AUTOMATIC RECALIBRATION OF QUANTUM DEVICES BY REINFORCING LEARNING

T. Crosta*[1], L. Rebón[2], F. Vilariño[1,3], J. M. Matera[4] and M. Bilkis[1]

[1]*Computer Vision Center (CVC), 08193 Bellaterra (Cerdanyola del Vallès), Spain*
[2]*Instituto de Física La Plata (IFLP), CONICET - UNLP, and Departamento de Ciencias Básicas, Facultad de Ingeniería, Universidad Nacional de La Plata (UNLP), La Plata 1900, Argentina*

[3]*Department of Computer Science, Universitat Autónoma de Barcelona (UAB), 8193 Bellaterra (Cerdanyola del Vallès), Spain.*
[4]*IFLP-CONICET, Departamento de Física, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, C.C. 67, La Plata 1900, Argentina*

Recibido: 03/05/2025 ; Aceptado: 09/09/2025

During their operation, due to shifts in environmental conditions, devices undergo various forms of detuning from their optimal settings. Typically, this is addressed through control loops, which monitor variables and the device performance, to maintain settings at their optimal values. Quantum devices are particularly challenging since their functionality relies on precisely tuning their parameters. At the same time, the detailed modeling of the environmental behavior is often computationally unaffordable, while a direct measure of the parameters defining the system state is costly and introduces extra noise in the mechanism. In this study, we investigate the application of reinforcement learning techniques to develop a model-free control loop for continuous recalibration of quantum device parameters. Furthermore, we explore the advantages of incorporating minimal environmental noise models. As an example, the application to numerical simulations of a Kennedy receiver-based long-distance quantum communication protocol is presented.

Keywords: Quantum Machine Learning, Quantum Control, Automatic re-calibration, Kenedy receiver

https://doi.org/10.31527/analesafa.2025.36.4.95-105 ISSN - 1850-1168 (online)

I. INTRODUCTION

Calibrating an experimental apparatus is a primitive and ubiquitous task in most areas of science and technology. In turn, sensor and detector devices constitute the way to extract information about the environment surrounding us and better understand reality via further post-processing of the acquired data. Thus, fully calibrating experimental devices is a primordial task and, in turn, an active research topic [1-11]. In this manuscript, we study the recurrent calibration of devices whose deployment environment is challenging to be modelled. Examples of this are scenarios that heavily vary with time in a way that is hard to predict, *e.g.* turbulent atmosphere [12-17], hydrological models [18, 19] or non-isolated magnetometers [20, 21] among many others. For such settings, where state-of-the art technology is being used to push forward the boundaries of scientific discoveries at a considerable resource overhead, it is of utmost importance to develop techniques that are ready to adapt the device configuration to the experimental condition at hand. In this regard, a plethora of artificial-intelligence techniques have recently been developed in the context of sensor calibration [1, 2, 8, 22-29], change-point detection [30-32] and malfunctioning device identification [33-38]. Our main contribution is to provide a framework for re-calibrating quantum devices. Based on this, we present the method applied to a quantum-classical long-distance communication by laser pulses, *e.g.* satellite-ground or optical-fiber communication. Overall, the success of most machine-learning (re)calibration schemes considered in literature rely either on perfect knowledge of device's functioning condition, access to huge amount of data for training purposes or limiting the dynamics of the system. Such assumptions constitute a double-edged sword when deploying the device on (potentially adversary) experimental conditions: while correct configurations can be granted if the machine-learning model was trained on data resembling the experimental conditions, there is a high probability of remaining off-calibrated otherwise.

Here, we depart from such a notion of similarity between training and deployment scenarios, by considering a hybrid scheme consisting of a pre-training round complemented with a reinforcement-learning stage. The latter fine-tunes the configuration, so the device can be adapted to the specific (and potentially unexplored) experimental conditions at hand; this is done by modifying device controls, as shown in Fig. 1.

The success of our method hinges on the capabilities of devising an approximate model of the setting's dependence with respect to changes in its surroundings (which we indistinctly call environment). Such approximate model is to be thought as a simplified description of the environment, *e.g.* captured by very few variables. While not expected to be fully accurate — not retrieving the exact device configuration for each specific experimental condition—, it shall be thought of as *ansatz* for controls initialization. The accuracy of this initialization relies on the capacity of the approximated model to capture relevant features of device behavior given the experimental condition at hand. As

[*] tomycrosta@gmail.com

a rule of thumb, the more complex such ansatz, the more accurate the description is expected to be. However, a trade-off rises; While more complex models tend to be data-consuming until reaching optimal calibrations, tailoring the model to specific experimental conditions will inherently induce a bias towards a sub-set of deployment scenarios. Thus, the goal of the pre-training round is to suggest control initialization values by using a small number of quantities that can easily be estimated out of few experiments. The control values are then improved by means of a complementary reinforcement-learning method, which adapts the control values to the specific experimental condition in a model-free way. On top of the calibration mechanism, the value of a *decalibration witness* is continuously monitored during deployment, which allows the *agent* to experimentally detect that the device entered an off-calibration stage, and thus re-initiate the calibration process. This work is a step further towards developing a fully automatic re-calibration of quantum detectors through machine-learning techniques. Importantly, we remark that neither the framework nor the method is specially biased towards the quantum realm, and can potentially be applied to other control problems beyond the quantum-technology scope.

