Generalized autonomous optimization for quantum transmitters with deep reinforcement learning

Yuen San Lo[1,*], Robert I. Woodward[1], Taofiq K. Paraiso[1], Rudra P.K. Poudel[1], Andrew J. Shields[1]
[1]*Toshiba Europe Ltd, Cambridge, UK*

ABSTRACT

Precise control of system parameters and extensive optimization play a crucial role in enabling quantum information technologies. As a further challenge, when targeting practical manufacturable systems, the presence of manufacturing variations in components necessitates individual optimization for each system. To address this challenge, we develop a generalisable optimisation framework based on deep reinforcement learning (RL). By applying our method to real-world quantum transmitters based on optical injection locking (OIL), we demonstrate that our RL agent can autonomously identify the optimal operating regions, and generalise its knowledge for new quantum transmitters of the same type. This work presents a new avenue for efficient optimisation of complex systems using modern RL algorithm.

Keywords: Deep reinforcement learning, quantum communication, system optimization, quantum transmitters, optical injection locking, generalization, machine learning, applications of reinforcement learning

1. INTRODUCTION

Quantum key distribution (QKD) allows two users to exchange secret keys with security guaranteed by the fundamental laws of physics, where no computational assumptions are imposed on potential eavesdroppers [1, 2]. Such technology gains significant interest as traditional cryptographic scheme become increasingly susceptible in the face of quantum computing advancements. As the adoption of QKD technology accelerates, there is an increasing need for more robust and reliable systems. Recently, optical injection locking (OIL) has emerged as a promising technique to realise high-rate and robust quantum transmitters [3], where quantum states are prepared and encoded at gigahertz rate before distributing them to a receiver. Although OIL presents several appealing characteristics, its underlying laser dynamics exhibit significant complexity [4] due to the intricate interplay between various control parameters. Achieving stable locking conditions for low-noise and high-coherence output in OIL systems presents a challenge due to the need to simultaneously optimise multiple, interdependent parameters. This complexity is compounded by the intrinsic variations of individual lasers and manufacturing tolerances, which lead to deviations in optimal operating points, even within the same component model. Consequently, optimal parameters identified for one system are typically not directly transferable to another, necessitating the needs for individual system optimisation. It is therefore highly desirable to develop an efficient method that enables automatic and reliable system tuning.

Recently, an autonomous optimisation approach has been developed using genetics algorithms (GAs) to optimise the quantum transmitter based on optical injection locking [5]. It was shown that the method can successfully tune the system to its optimal state without any human intervention. However, a limitation arises: as the properties of every system are slightly different—for example, the variations in the lasing thresholds and emission frequencies of the lasers—the whole optimisation procedure needs to be repeated for every new system, even if it is similar to the previous ones. This is because GAs, at their core, are search-based and lack the ability to learn from past optimisations and apply this knowledge to new but related scenarios. A more efficient solution would involve algorithms that not only adapt during the optimisation process but also generalise this learning to apply it to new systems with similar dynamics. To address this, we have developed a novel framework employing deep reinforcement learning (RL). This framework enables the system to learn from its experiences and apply this knowledge to optimise similar systems, making the process more efficient and generalisable. This is particularly relevant for chip-based systems [6] where a large number of chips with the same design are fabricated.

