the Recurrent Deterministic Policy Gradient (RDPG) algorithm.³³

In this case, the state for closed-loop AO can consist of only the most recent observation of the wavefront o_t and the previous DM commands a_{t-1} :

$$s_t = (o_t, a_{t-1}) . \quad (11)$$

2.5 Algorithm Overview

In summary, the algorithm can be described by three main steps that are iteratively performed:

1. Collect training sequences $(o_1, a_1, r_1, \dots, o_t, a_t, r_t)$ by running closed-loop and storing the data in the replay buffer.
2. Train the critic using TD learning on the observed data.
3. Improve the control law using the gradients obtained from the critic.

The pseudo-code describing the full algorithm can be found in Algorithm 1.

Algorithm 1: Recurrent Deterministic Policy Gradient for closed-loop AO (based on Heess et al.³³)

Initialize critic parameters ω and actor parameters θ randomly

Initialize target network parameters $\omega' = \omega$, $\theta' = \theta$

Initialize replay buffer

Initialize σ_0

for $k=1, \text{number of episodes}$ **do**

Reset controller memory states $m_t = 0$ and deformable mirror

for $t=1, T$ **do**

Receive wavefront measurement o_t , reward r_t , construct state $s_t = (o_t, a_{t-1})$ and select incremental DM commands $a_t = \pi(s_t) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma_k)$

end

Save trajectories $(o_1, a_1, r_1, \dots, o_T, a_T, r_T)$ in replay buffer

Decay exploration noise $\sigma_k = \frac{\sigma_0}{1+\zeta k}$

for $j=1, \text{number of training steps}$ **do**

Randomly sample N batches of length l from the replay buffer

For each batch i construct state histories $h_t^i = ((o_1^i, a_0^i) \dots, (o_t^i, a_{t-1}^i))$

Calculate the critic targets using the Bellman equation:

$$y_t^i = r_t^i + \gamma Q'_{\omega'}(h_{t+1}^i, \pi'_{\theta'}(h_{t+1}^i))$$

Calculate the sampled critic loss gradient with TBTT:

$$\nabla_{\omega} L(\omega) \approx \frac{1}{Nl} \sum_{i=1}^N \sum_{t=1}^l (y_t^i - Q_{\omega}(h_t^i, t^i)) \nabla_{\omega} Q_{\omega}(s_t^i, a_t^i)$$

Calculate the sampled Policy Gradient with TBTT:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{Nl} \sum_i^N \sum_t^l \nabla_a Q_{\omega}(h_t^i, a_t^i) \nabla_{\theta} \pi_{\theta}(h_t^i)$$

Use a gradient-based optimizer (e.g. Adam) to update the actor parameters θ and critic parameters ω .

Update the target network parameters:

$$\theta' = \tau \theta + (1 - \tau) \theta'$$

$$\omega' = \tau \omega + (1 - \tau) \omega'$$

end

end

3. SIMULATION RESULTS

3.1 Simulation setup

As a proof of concept we apply the algorithm to tip-tilt control in simulations. We simulate an idealized optical system using the `hcpy`³⁵ package for python. Our system has an unobscured, circular aperture of 1 meter in diameter and operates at a wavelength of 1 μm . The observation o_t is the center of gravity (x_t, y_t) of the focal plane image in units of λ/D . The measured center of gravity along with the previous DM commands a_{t-1} are fed into a controller that controls tip and tilt in closed loop. To represent the servo-lag of estimating the wavefront tip and tilt, calculating and applying the control commands, we delay the DM commands by a discrete number of frames:

$$\text{DM}_t = \text{DM}_{t-1} + a_{t-\tau}, \quad (12)$$

where τ is the servo lag. This is equivalent to the controller seeing the state from τ iterations ago, which implies a delay in obtaining the reward. Throughout this section we assume an AO system operating at 1 kHz with a servo lag of 3 frames and do not consider any detector noise. Furthermore, we only consider tip-tilt errors and ignore all higher order modes in the simulations.

3.1.1 Algorithm setup

To minimize the residual tip-tilt jitter we choose the squared PSF center deviation as the reward function:

$$r_t = -\frac{x_t^2 + y_t^2}{b^2}. \quad (13)$$

Here, b is a scaling factor for which we use a value of $4\lambda/D$ in the simulations. We use the same Neural Network architecture for both the actor and the critic. The input consists of the wavefront measurement $o_t = (x_t, y_t)$ and the previously applied tip-tilt commands $a_t = (a_{x,t-1}, a_{y,t-1})$. We use a Long Short-Term Memory³⁴ (LSTM) cell with 64 neurons as the recurrent layer in both the actor and the critic. For the critic the DM commands a_t are appended to the output of the LSTM. After that, we have a fully connected layer with 64 neurons for both networks. Finally, the output of the actor consists of two neurons, the additional tip and tilt commands, while the output of the critic is a single neuron that estimates the expected future return $Q(s_t, a_t)$. The LSTM uses a tanh activation function for the input and output gates and a hard sigmoid for the recurrent gates while the fully connected layer uses the ReLU activation function. The output of the actor again uses a tanh activation function to constrain the incremental DM commands between -1 rad and +1 rad. The output of the critic has a linear activation function. We use the gradient-based Adam optimizer algorithm³⁶ with default parameters except for the learning rate. After every 500 iterations (an episode), we reset the input turbulence and DM shape. To reduce the correlation of the data in a batch early on, we only start training after a certain number of episodes, the warmup. We use a warmup of 5 episodes, which is equivalent to 2500 iterations, and the exploration law given in Eq. 10. The hyperparameters of the RDGP algorithm are given in Table 2.

