Wavefront Sensorless Adaptive Optics using Reinforcement Learning

- Soft Actor-Critic Controller

Presented by: Runnan Zou

Outline

- Adaptive Optics in Free-space Satellite-to-Ground Communication
- Wavefront Sensor-based and Wavefront Sensorless Adaptive Optics
- Background of Reinforcement Learning
- Soft Actor-Critic Controller
- Results of Simulation
- Conclusion & Future work
- Our team

Motivation and Contribution

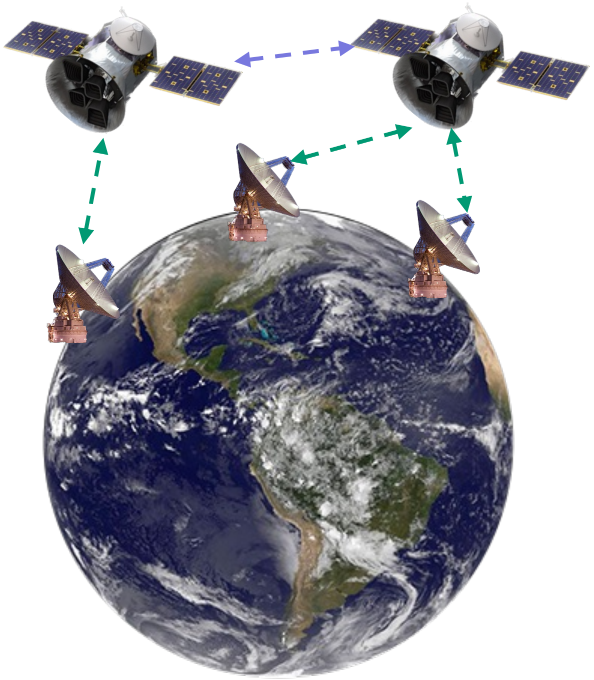
- Developing a budget-friendly wavefront sensorless adaptive optics system.
- Online model-free off-policy reinforcement learning framework, soft actor-critic controller.
- Tuning hyperparameters of soft actor-critic algorithm
- Simulations in both static and dynamic atmospheres

Free-space Satellite-to-Ground Communication

- Optical communication
- Shorter wavelength than radio waves.

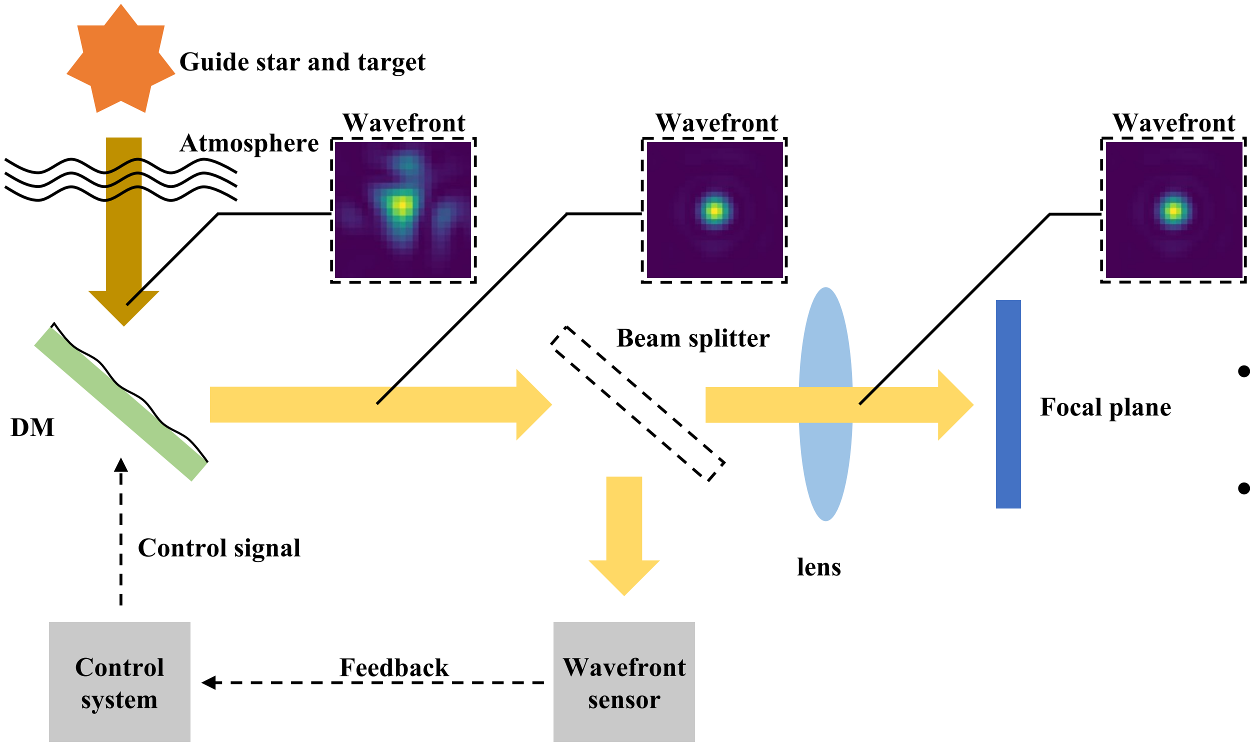
- Pros:
  - Concentrated power
  - Secured connection
  - High data transmission rate

