PROCEEDINGS OF SCIENCE

Partial Observability and Domain Randomization in RL-Based Strategy for Optical Cavity Locking Optimization

Andrea Svizzeretto[a,b,*] **and Mateusz Bawaj**[a,b]

[a] *Department of Physics, Università degli Studi di Perugia, Piazza Universtà 1 06123, Perugia, Italy*

[b] *INFN Perugia, Via A. Pascoli 06123, Perugia, Italy*

E-mail: andrea.svizzeretto@dottorandi.unipg.it

The purpose of these proceedings is to report on developments of RL-based cavity locking strategy carried out after the presentation given at the MODE workshop 2025. Reflecting on discussions held during the workshop, we examine how concepts such as partial observability and domain randomization critically impact the training process and the generalization of control policies in deep reinforcement learning. We keep the focus on the topic of our work, concerning reinforcement learning (RL) techniques for lock acquisition optimization in gravitational wave (GW) detection layout. Future directions are outlined to improve robustness and real-world applicability, including domain-randomized parameters and memory based architectures for addressing partial observability.

Fifth MODE Workshop on Differentiable Programming for Experiment Design (MODE2025)
8-13 June 2025
Kolymbari, Crete, Greece

*Speaker

 ISSN 1824-8039 . Published by SISSA Medialab.

PoS(MODE2025)035

1. Introduction

1.1 Context and Challenges

The detection of GW relies on large-scale laser interferometers such as Advanced Virgo Plus and LIGO, designed to measure displacements between suspended mirrors along the arms with a very high sensitivity. These detectors are characterized by a complex optical layout involving multiple Fabry-Perot (FP) cavities, mode cleaners and sophisticated sources of light including quantum squeezed states sources. As a consequence, the complexity behind the layout leads to a large number of parameters that have to be optimized in order to reach higher detection sensitivity. This optimization can be carried out along various design aspects such as: topological, optical (*e.g.* cavity lengths, mirror reflectivities), mechanical (*e.g.* suspension resonances, mirror masses) and control-related (*e.g.* actuator placement, digital filters, locking strategies). Regarding the latter, there is strong interest in optimizing locking strategies, as they play a crucial role in bringing the detector to its working point, *i.e.* the operational state where scientific data can be collected. This is relevant for maintaining high duty cycle during observation and it speeds up commissioning. Because of the complexity of the systems, caused especially by the suspended components, several non linearities can arise. Non linear effects can be produced by mirrors moving and crossing resonances with high speed over a certain threshold. Also they can be caused by radiation pressure or thermal deformations of the mirrors. Since the Pound-Drever-Hall (PDH) technique, the main method used to perform lock acquisition, can be applied only in the linear region of the error signal, there is a need of driving the system from the non linear and unstable regime towards the linear regime through the actuators, so that classic control can be enabled [1].

1.2 Goal and purposes

Deep Reinforcement Learning proved to be very suitable for addressing several non linear control problems in various fields of application, as autonomous driving [2], robotics [3], chemical processes [4] and many others [5]. Moreover, RL technique has been already applied in GW detection context and optical layouts with promising results [6–8]. These previous studies highlight the potential of machine learning (ML) and RL-based approaches in handling complex dynamics, which are common in precision optical systems as the ones in GW experiments. Building upon this foundation, our work explores the use of deep reinforcement learning to optimize the control strategies required to lock optical cavities, addressing both non-linear behaviours intrinsic to these systems, with the final goal of applying this technique to a real set up in the case of suspended free moving mirrors. To address our objective, we employed a Gymnasium-compatible environment [9] powered by a custom-built simulator. With the latter, we are able to reproduce optical signals, cavity power and PDH, affected by ringing effect for a wide range of cavities. Within this framework, we formulate the task as a continuous control problem used to train a Deep Deterministic Policy Gradient (DDPG) agent. In the following paragraphs we describe key ideas for development, particularly concerning partial observability (PO) and domain randomization (DR) and we discuss their potential integration into our experimental framework.