Table 2. Hyperparameters

Parameter	Value
Actor learning rate	10^{-5}
Critic learning rate	10^{-3}
Target network soft update τ	10^{-3}
Discount factor γ	0.99
Batch size	64
TBTT length l	50
Initial exploration σ_0	0.3 rad
Exploration decay ζ	0.005
Episode length	500 iterations
Number of training steps per episode	500

3.2 Vibration Suppression

Here we consider input turbulence consisting of three pure vibrations along both the x and y directions. Along the x direction we have vibrations with frequencies at 13 Hz, 37 Hz and 91 Hz and along the y direction at 11 Hz, 43 Hz and 87 Hz. The gain of the integrator is optimized by running in closed loop for 1 second and choosing the gain that gives the lowest residual root mean square (RMS) center deviation:

$$\text{RMS} = \sqrt{\langle x_t^2 + y_t^2 \rangle}. \quad (14)$$

Figure 1 shows the evolution of the RMS during the training of the RL controller. It shows that after a few thousand iterations the RL controller outperforms the integrator and eventually reaches an average RMS that is a factor 6 lower than for the integrator. The figure also shows the temporal PSD of the residuals along the x -direction for the fully trained controller without exploration noise. This PSD is estimated using Welch's method³⁷ for a simulation of 10 seconds. The integrator reduces the power of the frequency at 13 Hz, but the other vibrations fall outside its correction bandwidth, even adding power to the 91 Hz vibration. The RL controller almost completely removes the power from all three vibrations. The most power is left at a frequency of 13 Hz. This may be explained by the fact that the period of this vibration is longer than the optimization length of 50 ms. The residuals in the x -direction for a 500-ms simulation for the RL controller, integrator and input spectrum are shown in Fig. 1.

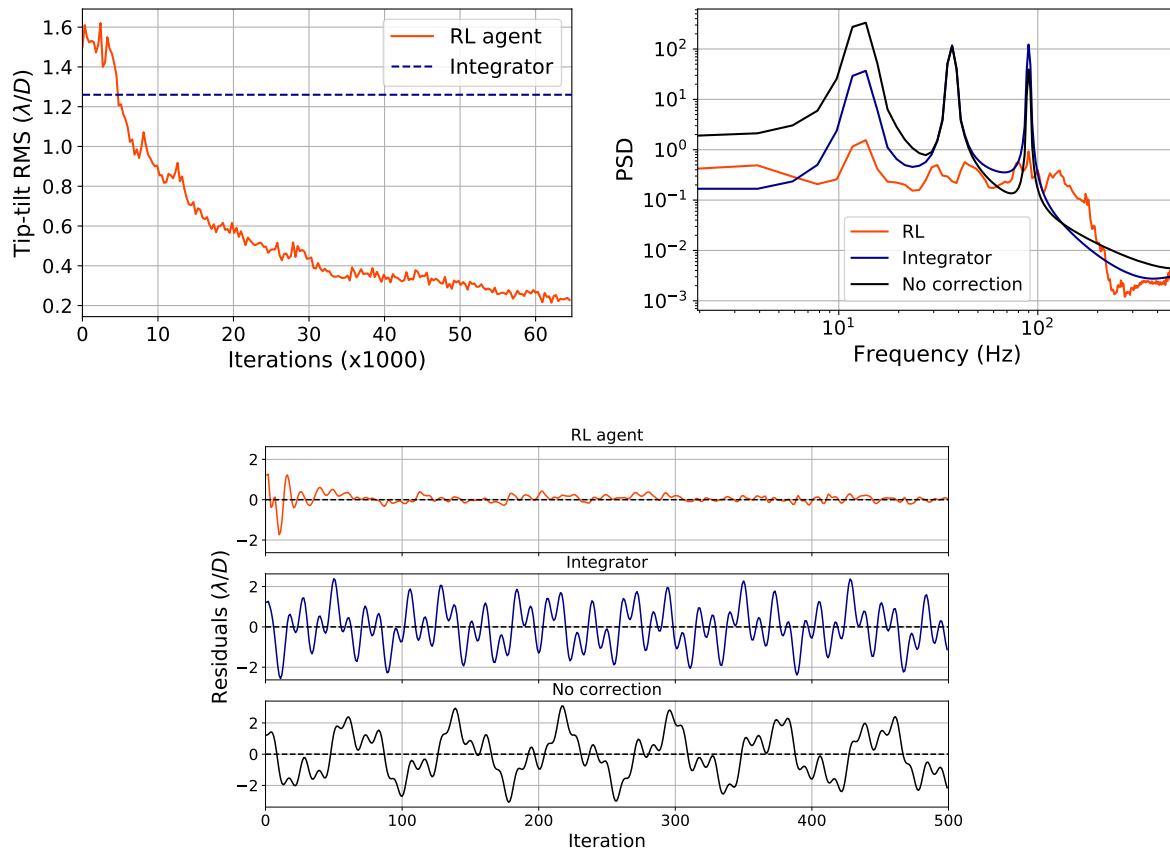


Figure 1. **Top left:** Training curve of the Reinforcement Learning controller under input turbulences consisting of three pure vibrations along both directions. Also shown is the average performance of an integrator with optimal gain. **Top right:** Temporal PSD of the residuals in the x -direction for a simulation of 10 seconds. Note that the lines of the input spectrum and integrator overlap at the peak at 37 Hz. **Bottom:** Residuals for the different controllers in the x -direction for a 500 ms simulation.