- Cons:
  - Affected by atmospheric turbulence



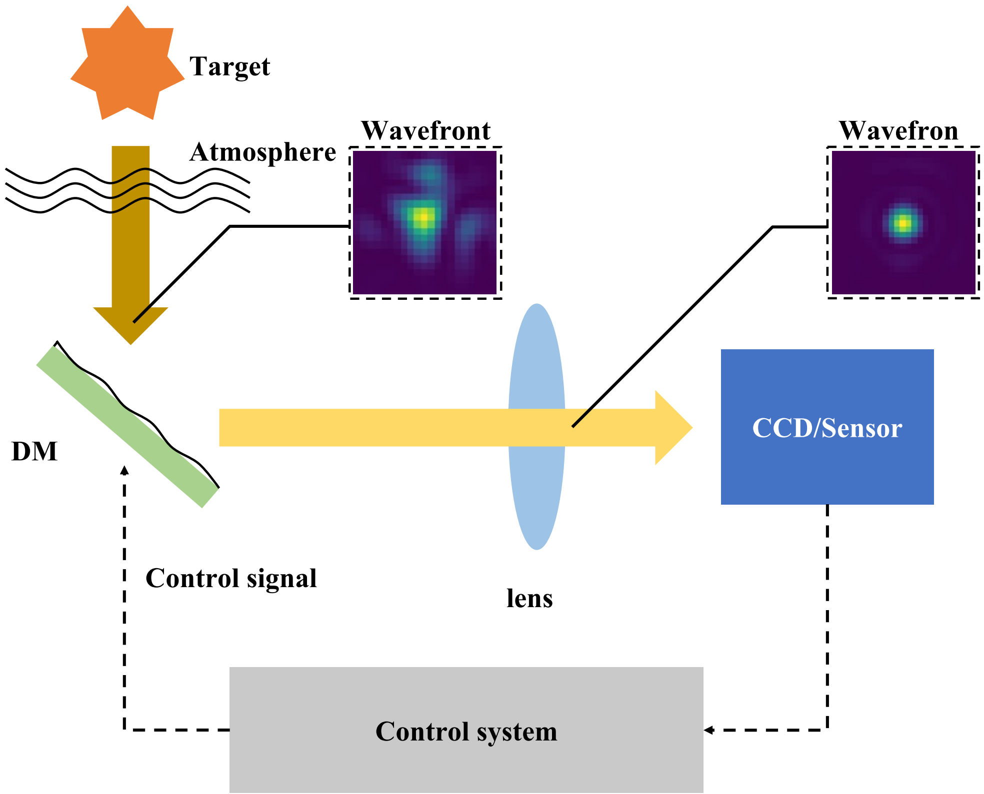
Photograph: NASA

Wavefront Sensor-based Adaptive Optics



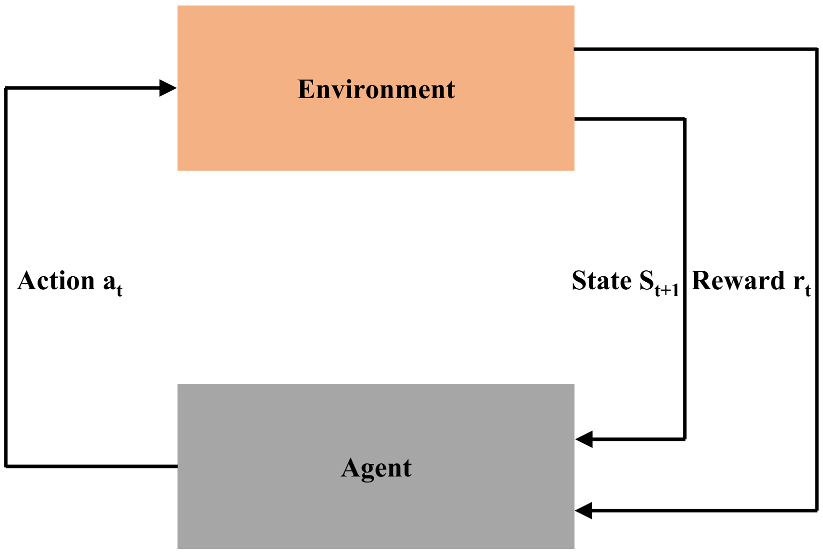
- Pros:
  - Precise feedback
- Cons:
  - Complex structure
  - Expensive
  - Require calibration
  - Slower

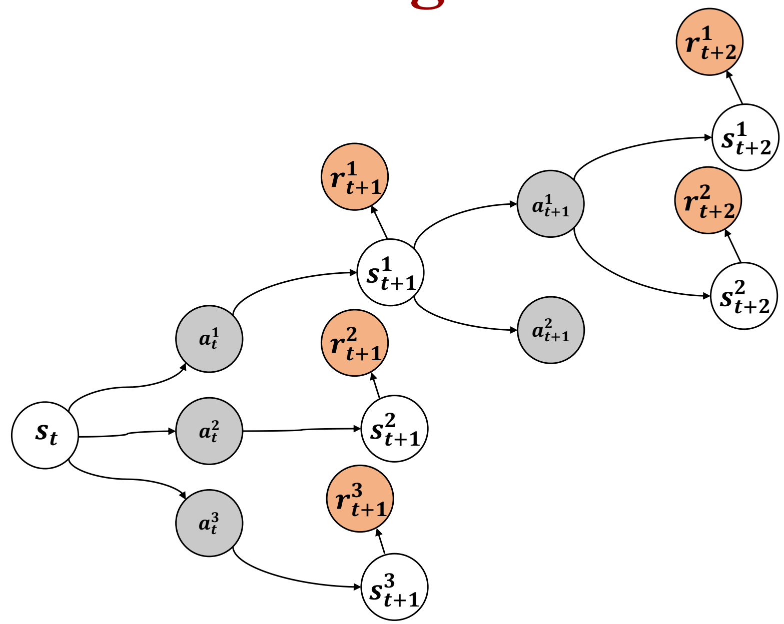
Wavefront Sensorless Adaptive Optics



- Pros:
  - Simple structure
  - Budget-friendly
- Cons:
  - Requirement of good controller
  - Less stubborn in severe condition

