Adaptive Optics Control with Reinforcement Learning: First steps

Bartomeu Pou*†, Eduardo Quiñones*, Mario Martín†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {bartomeu.poumulet, eduardo.quinones}@bsc.es, mmartin@cs.upc.edu

Keywords—Reinforcement Learning, Adaptive Optics, Non-linear Control, Machine Learning.

I. Extended Abstract

A. *Introduction*

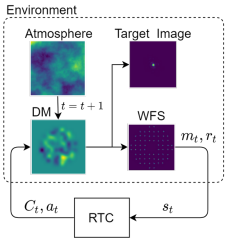
When planar wavefronts from distant stars traverse the atmosphere, they become distorted due to the atmosphere's inhomogeneous temperature distribution. Adaptive Optics (AO) is the field in charge of correcting those distortions allowing high-quality observations of distant targets. The AO solution is composed of three main components: a deformable mirror (DM) that corrects the deformation in the wavefront, a wavefront sensor (WFS) that allows characterising the current turbulence in the wavefront and a real time controller (RTC) that issues commands to, via the deformation of the DM, correct the wavefront. Usually, the operations are performed on closed-loop with stringent real-time requirements (in the order of $10^3 - 10^4$ actions per second). At each iteration, the WFS observes the wavefront after being corrected by the DM and the RTC issues the commands to correct for the evolution of turbulence and previous uncorrected errors (Figure 1 left).

One of the primary sources of error for an AO control algorithm is the temporal error. The delay between characterising the turbulence with the WFS and setting the desired commands in the DM creates the need that any successful control approach must take into account past commands and the probable evolution of the atmosphere in this gap of time. To do that, the most common approach in AO are variants of Linear Quadratic Gaussian (LQG) with Kalman filters with one of its initial iterations presented in [1]. Usually, a linear model of the system's evolution is built with a set of parameters that are usually fitted based on observations or on theoretical assumptions, which limits the capability of the system to correct the turbulence.

In this paper, we present a novel solution based on Reinforcement Learning (RL), based on a reward signal to be optimised, that does not need any previously built model (as LQG) and is non-linear. RL has been already applied in the domain of AO, however, it has been limited to WFS-less systems (e.g. [2]) or, more recently, to control a very limited number of actuators [3]. This work's main practical objective is to be applied in the 8.2 m Subaru telescope (located in Hawaii), which includes thousands of actuators.

B. *AO Control: Integrator with gain*

The traditional AO control algorithm is the integrator with gain. At each iteration, the WFS obtains a vector of measurements, m, where each element indicates a local deviation from the seen wavefront to a planar wavefront. The relationship between m and commands in the DM at a timestep, t, can be approximated as a linear relationship:



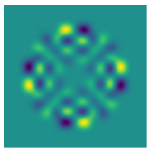


Fig. 1: Left: AO closed-loop. Right: Mode example.

$$m_t = D \cdot c_t \tag{1}$$

Where D is the interaction matrix obtained with a least squares approach method on the loss $||m - Dc||^2$. By inverting the interaction matrix in equation (1), we obtain the commands to be applied to the DM to correct the current wavefront deviations on ideal conditions. To deal with non-ideal issues, such as the temporal error, integration with past commands, C, with a gain factor, g, is used:

$$C_t = C_{t-1} + gc_t \tag{2}$$

C. *Adaptive Optics as a Reinforcement Learning problem*

RL [4] is concerned of finding a function (called policy, $\pi(\theta)$ parametrized with weights θ) that maximises the expected cumulative reward (r) obtained by interacting with an environment. To do so, RL maps the state describing the environment (s) to actions (a) with the objective of obtaining the optimal policy, $\pi^*(\theta)$.

$$\pi^*(\theta) = arg\,\underset{\theta}{max}\,\underset{env}{\mathbb{E}}\left[\sum_i r_i(\pi(\theta))\right] \tag{3}$$

Concretely, RL requires the following elements:

1) The states, s: Defined as the union of the integrator commands, c, which will give information of current perturbation in the atmosphere and past integrated commands, C, which will give information of commands that will be applied in the next timesteps (due to delay), and the evolution of the atmosphere at every time step t: $s_t = (c_t, C_{t-1}, C_{t-2}, ..., C_{t-n})$.