3.3 Power-law Disturbances

Next, we consider input disturbances following a temporal power law with a power-law index of $-8/3$ along both directions. For every episode, we generate random power-law time series for 500 ms, neglecting frequencies lower than 2 Hz. The training curve is shown in Fig. 2. We again observe better RMS performance for the RL controller as compared to an integrator with optimized gain. We test the trained controller by running in closed loop for 10 s and calculating the temporal PSD of the residuals, which are also shown in Fig. 2. It shows that the RL controller has a larger rejection power at lower frequencies. The residual PSD of the integrator exhibits a bump around 70 Hz, which is the result of the integrator overshooting. A lower gain would reduce this overshooting but would also decrease its rejection power at lower frequencies. The RL controller can account for already applied commands that are not yet visible in the residual measurements due to the servo lag, allowing it to use a higher gain without overshooting.

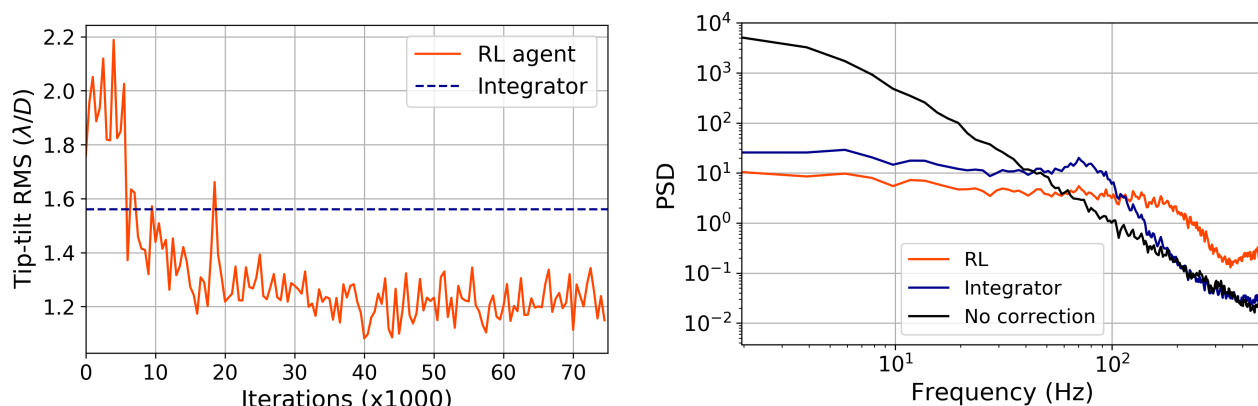


Figure 2. **Left:** Training curve of the Reinforcement Learning controller under power-law input turbulence and the average performance of an integrator with optimal gain. **Right:** Temporal PSD of the residuals in the x-direction for a simulation of 10 seconds.

4. LAB RESULTS

4.1 Experimental Setup

The results in the previous section are for idealized simulations. To validate our results for tip-tilt control we now test the algorithm in the lab. The lab setup is sketched in Figure 3. The first lens focuses a 633-nm He-Ne laser onto a single-mode fiber. The fiber output is then collimated, and a diaphragm defines the input aperture. A 4f system reimages the aperture onto a Holoeye LC2002 Spatial Light Modulator (SLM) with 800x600 pixels. The pupil covers approximately 200x200 pixels on the SLM. Right before the SLM we define the polarization with a polarizer. After a second polarizer we focus the beam onto the camera. The resulting Point Spread Function is also shown in Figure 3.

We use the SLM as both the turbulence generator and the corrector. The polarizers are rotated such that the SLM mainly manipulates the phase profile with minimal amplitude modulation. The voltages of the SLM are controlled with 8-bit resolution. As the response for both small and large voltages becomes highly nonlinear, only values between 25 and 225 are used. Tip and tilt aberrations are introduced by applying a gradient on the SLM at the location of the pupil. Figure 3 shows the calibration of the SLM for the resulting shift of the PSF in the focal plane image. The coefficient refers to the amplitude of the gradient profile, where 1 corresponds to the maximum possible gradient in the pupil. We see that the dynamic range is a shift of approximately $0.1 \lambda/D$. The refresh rate of the SLM is 30 Hz according to the specifications of the manufacturer. We use the SLM at a frequency of approximately 18 Hz. We have attempted to measure the delay of the control loop but this appeared to not be constant. This might be the result of varying processing speeds in the pipeline. On average, there is a delay of 0.1 to 0.15 seconds, which is equivalent to 2-3 frames at 18 Hz. For the RL algorithm we use the same architectures, hyperparameters and reward function as in the simulations (see Section 3.1.1), except for the reward scaling b , for which we now use a value of $0.05 \lambda/D$.