2. Partial observability

In several real-world control scenarios, the agent operates under PO, meaning it does not have direct access to the full underlying state of the environment, but only on observations that are often incomplete, noisy, leading to an harder decision-making process and, as a consequence, a decrease of agent's performance. This limitation is typically introduced by sensor noise, measurement delays or reduced sensing capabilities. This effects are particularly relevant in physical systems where high-precision measurements and complexity are involved.
While fully observable (FO) environments allow the agent to act based only on the current state, PO settings generally require the integration of information over time to infer the latent dynamics. Such conditions are commonly encountered in various domains, including robotics (*e.g.* sensor occlusions or restricted field of view), autonomous navigation (*e.g.* unreliable GPS signals) and, of course, optical systems.
In general, partially observable Markov decision processes (POMDPs) can be described by a tuple of parameters and spaces that together define the interaction between an agent and its environment. Formally, the tuple is composed by $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, R, O, \gamma, \rho \rangle$, where each element has a specific role:

- $\mathcal{S}$: the set of latent states representing the true configuration of the system. For an optical cavity, these might include the instantaneous cavity length, circulating optical power and mirror positions or velocities.
- $\mathcal{A}$: the set of actions available to the agent, such as actuator voltages for changing the positions of the mirrors or other physical inputs aimed at driving the system toward resonance.
- $\mathcal{O}$: the observation space, consisting of the limited information accessible to the agent, *e.g.* the PDH error signal, transmitted and reflected optical power or quadrant photodiode readings. These measurements do not directly reveal the full state of the system.
- $T(s' \mid s, a)$: the transition probability distribution describing the (potentially nonlinear) system dynamics.
- $R(s \mid a)$: the reward function.
- $O(o \mid s)$: the observation function, *i.e.* the probability of observing $o \in \mathcal{O}$ if the underlying state is $s \in \mathcal{S}$.
- $\gamma \in [0, 1]$: discount factor, which weights future rewards with respect to the immediate ones.
- $\rho = P(s_0 = s)$:, the probability distribution over the initial states. Since the true initial state of the system is unknown, at the beginning of the episode it is sampled from a probability distribution over possible initial states. This defines the starting point for the agent's interaction with the environment. In our setting, it may correspond to the probability function used to randomly place the cavity at different positions relative to its nominal length, thereby modeling realistic variations in initial conditions.

In Fig. 1 we show a possible scheme which represents the interaction between the agent and a partially observable environment in the case of high finesse FP cavity locking. At each timestep

t, the agent receives only the observation o_t. Let's consider it to be both, the optical transmitted power measured with a simple photo detector and the error signal, obtained through the PDH technique. The agent has no direct access to the latent state s_t, which is the positions and the velocities of the mirrors as well as the cavity length. In order to be controlled partially observable environment requires information over time, a history of recent environment changes. A useful distinction [10] can be made between the system history—the full sequence of latent states and actions, which determines the reward—and the observation history, which is all the agent can access. At the end of the cycle, the agent must select an action, by applying force $F(t)$ to the mirrors via the actuators, accordingly to the policy $\pi(a_t \mid h_t)$, where $h_t = (o_0, a_0, \ldots, o_{t-1}, a_{t-1}, o_t)$ is the observation history. The value function is formally defined over the system history. However, the agent's decision-making process is constrained to the information contained in the observation history. This mismatch between what determines the return and what is actually observed is a central challenge in solving partially observable Markov decision processs (POMDPs). A possible approach to address this is to reformulate the problem in the *belief-state space*, where the belief state $b_t(s) = P(s_t = s|o_{0:t}, a_{0:t-1})$, the probability function over the latent states given the agent's past observations and actions, acts as a sufficient statistic for the observation history. In practice, memory-based architectures such as recurrent neural networks (RNNs) or long-short term memorys (LSTMs) are commonly employed in RL algorithms to estimate or update the belief using the hidden state when the underlying dynamics or observation models are unknown or intractable. In this transformed setting, the POMDP becomes a sort of fully observable Markov decision process (MDP) in which each state corresponds to an approximated belief distribution over the original latent states.