The commands issued by the RTC are usually a vector of n_a dimensions ($C \in \mathbb{R}^{n_a}$) where each element of the command vector controls one actuator in charge of deforming the mirror. This way of handling the commands is said to be zonal as each value of the command vector only modifies a particular zone of the mirror. However, one can build a modal basis, e.g. by using

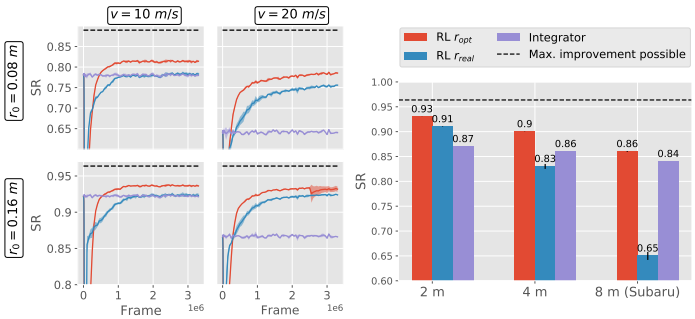


Fig. 2: Left: Training curves (77 modes, D=2 m). Right: Avg final performance (62 modes, r_0=0.16 m, v=20 m/s). Results averaged over 3 seeds.

the Zernike polynomials [5] (see Figure 1), to act globally in the whole DM with each element of the command vector. The usage of modal basis has two benefits: (1) we can just correct a subset of modes if the number of actuators to control is problematically high and (2) RL method performance depends on the feature definition [4]. Empirically, we have observed that a value of $n = 3$ for the state and using a modal basis for the commands C_t leads to better performance.

2) The actions, a: Defined as a correction applied to the commands computed with the "integration with gain" AO control, as follows: $C_t = C_{t-1} + gc_t + a$.

3) The reward, r: Defined as a function that determines how well the turbulence distortion has been corrected. To do that, we apply two different strategies: (1) A reward based on the spatial variance of the wavefront phase ϕ, $r_{opt,t} = -var(\phi_t)$, in which a variance of 0 indicates that all the wavefront points are on phase, hence, the wavefront is planar; and (2) a reward based on the average measurements, m_t, squared: $r_{real,t} = -avg(m_t^2)$. The former strategy is optimal but unrealistic as it is not possible to get the variance of the wavefront at each timestep; the latter provides an approximation of the former strategy but can be obtained at each time step.

4) The algorithm: We choose Soft Actor Critic, which slightly modifies eq. 3 to include the entropy of the policy, $\pi(\theta)$, as a regularisation term [6].

D. Results

This section evaluates the AO RL controller in a different range of atmospheric conditions, specifically, different values for Fried parameter, r_0, which a lower value denotes a higher strength of turbulence, and wind speed, v, which a higher value will drive up the temporal error. Moreover, it evaluates the performance of RL when increasing the telescope diameter, D, and thus the complexity of the problem as the number of actuators to control, and the number of measurements of the WFS to process, increases as well, when considering the optimal (r_{opt}) and the realistic (r_{real}) rewards. The performance is measured in terms of Strehl Ratio ($0 \leq SR \leq 1$), the ratio between the peak intensity of the target image and its theoretical maximum. The experiments presented use an AO control simulator named COMPASS [7], including the simulation of the atmosphere and the AO control, executed on a IBM Power9 8335-GTH CPU (40 cores) with a GPU NVIDIA V100 (16 GB).

Figure (2) *left* evaluates different atmospheric conditions. We can observe that the RL agent outperforms the traditional integrator and is both robust to variations of v and r_0 with both reward functions, i.e., r_{opt} and r_{real}. RL agent's quasi-constant performance in terms of wind speed may indicate that we are solving mainly temporal error.