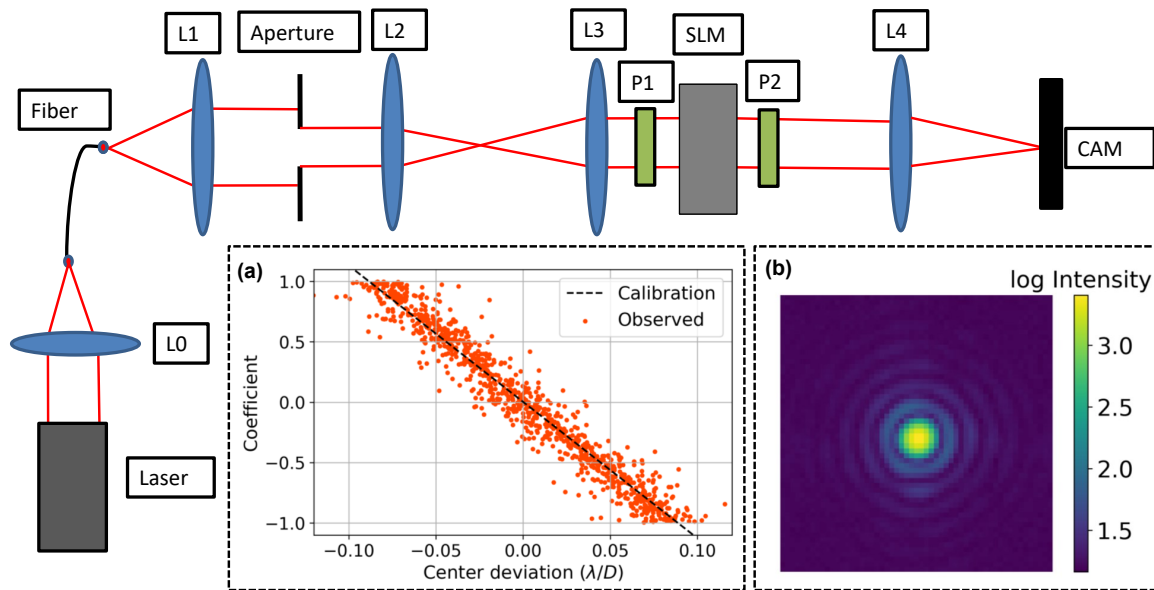


Figure 3. Lab setup to test the RL control algorithm with lenses L0-L4, Polarizers P1 and P2, a Spatial Light Modulator (SLM) and a camera (CAM). Insets: (a) Measured tilt in the focal plane as a function of the applied phase gradient on the SLM. (b) Measured Point Spread Function for the setup.

4.2 Vibration Suppression

We again test the ability of the algorithm to filter out a combination of 3 vibrations along both directions. Figure 4 shows the training curve and the residual temporal PSD for the trained agent and an integrator with optimized gain. It also shows the residuals for a 500-iteration measurement. These results confirm our findings in the simulations as we again see a much reduced power in the vibrations. There is a little more power left in the vibrations as compared to the simulations, and the training time is longer. This is likely the result of the additional noise in the lab, variable control-loop delay and dynamics of the SLM.

4.3 Identification and mitigation of a varying vibration

So far, we have only considered stationary input spectra. Once the input spectrum changes (e.g. the frequency of the vibrations changes), we would have to retrain the controller. However, Recurrent Neural Networks should be able to identify the relevant parameters online, without retraining. To test this, we train the controller with a single but randomly varying input vibration along the x -axis. After every 500 iterations we reset the vibration parameters: the amplitude is randomly sampled between 0 and 1 and the frequency between 0 and 1.8 Hz. The training curve is shown in Fig. 4.3. Once it has converged there is very little variance in the RMS, meaning that the performance is mostly independent of the amplitude and frequency of the input vibration. We compare the performance of the trained controller to that of an integrator with a fixed gain of 0.6. We apply a vibration with a fixed amplitude of 1 and a specific frequency and run closed loop for 2000 iterations. The residual RMS as a function of the frequency of the vibration is shown in Fig. 4.3. We see that the RL controller is able to reduce the RMS for all frequencies. On the other hand, the integrator amplifies the vibrations starting at a frequency of 1.25 Hz. This is the result of the vibration falling outside the correction bandwidth for this gain. However, the RNN is able to adaptively change its gain based on the observed sequence of measurements. This demonstrates that the RNN can learn to identify the relevant parameters of the turbulence, in this case the amplitude and frequency of the vibration, without updating the control law.