*yuen.lo@toshiba.eu

2. QUANTUM TRANSMITTER DESIGN

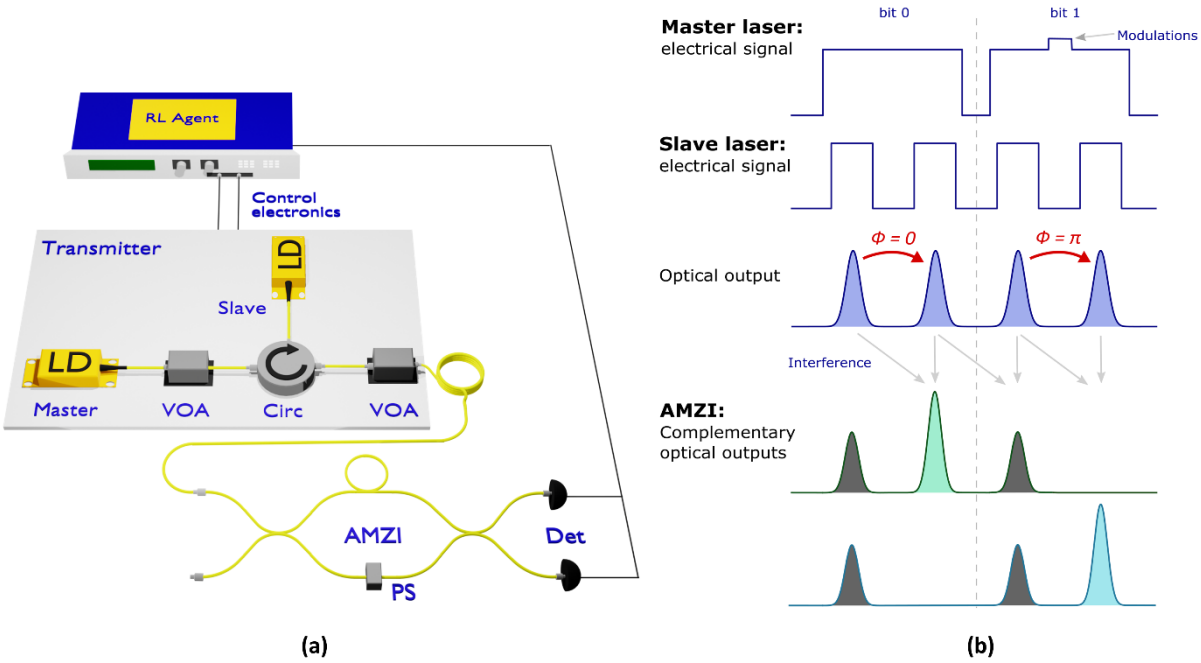

Fig. 1: (a) Experimental setup. VOA variable optical attenuator, Circ circulator, AMZI asymmetric Mach-Zehnder interferometer, PS phase shifter, Det detection. (b) Direct phase encoding scheme of the quantum transmitter.

The experimental setup is shown in Fig. 1a. The transmitter consists of two distributed feedback (DFB) lasers, arranged in an OIL configuration. Light is injected from the 'master' laser into the cavity of the 'slave' laser via an optical circulator. The injection power is controlled by using a variable optical attenuator (VOA). The temperature of each laser is stabilised by an integrated thermoelectric cooler. The RF signals supplied by an arbitrary waveform generator and the DC bias supplied by the current source are combined using a bias-tee to drive each laser. The master laser is gain-switched at 1 GHz, generating a train of long pulses with a duration of ~850 ps. The master pulses are injected into the slave laser which is gain-switched at 2 GHz, generating a train of short pulses with ~70 ps duration. The driving signals of the two lasers are temporarily aligned such that each master laser pulse coherently seeds the generation of two slave laser pulses. This pulse-pair, generated under the same master pulse, represents the early and late time bins for one clock cycle, forming the basis for time-bin encoding.

The relative phase between the two slave laser pulses is used to encode the bit values. Here we perform direct phase modulation where the relative phase can be modulated without the need for external modulators [7,8]: by adding a small perturbation on the driving signal of the master laser, the carrier density will be changed. As a result, the emission frequency will change which in turns changes the phase evolution. Due to optical injection locking, the slave laser inherits the phase of the master laser. As the modulation is located between the two slave pulses, this induced phase change is subsequently transferred onto the relative phase between the two slave pulses, as schematically shown in Fig. 1b. An asymmetric Mach-Zehnder interferometer (AMZI) with a delay line of 500 ps is used to decode the relative phase between the slave laser pulses. The outputs of the AMZI are measured with detectors and analysed to compute the quantum bit error rate (QBER) in a proof-of-concept experiment. A PC is used to control all of the electronics and access the measurement data.

3. GENERALISABLE AUTONOMOUS OPTIMIZATION WITH DEEP RL

3.1 Formalising the optimisation problem in a RL framework

RL is a general formalism that studies optimal decision making in sequential processes. The RL problem is formulated in terms of an agent interacting with an environment in discrete time steps. At each time step t, the agent observes the current state s_t of the environment and selects an action a_t. Following this action, the environment provides an immediate reward r_t and transitions to a new state s_{t+1}. The agent's goal is to maximize the cumulative rewards it receives over the course of its interactions with the environment. To achieve this, the agent learns through trial and error, collecting information from the environment and determining the best action to take in response to each observation.

To apply RL in solving our optimisation problem, first we frame the task as a pathfinding problem, where the agent needs to learn to find the best path moving from a starting point to an end point in a grid world. The objective is for the agent to learn a policy that maximizes the total reward. With this framework, we can consider the laser optimisation landscape as the discrete grid world where the agent needs to locate the optimum operating position, navigated by the system performance as a form of rewards. The metric that quantifies the system performance of the quantum transmitter is the QBER, which depends on the driving parameters of the transmitter. To generalise the optimisation, we train the RL agent with various similar environments that share the same underlying mechanism. As such, our agent will be exposed to different environment during its training, and will thus be able to derive a generalised policy to navigate in this type of environment, to autonomously minimize the QBER.