Figure (2) *right* compares the performance of the RL agent when controlling 62 modes while increasing the telescope diameter with a fixed atmospheric configuration. While the RL agent with a limited number of modes performs better when compared with the integrator, the agent is incapable of scaling to bigger diameters with the realistic reward function, r_{real}. It therefore remains as future work to derive a more efficient reward function. Furthermore, while the use of a modal basis allows to significantly reduce the state's size and so avoid the problem of the curse of dimensionality [4], it remains as a future work as well, to control a higher number of modes. In addition to performance, we must take into account the inference time. Currently, for the given machine and 62 modes, the inference time is $\sim 1.2\ ms$ which is below the threshold of $2\ ms$ to not increase the delay as to affect proper operation of the telescope. However, we must take into account that for large telescopes we will probably end up controlling a higher number of modes hence increasing the inference time.

E. Conclusion

We have presented a novel AO control based on RL that outperforms traditional controllers in a set of limited experiments. However, we must address some challenges before its application in the real world.

II. Acknowledgment

This research has been conducted in collaboration with Dr. Damien Gratadour (Paris Observatory, PSL University and Australian National University).

References

[1] R. N. Paschall and D. J. Anderson, "Linear quadratic gaussian control of a deformable mirror adaptive optics system with time-delayed measurements," *Applied optics*, vol. 32, no. 31, pp. 6347–6358, 1993.

[2] K. Hu, Z. Xu, W. Yang, and B. Xu, "Build the structure of wfsless ao system through deep reinforcement learning," *IEEE Photonics Technology Letters*, vol. 30, no. 23, pp. 2033–2036, 2018.

[3] R. Landman, S. Y. Haffert, V. M. Radhakrishnan, and C. U. Keller, "Self-optimizing adaptive optics control with reinforcement learning," in *Adaptive Optics Systems VII*, vol. 11448. International Society for Optics and Photonics, 2020, p. 1144849.

[4] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[5] R. J. Noll, "Zernike polynomials and atmospheric turbulence," *JOsA*, vol. 66, no. 3, pp. 207–211, 1976.

[6] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.

[7] F. Ferreira, D. Gratadour, A. Sevin, N. Doucet, F. Vidal, V. Deo, and E. Gendron, "Real-time end-to-end ao simulations at elt scale on multiple gpus with the compass platform," in *Adaptive Optics Systems VI*, vol. 10703. International Society for Optics and Photonics, 2018, p. 1070347.

Bartomeu Pou is a PhD student at Barcelona Supercomputing Center with research focused on applying reinforcement learning to adaptive optics in large telescopes. He has a background in physics (bachelor's) and artificial intelligence (master's) and previously has worked on Accenture as a data scientist on the domains of supply chain and healthcare.

Adaptive Optics Control with Reinforcement Learning: First steps

B. Pou[1,2], E. Quiñones[1], D. Gratadour [3,4], M. Martín[2]

[1] Barcelona Supercomputing Center (BSC).
[2] Universitat Politècnica de Catalunya (UPC).
[3] Research School of Astronomy and Astrophysics, Australian National University.
[4] LESIA, Observatoire de Paris, Universite PSL, CNRS, Sorbonne Universite, Univ. Paris Diderot, Sorbonne Paris Cite.

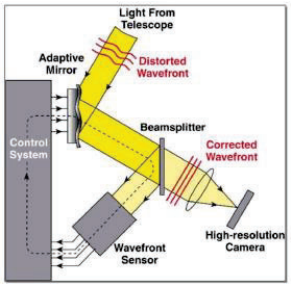
1. Motivation

- In ground-based telescopes, the **light from distant stars is distorted** due to small variations of index of refraction in the atmosphere
- **Adaptive Optics** (AO) systems are responsible of correcting the distortion by means of a **deformable mirror (DM).**

Image of Ground-Based Telescopes
Credit: Claire E. Max, UCSC

2. Adaptive Optics (AO)

- An AO system characterizes the **distortion (m_t)** using a **Wavefront sensor (WFS)**
- **A Real-time Controller (RTC)** computes commands to the **DM actuators (c_t)** to correct observed distortion, considering the following **linear relationship**:
  - (1) $c_t = R \cdot m_t$
  - (2) $C_t = C_{t-1} + g \cdot c_t$
- **The RTC have high-performance and real-time requirements**
  - Commands must be issued every ~2ms to ensure the correct operation



Credit: Claire E. Max, UCSC.