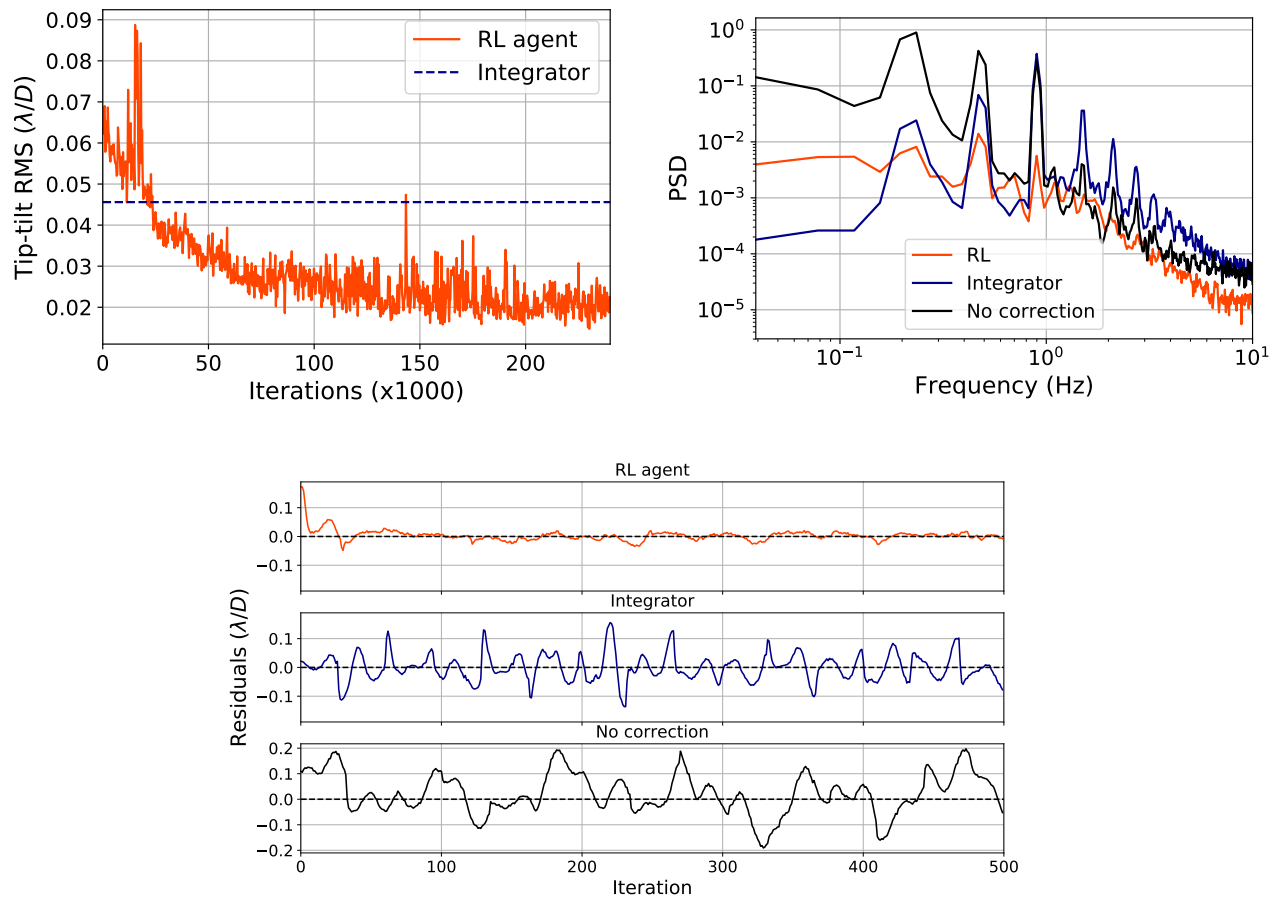


Figure 4. **Top left:** Training curve of the Reinforcement Learning controller for an input spectrum consisting of 3 vibrations along both directions in the lab. Also shown is the average performance of an integrator with optimal gain. **Top right:** Temporal PSD of the residuals in the x-direction for a measurement of 2500 iterations. **Bottom:** Residuals for the different controllers in the x-direction for a 500-iteration measurement.

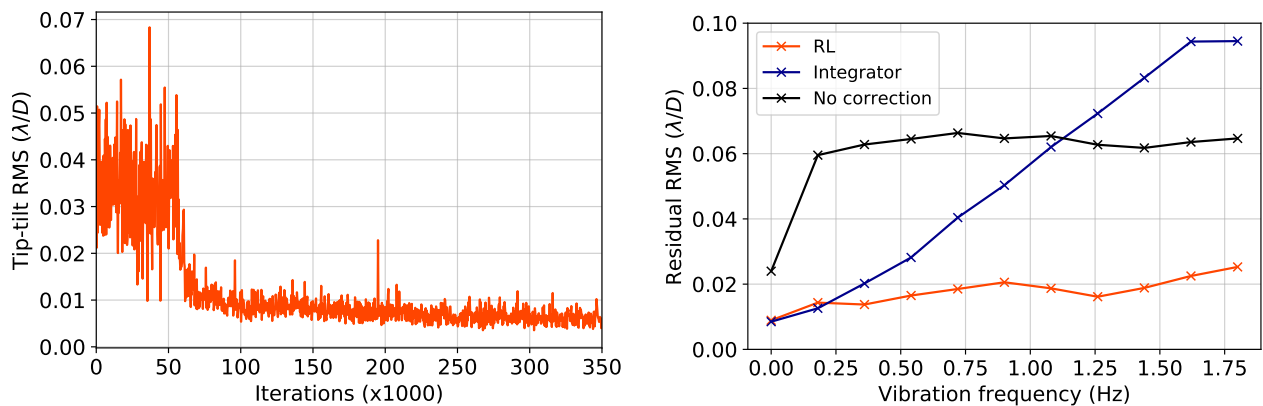


Figure 5. **Left:** Training curve of the Reinforcement Learning controller for a single vibration with random amplitude between 0 and 1 and random frequency between 0 and 1.8 Hz. **Right:** Residual RMS as a function of vibration frequency for a 2000-iteration measurement.