In the conventional pathfinding settings [9], the dimensions of the grid world naturally translate to the optimisation parameters of interest, and the current parameter values are represented by the coordinate of the agent in the grid world (which can be multidimensional). With these settings, however, the RL agent inevitably fails. This is because the observation of the agent is tied to the coordinates in the environment. During training, the agent tries to find the best route leading to the destination in terms of a set of coordinates in the environment. When the environment has changed, the agent simply has no way to distinguish which environment it is currently in. To overcome this problem, we reconstruct the representation of the agent's observation. Instead of using its own coordinate in the grid, we use its surrounding terrain in the parameter space as its observation [10], as shown in Fig 2. The settings are summarised as below:

1. Environment = Parameter space of optical injection locking dynamics
2. Observation = $\sum_i \sum_j f\,(x+i, y+j)$ where (x, y) is the coordinate of the agent and i, j iterates over the coordinates of the local terrain surrounding the agent in the environment.
3. Action = {+1, +0, -1}
4. Rewards, r = $-QBER + 100s$, s = 1 if $QBER \leq QBER_{optimum}$ else s = 0

3.2 Experimental results

To verify our approach, we use three sets of OIL systems to train the agent to locate optimum region and use one other set for validation. During training, in each episode, one of the laser sets is selected randomly for the agent to control. We optimise for four parameters: master laser DC bias current, slave laser DC bias current, injection power and slave laser wavelength (as set by the temperature tuning). In this work, we use the soft actor-critic (SAC) algorithm [11], which is an off-policy actor-critic deep RL algorithm, to learn the policy. SAC-based policy is trained to maximize a trade-off between expected reward and entropy, a measure of randomness in the learned policy. This is implemented by using the Stable Baselines3 library, a widely-used toolkit for reinforcement learning algorithms in Python [12]. Fig. 3 shows the success rates in tuning the three sets of lasers during training. A "success" episode is achieved when the system is tuned to a QBER below 3.5%. The variation in these curves shows that the agent is trying to derive an optimal policy that can simultaneously optimise these three sets of lasers. For example, after 2000 episodes, the agent becomes very good at tuning set 1, but struggle at tuning set 3. After 4000 episodes, the agent is able to find a way to optimise all of them, suggesting generalisation in its optimisation strategy.

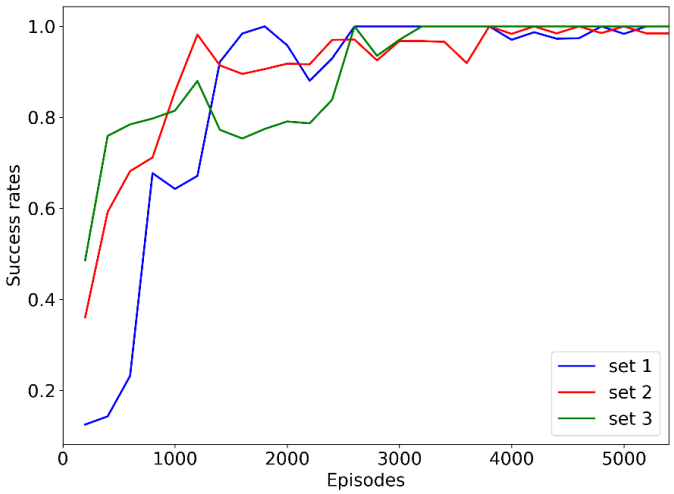

Figure 3: Learning curves during training. Each curve shows the success rate of the corresponding laser set.

After training, the agent is applied to the validation set. Fig. 4 shows 8 trials of optimizations performed by the agent, showing that starting from a random state, the agent is capable of tuning the laser to a QBER below 3.5% (our chosen threshold for success, determined as a reasonably low QBER for good QKD system operation) with a high success rate, despite the fact that the agent has not encountered this laser set before. Moreover, it can locate the optimal state within ~10 steps (less than 5 min). Note that here the QBER is measured with an oscilloscope to facilitate fast data acquisition, although similar results would be expected using single-photon detectors with time-tagging electronics. An interesting observation here is that sometimes the QBER temporarily increases, then reduces again to the optimum value. This underscores a key aspect of reinforcement learning: an agent does not merely follow the local landscape in the parameter space, akin to what we see in gradient descent. Instead, the agent predicts the long-term outcomes, which in this case means identifying and navigating towards the location of optimal region with much bigger rewards, sometimes this means choosing paths that are initially less rewarding but potentially lead to greater overall rewards.