In the work presented during the MODE workshop *Constrained Charged Particle Tracking with Multi-Agent Reinforcement Learning*, authors [11] deal with PO by using decentralized multi-agent framework, where each one of them can see only a small part of the full tracking problem. To overcome this limitation, the agents are trained together using shared information, even though they act independently during inference. This helps them learn how to cooperate and make better decisions, even with limited local observations. Thinking along similar lines, in our framework each agent could be responsible of acting on the mirror or the correspondent suspension through just one actuator, sharing information from one photo-detector.

Taking PO into account is crucial for realistic control problems, where sensing is indirect, system dynamics are non-linear and robust performance must be maintained in the presence of noise and uncertainty—conditions that can naturally arise in high-finesse optical cavities and interferometric detectors. However, solving POMDPs is computationally challenging. Determining an optimal policy is PSPACE-complete, meaning that the required memory grows polynomially with the problem's size. Furthermore, as already mentioned, optimal policies are generally history-dependent, requiring knowledge of the entire past rather than only the latest observation.

In many practical cases, one instead seeks a near optimal policy, where the resulting value function differs only slightly from the optimal one. Yet, when observations lack enough information to distinguish between certain latent states, learning becomes statistically hard, as identifiability is lost. This difficulty increases sample complexity and computational demands, particularly in overcomplete or undercomplete settings—where the observation space has, respectively, fewer or more elements than the state space.

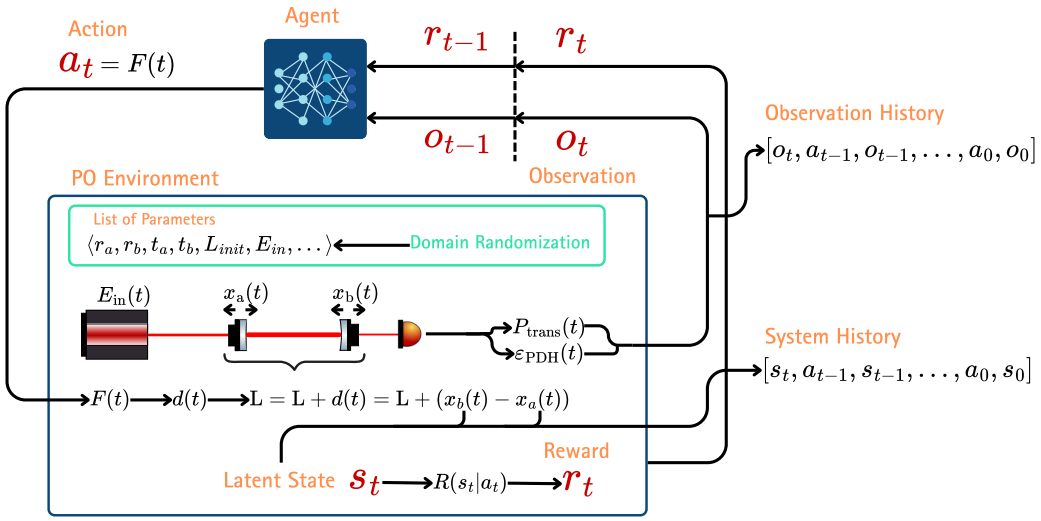

Figure 1: POMDP scheme of an optical cavity application. The environment consists of a rigid optical cavity with movable mirrors, an incident laser and a photo-detector. a_t corresponds to an impulse sent to the actuators, altering mirror positions and thus the cavity length. The photo-detector measures the optical power and, via a running PDH loop (not shown), produces an error signal. These signals form the observation, the only information available to the agent. The true cavity state s_t—mirror positions/velocities or cavity length—is hidden and unobservable. Latent states and observations contribute to the system and observation histories, respectively. In teal green, a list of parameters potentially varied during domain randomization.