4.4 Power-law Disturbances

We also test the performance for a power-law input spectrum. The training curve and residual temporal PSD are shown in Fig. 6. It takes many iterations before it converges. This is likely because of the noisy response of the SLM. Comparing the residual PSD to that of an integrator with optimized gain, we see results that are similar to the simulations, although the improvement is smaller. We again observe an increased rejection power at low frequencies without the bump due to overshooting.

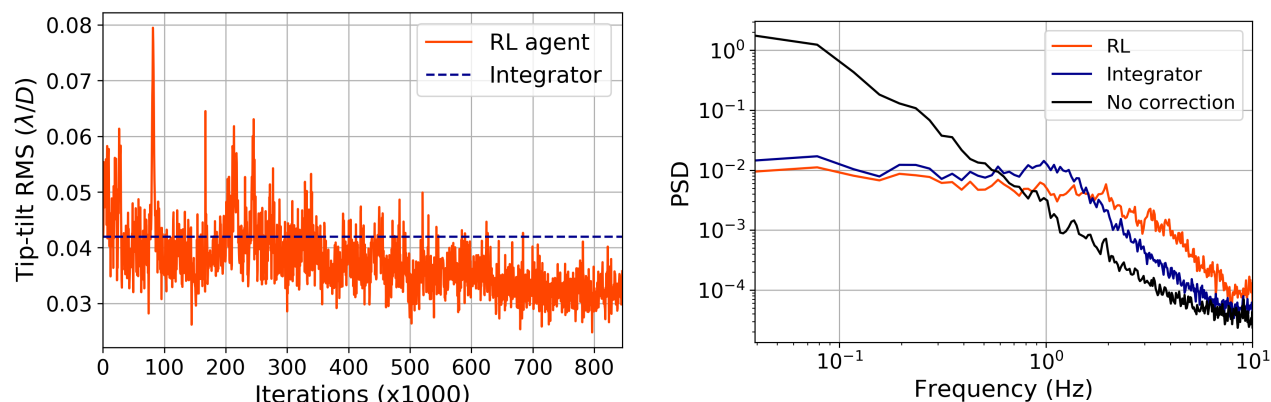


Figure 6. **Left:** Training curve of the Reinforcement Learning controller under power-law input turbulence in the lab. Also shown is the average performance of an integrator with optimal gain. **Right:** Temporal PSD of the residuals in the x-direction for a measurement of 2500 iterations.

5. CONCLUSIONS & OUTLOOK

In conclusion, model-free Reinforcement Learning is a promising approach towards data-driven optimal control for adaptive optics. We have shown how a closed-loop Recurrent Neural Network controller can be trained using the Recurrent Deterministic Policy Gradient algorithm, without any prior knowledge of the system dynamics and disturbances. This RNN only needs the most recent wavefront measurement and the previous DM commands at every time step. First, we applied the algorithm to tip-tilt control for a simulated, ideal AO system. We demonstrated that our approach can learn to mitigate a combination of tip-tilt vibrations, reducing the residual RMS by a factor of ~ 6 as compared to an optimal-gain integrator. We also show a $\sim 25\%$ decrease in residual RMS for power-law input turbulence compared to the optimal gain integrator. These experiments were then repeated in a lab setup. For the vibration mitigation we observe a decrease in RMS by a factor 2.2 and a $\sim 25\%$ decrease in residual RMS for power-law input turbulence. Furthermore, we demonstrated that a single Recurrent Neural Network can identify and correct a vibration with varying amplitude and frequency without updating the control parameters.

5.1 Towards full wavefront control

In future work we will test the algorithm for the predictive control of a high-order DM. Our approach is almost directly applicable to full wavefront control, as we will only have to modify the reward function and the model architectures. To optimally make use of the spatio-temporal structure of the input data in full wavefront control, we propose to combine Recurrent Neural Networks and Convolutional Neural Networks (CNNs). As the reward for the full wavefront control, one could use a modal approach, where the modal residuals are separately minimized. In analogy to our tip-tilt approach here, we can also locally minimize the spot deviation of a Shack-Hartmann wavefront sensor. Since we can use an arbitrary reward, it is, in principle, also possible to use a focal-plane quantity such as the Strehl ratio or contrast. However, the accurate estimation of these quantities at kHz frequencies might be an issue in practice.

We showed that a single RNN can identify and mitigate vibrations with different frequencies and amplitudes without requiring an update of the control law. In a similar way, we expect it to be able to track changes in

wind speed and direction when trained using data obtained during a variety of atmospheric conditions, as was shown in 22. This means that it would not require real-time updating of the control law. Since training of the controller can also be done completely separate from the data collection, this can, for example, even be done during the day. This implies that the computational demand of updating the controller is not an issue. Once a general model is learned, it would only require retraining once something in the instrument changes.

In this work we started with a randomly initialized control law. However, we already empirically know what a reasonable control law will be. We could therefore already initialize these weights such that the output is equal to that of, for example, an integrator. Furthermore, we can also train on data collected using another controller. It is therefore possible to simply train on data collected using the current controller installed at the telescope. We can therefore already learn very reasonable models without even requiring time on the telescope, by learning from past data.