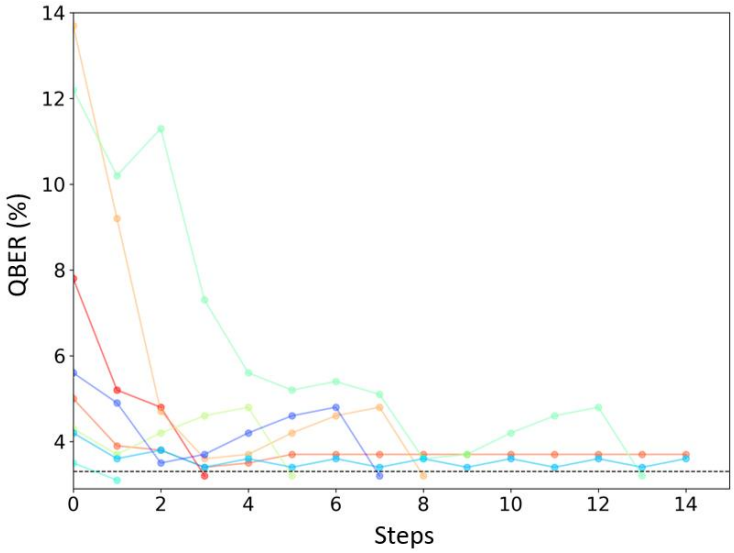

Fig. 4: QBER optimizations performed by the RL agent on the validation laser set. Each line represents a different experimental run of the optimization.

4. CONCLUSION

In conclusion, we have introduced a novel method for multiparameter optimization in complex systems, specifically focusing on the QKD transmitter based on OIL. By leveraging advancements in deep RL, the results show that our model can generalize the optimization across different systems of the same type, efficiently locating their optimal operating regions. Our work presents a new avenue for efficient optimization of complex systems, offering a pathway towards low-cost and scalable production processes and more widely available quantum technologies.

REFERENCES

[1] N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden. “Quantum cryptography”. Reviews of Modern Physics 74, 145 (2002)

[2] C. H. Bennett and G. Brassard. “Quantum cryptography: Public key distribution and coin tossing”. In IEEE Int. Conf. Comput. Syst. Signal Process, 175 (IEEE, 1984)

[3] T. K. Paraïso, R. I. Woodward, D. G. Marangon, V. Lovic, Z. Yuan, and A. J. Shields. “Advanced Laser Technology for Quantum Communications (Tutorial Review)”. Advanced Quantum Technologies 4, 2100062 (2021)

[4] E. K. Lau, L. J. Wong, and M. C. Wu. “Enhanced Modulation Characteristics of Optical Injection-Locked Lasers: A Tutorial”. IEEE J. Sel. Top. Quantum Electron 15, 618 (2009).

[5] Y. S. Lo, R.I. Woodward, T. Roger, V. Lovic, T.K. Paraïso, I. De Marco, Z.L. Yuan, A.J. Shields. “Self-Tuning Transmitter for Quantum Key Distribution Using Machine Intelligence”. Phys. Rev. Applied 18, 034087 (2022).

[6] Paraïso, T.K., Roger, T., Marangon, D.G. *et al.* A photonic integrated quantum secure communication system. Nat. Photon. **15**, 850–856 (2021).

[7] Z. L. Yuan, B. Fröhlich, M. Lucamarini, G. L. Roberts, J. F. Dynes, and A. J. Shields. “Directly Phase-Modulated Light Source”. Physical Review X 6, 031044 (2016).

[8] Y. S. Lo, R. I. Woodward, N. Walk, M. Lucamarini, I. De Marco, T. K. Paraïso, M. Pittaluga, T. Roger, M. Sanzaro, Z. L. Yuan, A. J. Shields. “Simplified intensity- and phase-modulated transmitter for modulator-free decoy-state quantum key distribution”. APL Photonics 8, 036111 (2023).

[9] B. D. Pena and D. T. Banuti, "Reinforcement learning for pathfinding with restricted observation space in variable complexity environments", AIAAScitech 2021 Forum, 2021.

[10] Gym Documentation. Frozen Lake. Gym Library. Retrieved 12 Jan 2024, from https://www.gymlibrary.dev/environments/toy_text/frozen_lake/

[11] T. Haarnoja *et al*, “Soft Actor-Critic Algorithms and Applications” arXiv:1812.05905 [cs.LG], 2018.

[12] Raffin, A. *et al*. “Stable-baselines3: reliable reinforcement learning implementations”. J. Mach. Learn. Res. 22, 1–8 (2021).