Background of Reinforcement Learning





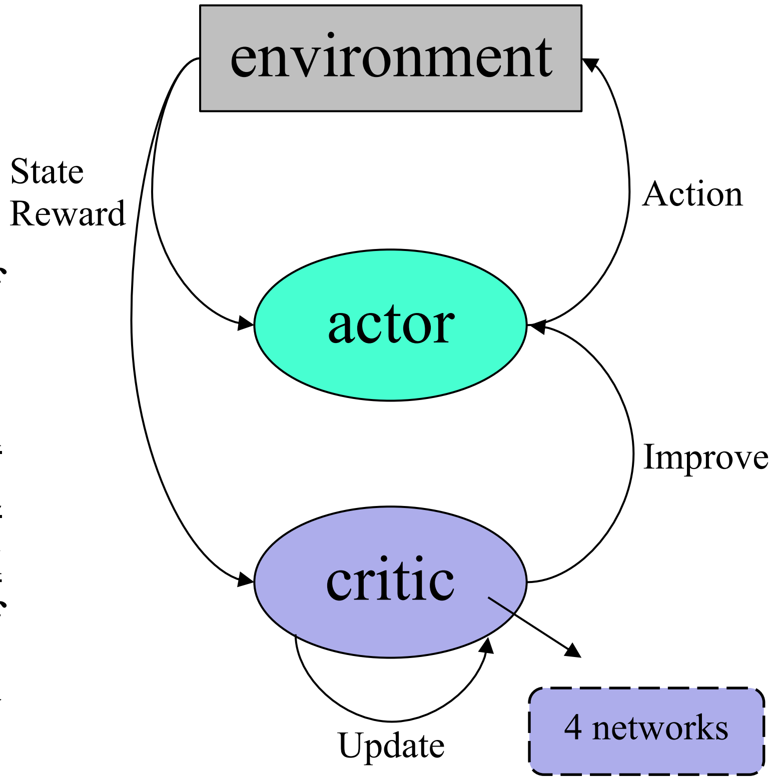
- Widely applied in sequential decision making
- Key components:
  - State
  - Action
  - Reward

- Aim at obtaining a long-term return:

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

The Soft Actor-Critic Controller

- Actor: Generating actions to achieve better performance.
- Critic: Evaluating the value of actions from actor while improving itself to make the assessment more reflective of the real situation.
- There are five networks including an actor network, a soft value network, a target value network, a soft Q network and a soft Q network for the purpose of stabilizing the learning process and preventing overfitting.



The Soft Actor-Critic Controller

Strehl ratio

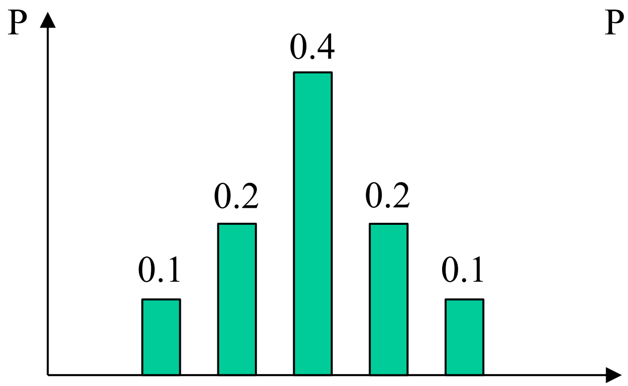
Temperature parameter, the weight of the information entropy

- The optimal policy:

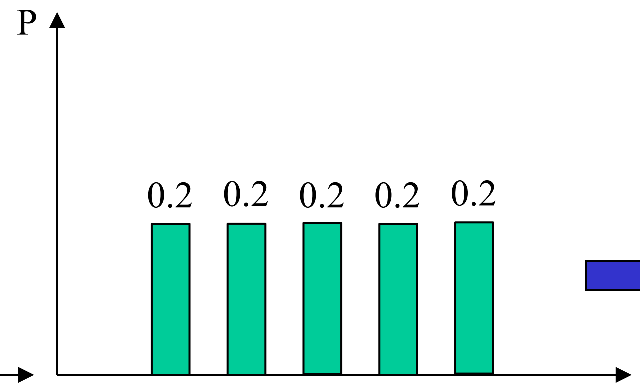
$$\pi^* = \underset{\pi}{\operatorname{argmax}} \sum_t E_{(s_t,a_t)\sim\rho(\pi)}[\boxed{r(s_t,a_t)} + \boxed{\alpha}\boxed{\mathscr{H}(\pi(\cdot\,|\,s_t))}]$$