While the computational demand is not an issue for tip-tilt control, it may become crucial for ELT-sized XAO systems. Even though the computation time is independent of the temporal order, the architecture used here would be expensive to apply to every mode. There is a trade-off between predictive power of the controller and the computational time. One could optimize the architecture given a constraint on the computation time for calculating the real-time commands.

Finally, another advantage is the algorithm's ability to learn a nonlinear control law. It was recently shown that CNNs are able to reconstruct the nonlinearities of Fourier-based wavefront sensors.²³ However, these reconstructors can only be calibrated on phase profiles spanned by the DM modes, while the response also depends on higher-order modes. Reinforcement Learning could allow us to calibrate a nonlinear controller on-sky. This may be useful for current and future instruments using a Pyramid Wavefront Sensor.

REFERENCES

- [1] Cantalloube, F., Farley, O. J., Milli, J., Bharmal, N., Brandner, W., Correia, C., Dohlen, K., Henning, T., Osborn, J., Por, E., Suárez Valles, M., and Vigan, A., "Wind-driven halo in high-contrast images: I. Analysis of the focal-plane images of SPHERE," *Astronomy and Astrophysics* **638** (2020).
- [2] Lozi, J., Guyon, O., Jovanovic, N., Takato, N., Singh, G., Norris, B., Okita, H., and Bando, T., "Characterizing vibrations at the Subaru Telescope for the Subaru coronagraphic extreme adaptive optics instrument," *4*(04), 1 (2018).
- [3] Hartung, M., Hayward, T., Saddlemyer, L., Poyneer, L., Cardwell, A., Cavedoni, C., Cho, M., Chilcote, J. K., Collins, P., Dillon, D., Galvez, R., Gausachs, G., Goodsell, S., Guesalaga, A., Hibon, P., Larkin, J., Macintosh, B., Palmer, D., Sadakuni, N., Savransky, D., Serio, A., Rantakyro, F., and Wallace, K., "On-sky vibration environment for the Gemini Planet Imager and mitigation effort," **9148**, 91480N (2014).
- [4] Sauvage, J.-F., Fusco, T., Petit, C., Meimon, S., Fedrigo, E., Suarez Valles, M., Kasper, M., Hubin, N., Beuzit, J.-L., Charton, J., Costille, A., Rabou, P., Mouillet, D., Baudoz, P., Buey, T., Sevin, A., Wildi, F., and Dohlen, K., "SAXO, the eXtreme Adaptive Optics System of SPHERE: overview and calibration procedure," *Adaptive Optics Systems II* **7736**, 77360F (2010).
- [5] Lloyd, J. P. and Sivaramakrishnan, A., "Tip-Tilt Error in Lyot Coronagraphs," *The Astrophysical Journal* **621**, 1153–1158 (2005).
- [6] Mawet, D., Pueyo, L., Moody, D., Krist, J., and Serabyn, E., "The Vector Vortex Coronagraph: sensitivity to central obscuration, low-order aberrations, chromaticism, and polarization," *Proc. of SPIE* **7739**, 773914 (2010).
- [7] Gendron, E. and Lena, P., "Astronomical adaptive optics. I. Modal control optimization," *Astronomy and Astrophysics* **291**(1), 337–337 (1994).
- [8] Gendron, E. and Lena, P., "Astronomical adaptive optics. II. Experimental results of an optimized modal control," *Astronomy and Astrophysics Supplement Series* **111**(111), 153 (1995).
- [9] Paschall, R. N. and Anderson, D. J., "Linear quadratic Gaussian control of a deformable mirror adaptive optics system with time-delayed measurements," *Applied Optics* **32**(31), 6347 (1993).
- [10] Correia, C., Véran, J.-P., and Herriot, G., "Advanced vibration suppression algorithms in adaptive optics systems," *Journal of the Optical Society of America A* **29**(3), 185 (2012).