3. Domain randomization

Another topic which has been a subject of further thoughts during the MODE workshop is DR, a popular algorithm used to address sim-to-real transfer properly, reducing the number of real-world interactions. Several studies has been conducted in various domains in that sense [12]. The core idea of DR is to expose the agent to a wide range of variations in the environment during training, such as changes in dynamics, sensor noise, object properties, by randomly sampling these parameters within predefined ranges or using some probability distributions. By learning to perform well across different scenarios, the agent becomes less sensitive to discrepancies between the simulated and real-world environments, reducing the so-called reality gap. This gap is also of primary importance in our setting. Our ultimate objective is to deploy an agent on a suspended optical setup, where it will inevitably encounter several dynamics and sources of uncertainty for the first time. Consequently, exposing the agent in a pre-training session to selected aspects of real-world dynamics in advance can be helpful for improving performance during subsequent training in the physical environment. In that sense, the development of a simulator reliable as much as possible becomes strongly needed. Utopically, we may think of a digital-twin of the reality. This is why a large part of our work has been focused on building a custom real-time simulator capable of reproducing the ring-down effect [13, 14] of an high-finesse rigid cavity, based on the field equation in [15]. New improvements of the simulator in the direction of decreasing the reality gap have been implemented recently and further ones will be added to take into account the dynamics of free suspended mirrors. Some challenges that we already partly investigated [13] are related to delays, key aspect while dealing with the inference through AI hardware, ADC/DAC system and actuation

time. Such delays can contribute in making the environment less observable and, as a consequence, the necessity of implying some efficient technique against PO. In fact, according to [12], policies trained with DR can be represented as some memory-augmented structure as RNN or LSTM. This kind of architecture, integrating a history of past observations, allows for better identification of the underlying properties of the system, partly tackling PO issue. Integrating these aspects in the simulation can make a significant difference in reality model-deployment. Nonetheless, DR can easily be a game changer in order to pursue the goal of sim-to-real transfer.
Many possible techniques can be implied to implement DR in simulated environments. Looking at Fig. 1, DR in our setting may consist of randomly varying, at each training episode, key parameters of the optical cavity such as mirror reflectivities, transmittivities, initial cavity length and incident electric field, according to a chosen probability distribution. To mitigate issues coming from overly conservative policies as hindrance in exploration, each packet of the current parameters can be appended to the policy input, applying in that way a parameter-conditioned DR [16]. Another possibility is to analyse the most experimentally relevant non-linear cavities, select a representative ensemble and then train the agent by sampling a different cavity at each episode according to the probability distribution.
On the other hand, borrowing some other possible technique from robotics applications, it seems that enlarging too much the randomization ranges can be counterproductive. Instead, modest parameter variations during the simulation can allow acting on real hardware even without robust details about sensor delays and friction. For sure in suspended setting, where several splice and friction points are present, this strategy can help the agent in generalization.

4. Outlook and conclusion

In this document, we broke down ideas and concepts emerged during and after the MODE workshop 2025. We reported on how partial observability (PO) can occur and be problematic in our framework where the final goal is the deployment of an RL agent capable of achieving an optimal lock acquisition in non linear regime of high finesse optical cavities. Shortly, memory-based architectures will be adopted and integrated within the framework to address this challenge. Further developments on the simulation side will certainly be required to capture the dynamics of the mirrors, bringing us closer to the final goal of the sim-to-real transfer. Moreover, following the trail of the work presented by Tobus Kortus [11], Decentralized Multi Agent Reinforcement Learning can be an optimal technique to address PO in our setting as well. On top of that, we plan to employ domain randomization (DR), being one of the most solid methods to address irreducible reality gaps, which are common in complex experimental designs.