- Entropy:

$$\mathscr{H}(X) = E[-\log p(X)]$$



Entropy = 0.64



Entropy = 0.70

More even distribution results in higher entropy

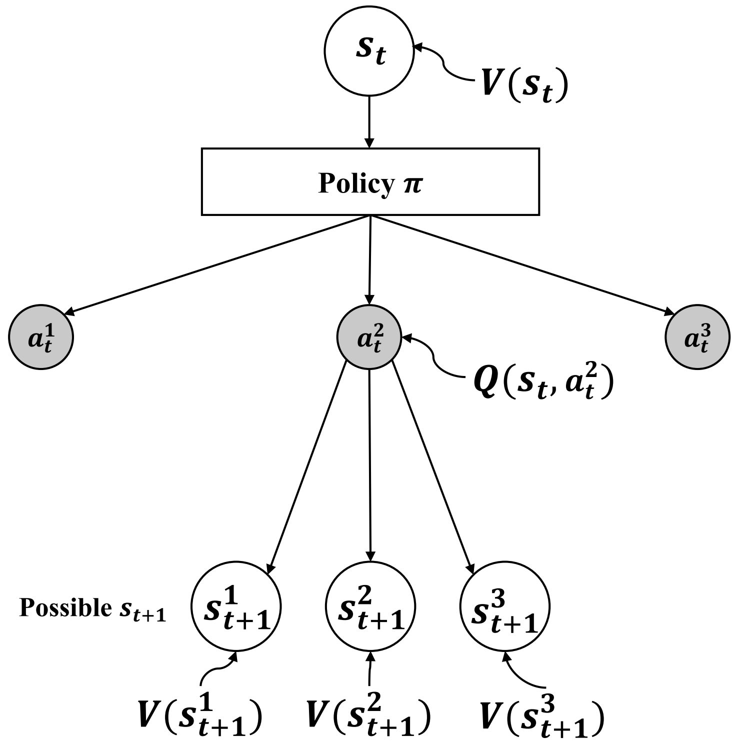
The Soft Actor-Critic Controller

- Policy evaluation:
- State-action value function (Q)
- State value function (V)

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

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

Then the sequence Q will converge to the soft Q-value of π as k → ∞.



The Soft Actor-Critic Controller

- Policy improvement:
- Update the policy to the exponential of Q function for an improved policy
- This particular choice of update can be guaranteed to result in an improved policy in terms of its soft value.
- Kullback-Leibler divergence

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

The Soft Actor-Critic Controller

- Update:

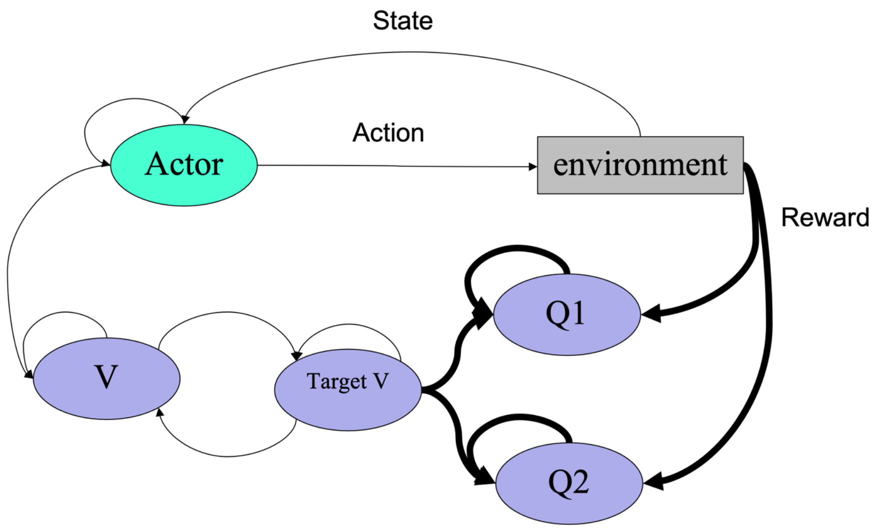
Algorithm 1: Soft Actor-Critic

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

The Soft Actor-Critic Controller

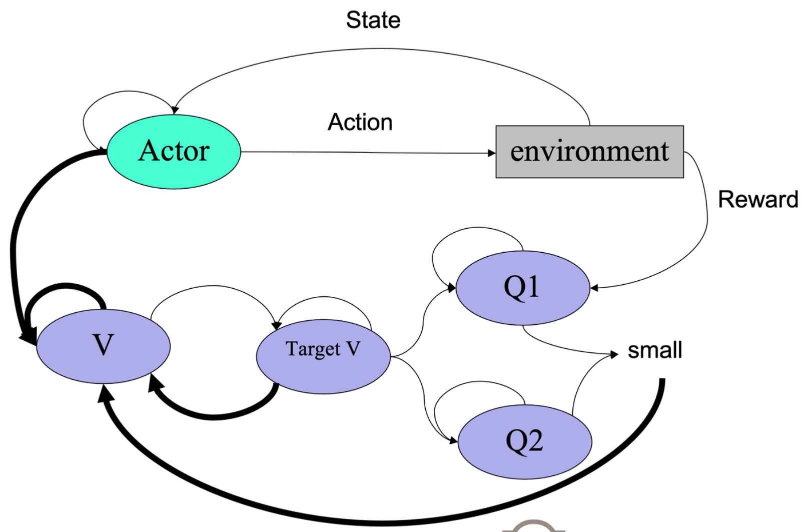
- Improvement of soft Q network:
- Performing gradient descent on the Bellman residual

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



- Improvement of soft value network:
- Performing gradient descent on the square residual error