- [11] Petit, C., Conan, J.-M., Kulcsár, C., Raynaud, H.-F., and Fusco, T., “First laboratory validation of vibration filtering with LQG control law for Adaptive Optics,” *Optics Express* **16**(1), 87 (2008).
- [12] Sivo, G., Kulcsár, C., Conan, J.-M., Raynaud, H.-F., Gendron, É., Basden, A., Vidal, F., Morris, T., Meimon, S., Petit, C., Gratadour, D., Martin, O., Hubert, Z., Sevin, A., Perret, D., Chemla, F., Rousset, G., Dipper, N., Talbot, G., Younger, E., Myers, R., Henry, D., Todd, S., Atkinson, D., Dickson, C., and Longmore, A., “First on-sky SCAO validation of full LQG control with vibration mitigation on the CANARY pathfinder,” *Optics Express* **22**(19), 23565 (2014).
- [13] Petit, C., Sauvage, J.-F., Fusco, T., Sevin, A., Suarez, M., Costille, A., Vigan, A., Soenke, C., Perret, D., Rochat, S., Barrufolo, A., Salasnich, B., Beuzit, J.-L., Dohlen, K., Mouillet, D., Puget, P., Wildi, F., Kasper, M., Conan, J.-M., Kulcsár, C., and Raynaud, H.-F., “SPHERE eXtreme AO control scheme: final performance assessment and on sky validation of the first auto-tuned LQG based operational system,” *Adaptive Optics Systems IV* **9148**, 91480O (2014).
- [14] Correia, C., Raynaud, H.-F., Kulcsár, C., and Conan, J.-M., “On the optimal reconstruction and control of adaptive optical systems with mirror dynamics,” *Journal of the Optical Society of America A* **27**(2), 333 (2010).
- [15] Poyneer, L. A., Palmer, D. W., Macintosh, B., Savransky, D., Sadakuni, N., Thomas, S., Véran, J.-P., Follette, K. B., Greenbaum, A. Z., Mark Ammons, S., Bailey, V. P., Bauman, B., Cardwell, A., Dillon, D., Gavel, D., Hartung, M., Hibon, P., Perrin, M. D., Rantakyrö, F. T., Sivaramakrishnan, A., and Wang, J. J., “Performance of the Gemini Planet Imager’s adaptive optics system,” *Applied Optics* **55**(2), 323 (2016).
- [16] Dessenne, C., Madec, P.-Y., and Rousset, G., “Modal prediction for closed-loop adaptive optics,” *Optics Letters* **22**(20), 1535 (1997).
- [17] Guyon, O. and Males, J., “Adaptive Optics Predictive Control with Empirical Orthogonal Functions (EOFs),” (2017).
- [18] Jensen-Clem, R., Bond, C. Z., Cetre, S., McEwen, E., Wizinowich, P., Ragland, S., Mawet, D., and Graham, J., “Demonstrating predictive wavefront control with the Keck II near-infrared pyramid wavefront sensor,” *Proc. of SPIE*, 34 (2019).
- [19] Van Kooten, M. A., Doelman, N., and Kenworthy, M., “Robustness of prediction for extreme adaptive optics systems under various observing conditions: An analysis using VLT/SPHERE adaptive optics data,” *Astronomy and Astrophysics* **636**, 81 (2020).
- [20] Van Kooten, M., Doelman, N., and Kenworthy, M., “Impact of time-variant turbulence behavior on prediction for adaptive optics systems,” *Research Article Journal of the Optical Society of America A* (2019).
- [21] Gray, M., Petit, C., Rodionov, S., Bocquet, M., Bertino, L., Ferrari, M., and Fusco, T., “Local ensemble transform Kalman filter, a fast non-stationary control law for adaptive optics on ELTs: theoretical aspects and first simulation results,” *Optics Express* **22**(17), 20894 (2014).
- [22] Liu, X., Morris, T., Saunter, C., De Cos Juez, F. J., González-Gutiérrez, C., and Bardou, L., “Wavefront prediction using artificial neural networks for open-loop adaptive optics,” *Monthly Notices of the Royal Astronomical Society* **496**(1), 456–464 (2020).
- [23] Landman, R. and Haffert, S. Y., “Nonlinear wavefront reconstruction with convolutional neural networks for Fourier-based wavefront sensors,” *Optics Express* **28**(11), 16644 (2020).
- [24] Norris, B. R. M., Wei, J., Betters, C. H., Wong, A., and Leon-Saval, S. G., “An all-photonic focal-plane wavefront sensor,” *Nature Communications* (2020), 1–9 (2020).
- [25] Swanson, R., Kutulakos, K., Sivanandam, S., Lamb, M., and Correia, C., “Wavefront reconstruction and prediction with convolutional neural networks,” (July), 52 (2018).
- [26] Radhakrishnan, V. M., Doelman, N., and Keller, C. U., “Optimization of contrast in adaptive optics for exoplanet imaging,” *Proc. of SPIE* (July 2018), 172 (2018).
- [27] Sun, H., Kasdin, N. J., and Vanderbei, R., “Identification and adaptive control of a high-contrast focal plane wavefront correction system,” *Journal of Astronomical Telescopes, Instruments, and Systems* **4**(04), 1 (2018).
- [28] Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A., “Deep reinforcement learning: A brief survey,” *IEEE Signal Processing Magazine* **34**(6), 26–38 (2017).

- [29] Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M., “Deterministic policy gradient algorithms,” *31st International Conference on Machine Learning, ICML 2014* **1**, 605–619 (2014).
- [30] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D., “Continuous control with deep reinforcement learning,” *4th International Conference on Learning Representations, ICLR 2016* (2016).
- [31] Goodfellow, I. J., Bengio, Y., and Courville, A., [*Deep Learning*], MIT Press, Cambridge, MA, USA (2016). <http://www.deeplearningbook.org>.
- [32] Wierstra, D., Foerster, A., Peters, J., and Schmidhuber, J., “Solving deep memory POMDPs with Recurrent Policy gradients,” (2007).
- [33] Heess, N., Hunt, J. J., Lillicrap, T. P., and Silver, D., “Memory-based control with recurrent neural networks,” (2015).
- [34] Hochreiter, S. and Jürgen Schmidhuber, J., “Long Short-term Memory,” *Neural Computation* **9**(8), 1735–1780 (1997).
- [35] “High Contrast Imaging for Python (HCIPy): an open-source adaptive optics and coronagraph simulator,” *Adaptive Optics Systems VI* **10703**, 152 (2018).
- [36] Kingma, D. P. and Ba, J. L., “Adam: A method for stochastic optimization,” *3rd International Conference on Learning Representations, ICLR 2015* (dec 2015).
- [37] Welch, P. D., “The Use of Fast Fourier Transform for the Estimation of Power Spectra,” *Digital Signal Processing* (2), 532–574 (1975).