AO Control-Loop

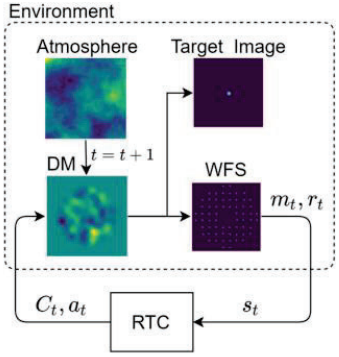
Large telescopes includes non-linear effects not captured by current RTC that diminishes the performance of telescopes

3. Reinforcement Learning (RL)

RL allows capturing non-linear effects not addressed by linear solutions

RL objective: find a function, $\pi(s)$, that maps **states (*s*)** to **actions (*a*)** that maximizes a cumulative **reward function (*r*)** via trial and error**.**

- ***a*** corresponds to a **correction term** to the linear RTC: $C_t = C_{t-1} + g \cdot c_t + a_t$
- ***s*** **=** $(c_t, C_{t-1}, C_{t-2}, \dots, C_{t-N})$ corresponds to the current linear and **previous commands** and provides information about:
  - Commands that will be executed in the future
  - Statistics of evolution of the atmosphere.
- **r** is based on spatial variance of the wavefront phase, ϕ_t and average of measurements squared: $r_{opt} = -var(\phi_t)$ and $r_{real} = -avg(m_t^2)$, respectively



AO loop with RL elements

Our RL agent does not consider a single DM actuator but **global orthogonal shapes in the DM** inspired in **Zernike polynomials [1].**

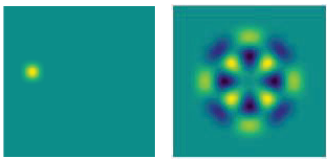
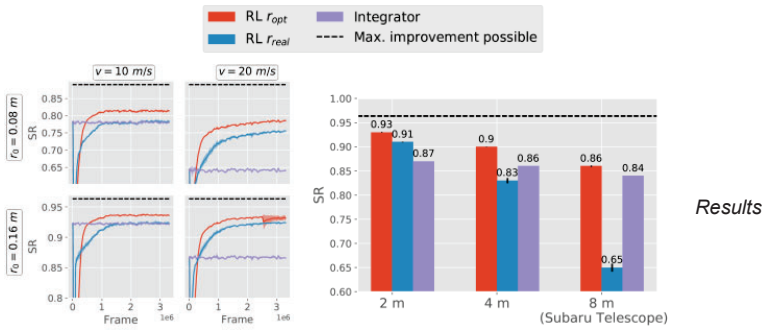


Image of DM shape.
(Left) acting on a single actuator.
(Right) acting on a single mode.

4. Results

Experiments:

- Simulated in **COMPASS** [2] (GPU-based high-performance AO simulations).
- Characterisation of different range of atmospheric conditions.
  - **Fried parameter**, r_0: inverse relationship to strength of turbulence.
  - **Wind speed**, v. Related to temporal error.
- Different diameter **(D)** of telescope.
- Measuring results in **Strehl Ratio** ((*worst*) 0 ≤ SR ≤ 1 (*best*)).



a) RL agent (77 modes) on 2m telescope with different atmospheric conditions.

b) RL agent controlling 62 modes with different D. Atmospheric conditions constant.

Results

5. Conclusions

- RL agent **tackles non-linear effects such as temporal error**.
- **Dimensionality problem**.
- Realistic reward function does not work for large telescopes.

6. Future work

- **Multi-agent system**: each agent controls a small amount of modes.
- **Preliminary results** show an **improvement** over the integrator with a large telescope and realistic reward function.

References:

[1] R. J. Noll, "Zernike polynomials and atmospheric turbulence,"JOsA,vol. 66, no. 3, pp. 207–211, 1976.

[2] F. Ferreira, D. Gratadour, A. Sevin, N. Doucet, F. Vidal, V. Deo, and E. Gendron, "Real-time end-to-end ao simulations at elt scale on multiple gpus with the compass platform," in Adaptive Optics Systems VI, vol. 10703.International Society for Optics and Photonics, 2018,p. 1070347.

Figure 1 and 2 extracted from https://www.ucolick.org/~max/289/ Lecture 1.