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



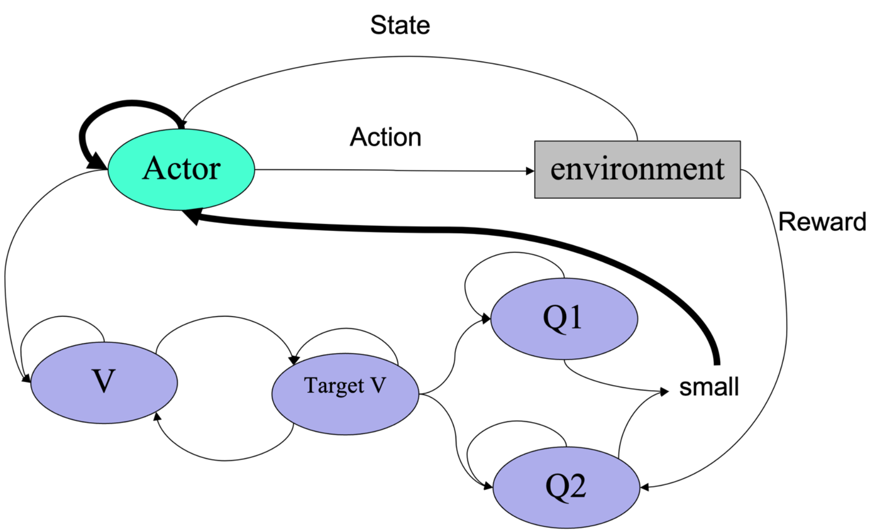
The Soft Actor-Critic Controller

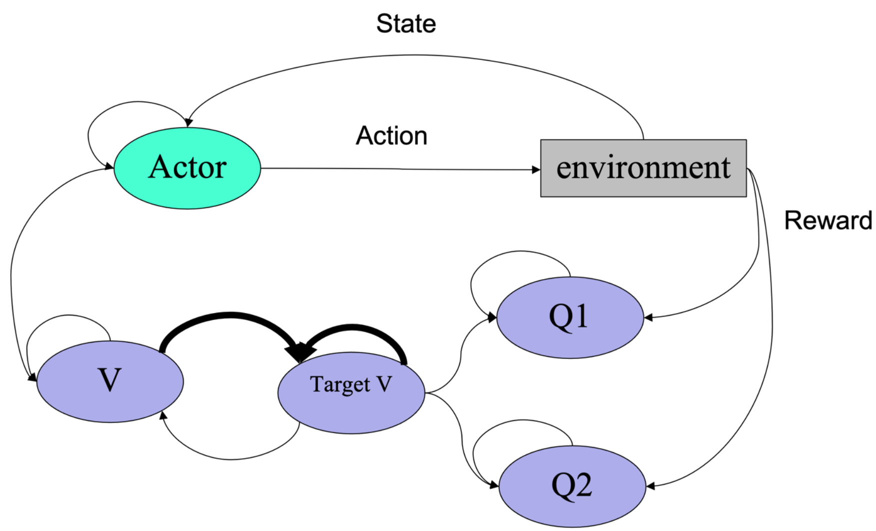
- Improvement of Actor network:
- Gradient descent.

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

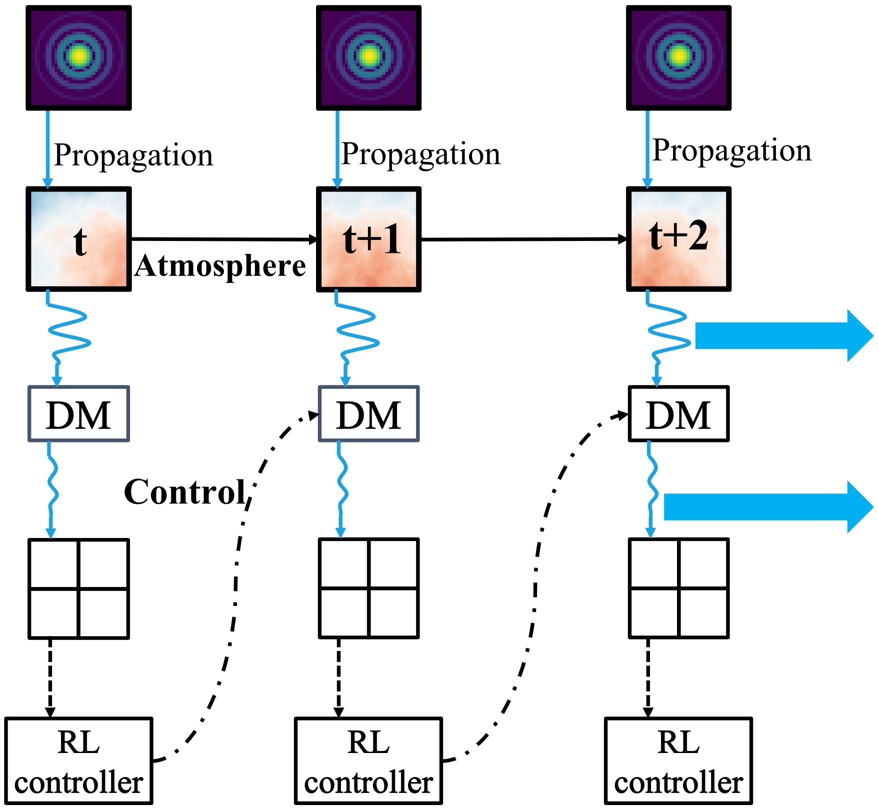
- Target V network

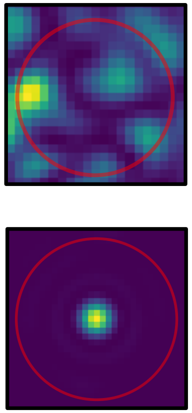
$$\bar{\psi} \leftarrow \tau \psi + (1-\tau)\bar{\psi}$$