Acknowledgements

Research is supported by Italian Ministry of University and Research (MUR) through the program “Dipartimenti di Eccellenza 2023-2027” (Grant SUPER-C), Italy

References

[1] D. Bersanetti, J. Casanueva Diaz, A. Allocca, H. Heitmann, D. Hoak, M. Mantovani et al., *New algorithm for the guided lock technique for a high-finesse optical cavity*, 2020. https://doi.org/10.1016/j.astropartphys.2019.102405.

[2] B.R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A.A.A. Sallab, S. Yogamani et al., *Deep Reinforcement Learning for Autonomous Driving: A Survey*, 2021. https://doi.org/10.48550/arXiv.2002.00444.

[3] C. Tang, B. Abbatematteo, J. Hu, R. Chandra, R. Martín-Martín and P. Stone, *Deep Reinforcement Learning for Robotics: A Survey of Real-World Successes*, 2024. https://doi.org/10.48550/arXiv.2408.03539.

[4] M. Bloor, A. Ahmed, N. Kotecha, M. Mercangöz, C. Tsay and E.A.D.R. Chanona, *Control-Informed Reinforcement Learning for Chemical Processes*, 2024. https://doi.org/10.48550/arXiv.2408.13566.

[5] Y. Li, *Reinforcement Learning Applications*, 2019. https://doi.org/10.48550/arXiv.1908.06973.

[6] N. Mukund, J. Lough, A. Bisht, H. Wittel, S.L. Nadji, C. Affeldt et al., *First demonstration of neural sensing and control in a kilometer-scale gravitational wave observatory*, 2023. https://doi.org/10.48550/arXiv.2301.06221.

[7] D. Sorokin, A. Ulanov, E. Sazhina and A. Lvovsky, *Interferobot: aligning an optical interferometer by a reinforcement learning agent*, 2021. https://doi.org/10.48550/arXiv.2006.02252.

[8] P.X. Ma and G. Vajente, *A Deep Learning Technique to Control the Non-linear Dynamics of a Gravitational-wave Interferometer*, 2023. https://doi.org/10.48550/arXiv.2302.07921.

[9] M. Towers, A. Kwiatkowski, J. Terry, J.U. Balis, G.D. Cola, T. Deleu et al., *Gymnasium: A Standard Interface for Reinforcement Learning Environments*, Nov., 2024. https://doi.org/10.48550/arXiv.2407.17032.

[10] L.P. Kaelbling, M.L. Littman and A.R. Cassandra, *Planning and acting in partially observable stochastic domains*, 1998. https://doi.org/10.1016/S0004-3702(98)00023-X.

[11] T. Kortus, R. Keidel, N.R. Gauger and J. Kieseler, *Constrained Optimization of Charged Particle Tracking with Multi-Agent Reinforcement Learning*, Jan., 2025. https://doi.org/10.48550/arXiv.2501.05113.

[12] X. Chen, J. Hu, C. Jin, L. Li and L. Wang, *Understanding Domain Randomization for Sim-to-real Transfer*, 2022. https://doi.org/10.48550/arXiv.2110.03239.

[13] A. Svizzeretto, *Optical cavity locking in non linear regime: A reinforcement learning strategy for future applications in gravitational wave detectors*, master's thesis, Università di Perugia, 2024.

[14] D. Bersanetti, *Application of the guided lock technique to advanced virgo's high-finesse cavities using reduced actuation*, 2018. https://doi.org/10.1140/epjp/i2018-11920-5.

[15] M. Rakhmanov, *Dynamics of Laser Interferometric Gravitational Wave Detectors*, phd thesis, California Institute of Technology, Pasadena, California, 2000.

[16] Y. Kawachi, *Quantifying and Visualizing Sim-to-Real Gaps: Physics-Guided Regularization for Reproducibility*, July, 2025. https://doi.org/10.48550/arXiv.2507.23445.