Adaptive Optics RL Environment

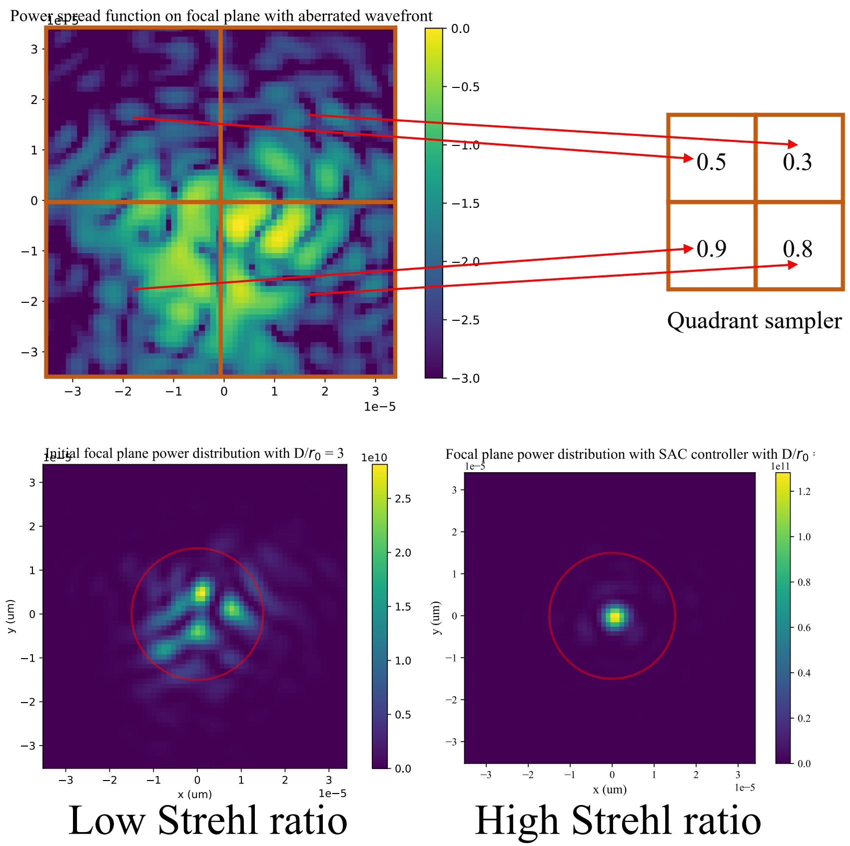




RL controller generates the next time step action while the changing of the atmosphere

Adaptive Optics RL Environment

- Action: Control signal of deformable mirror which manipulates the shape of the surface.
- State: Power distribution observed by quadrant sampler on focal plane with size of 2×2.
- Reward: The reward is built based on summation Strehl ratio on focal plane and the entropy of action distribution on deformable mirror.



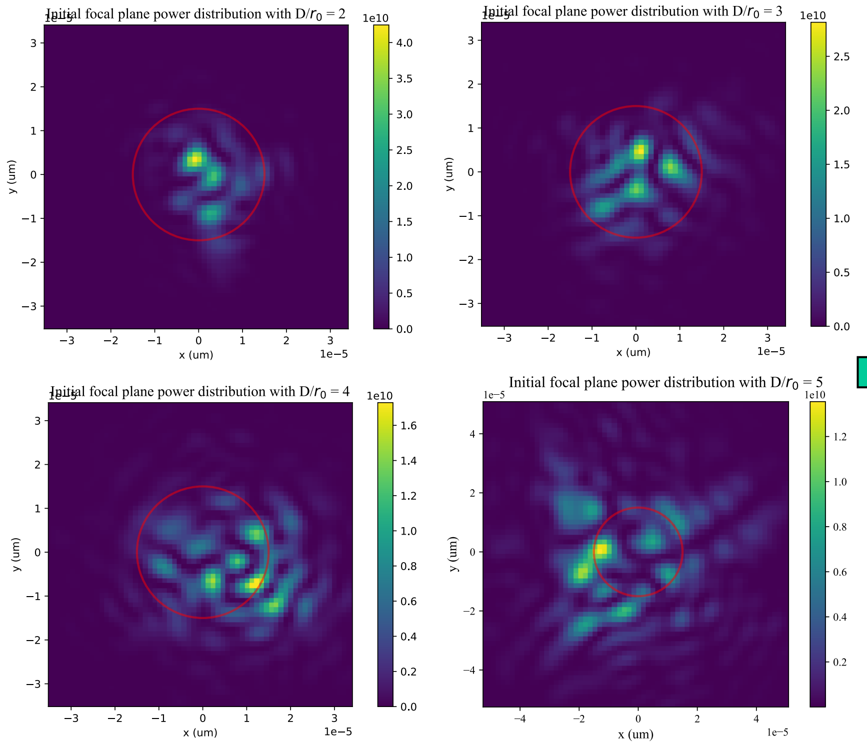
Experiment Setup

- Platform:
  - Compute Canada with CPU and GPU
- Atmosphere is depicted by D/r0
- System configuration:
  - 0.5m telescope
  - 1.5×10^{-6}m wavelength
  - 4×4 deformable mirror
  - 2×2 quadrant sampler
- The result is assessed by the value of Strehl ratio.

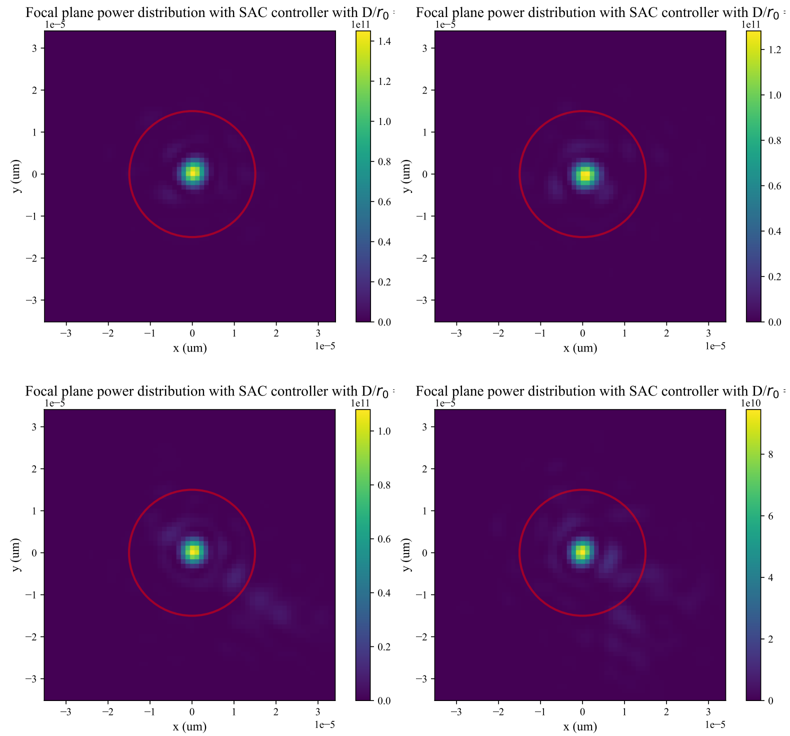
Hyperparameters	Value
actor learning rate	5×10^{-5}
critic learning rate	1×10^{-3}
discount factor	0.95
batch size	256
layer size	128

Static Atmosphere Simulation

- Before training



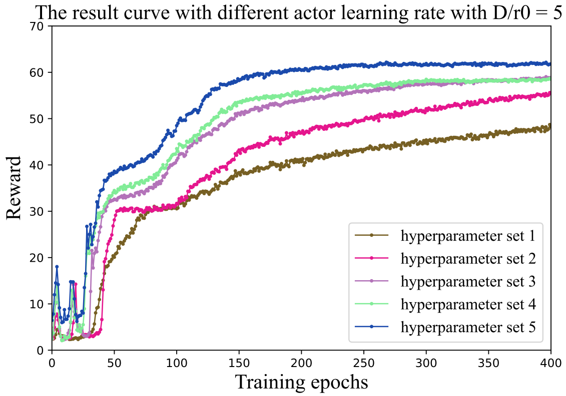
- After training

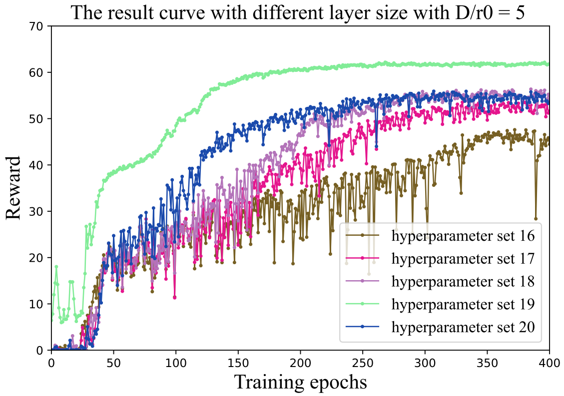


Power is concentrated to the centre of the fiber

Static Atmosphere Simulation

- Hyperparameter tunning of policy learning rate and layer size.
- Conduct by Compute Canada (META) and WandB
- https://github.com/CarlZOUbit/Para_sweep_CC



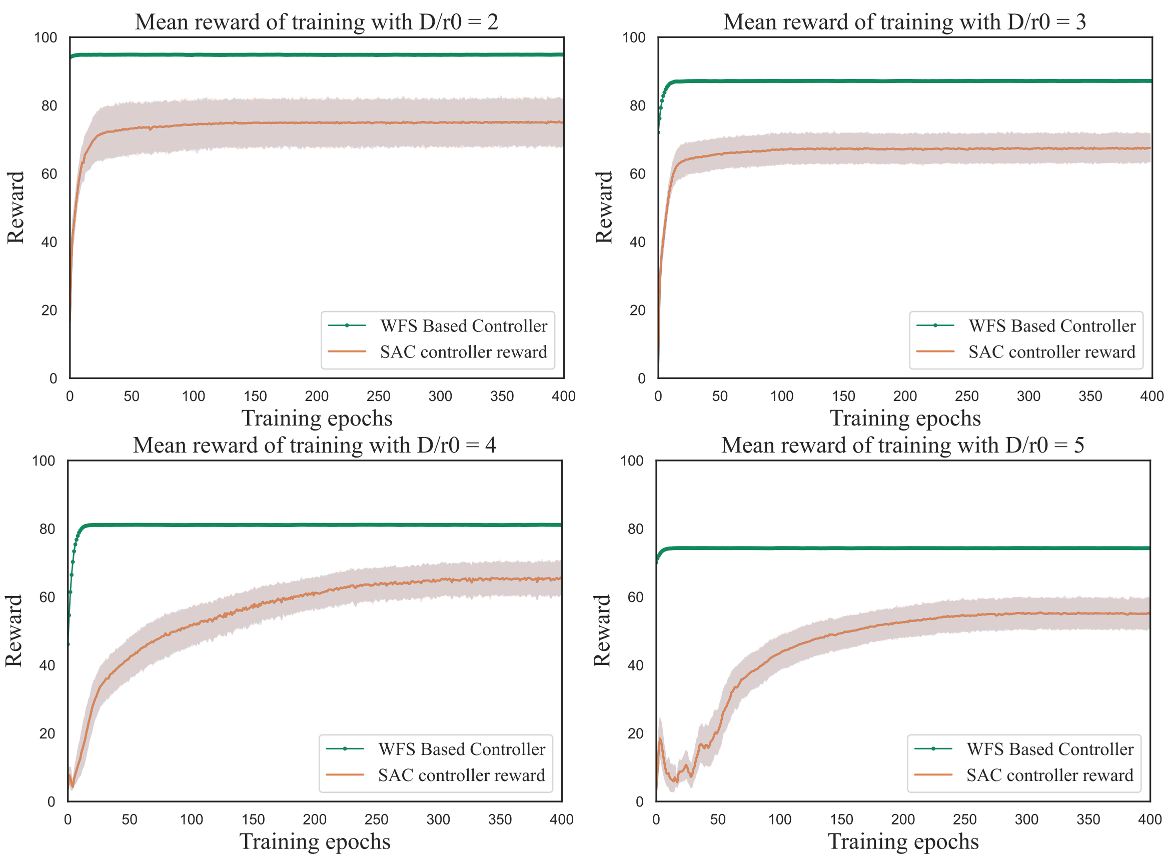


Hyperparameters	Set 1	Set 2	Set 3	Set 4	Set 5
actor learning rate	1×10^{-5}	2×10^{-5}	3×10^{-5}	4×10^{-5}	5×10^{-5}
critic learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
discount factor	0.95	0.95	0.95	0.95	0.95
layer size	256	256	256	256	256
batch size	128	128	128	128	128

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

Static Atmosphere Simulation

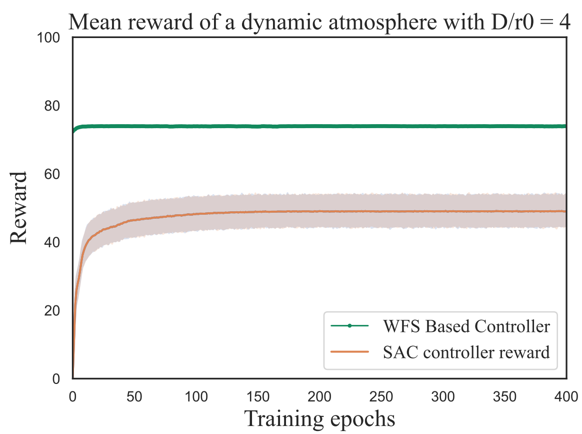
- Reward curve:

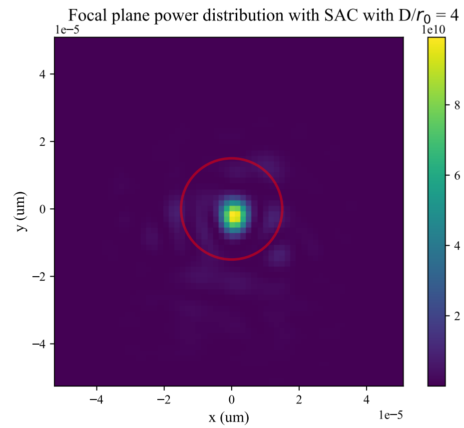


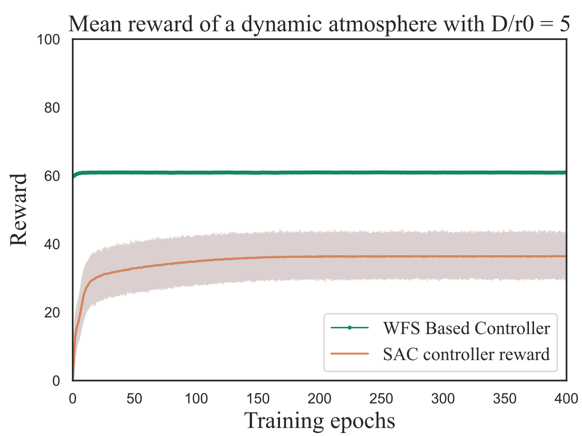
In static condition, relatively high Strehl ratio with low variance

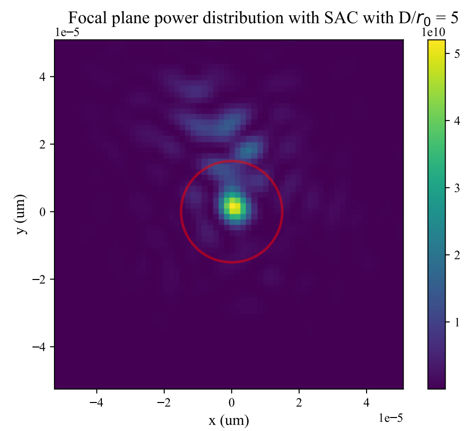
Dynamic Atmosphere Simulation

•









In dynamic condition which is not severe, relatively high Strehl ratio with low variance.

In severe dynamic condition, still could achieve 25%-40% Strehl ratio and concentrate the power to the centre of fibre.

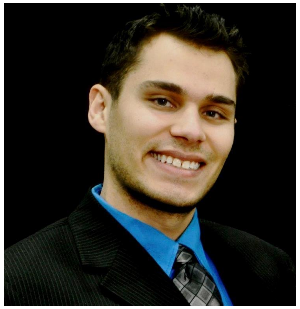
Conclusion & Future work

- A cost-effective wavefront sensorless adaptive optics system by integrating the soft actor-critic algorithm into the adaptive optics controller.
- Relatively high Strehl ratio in simulations of static and dynamic atmospheres

- Future work:
- Improving our algorithm to work with consideration of more factors in real-time environment.
- Improving the code and structure for a real-time application of the controller

Our Team

Dr. Davide Spinello
Associate Professor
University of Ottawa



Dr. Ross Cheriton
Associate Research Officer
National Research Council Canada

Dr. Colin Bellinger
Research Officer
National Research Council Canada

Mr. Payam Parvizi
PhD Candidate
University of Ottawa

Mr. Runnan Zou
Master Student
University of Ottawa