The manuscript is structured as follows. In Sec. II we present our re-calibration framework and described our method. In Sec. III we numerically analyze the performance of our re-calibration method in an emblematic long-distance quantum communication setting. Conclusions and future work are outlined in Sec.IV

II. The re-calibration framework

We consider a device whose controls are defined by continuous parameters $\boldsymbol{\theta} = \{\theta_1, ..., \theta_M\}$. As shown in Fig.1, our setting is a black-box device controlled by different knobs $k = 1, ..., M$, each associated to a control value θ_k. In the following, we define several quantities of interest.

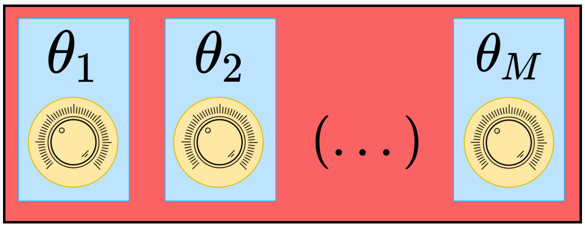

FIG. 1: We depict a device that needs to be calibrated. Here, the apparatus is controlled by different knobs defined by values $\boldsymbol{\theta} = \{\theta_1, ..., \theta_M\}$, and the aim is to tune such parameters in a way that the device is configured to optimally operate under experimental conditions $\mathscr{E}$.

Device configuration. A fixed set of parameter values $\boldsymbol{\theta}$ completely defines a device configuration.

Score function. The quality of a device configuration $\boldsymbol{\theta}$ is evaluated by a score function $S_{\mathscr{E}}(\boldsymbol{\theta})$. The value of the score function can be estimated—during a calibration stage—by means of N repeated experiments; each experiment i involves a quantum measurement and leads to a measurement outcome $\boldsymbol{n}_i$, whose value is generally of stochastic nature. Here, the full underlying model needed to describe outcomes probability distributions is denoted by $\mathscr{E}$, and generally involves an accurate description of noisy channels present in the setting at hand.

Effective score function. The underlying model $\mathscr{E}$ is generally inaccessible to the calibrating agent, and hence shall assume to be unknown to it. This is motivated by the fact that: *(i)* time-varying deployment conditions can be fundamentally hard to model, and *(ii)* even in the case of having full control of experimental conditions, quantum channel-tomography comes with a considerable sample overhead, implying that the total number of experiments and parameters required to reach near-optimal environment modelling (plus device calibration) would grow exponentially or be otherwise constrained to specific scenarios [39-42]. On the contrary, we do assume that a certain relationship exists between the *true* score function $S_{\mathscr{E}}(\boldsymbol{\theta})$ and its *effective version* $S_{\tilde{\mathscr{E}}}(\boldsymbol{\theta})$, *e.g.* an effective model $\tilde{\mathscr{E}}$ used by the agent is indeed able to capture certain relevant features of the score function. Effective models should be thought as an enhanced *control initialization* strategy. This notion applies for the case in which the device enters an off-calibration stage, and new control values should be found.

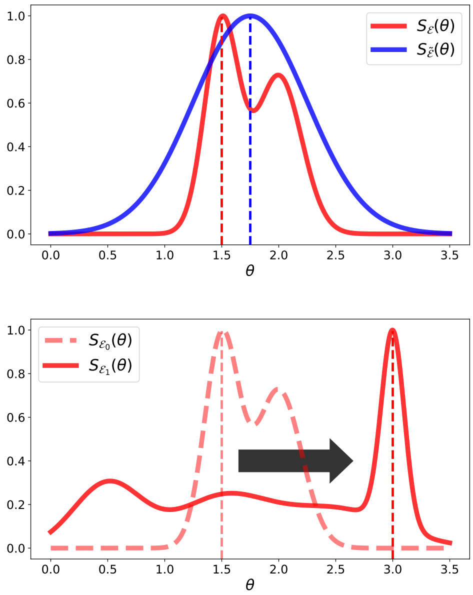

FIG. 2: Single-parameter device example. Top panel: the optimal calibration score $S_{\mathscr{E}}(\boldsymbol{\theta})$ is shown (dashed-red vertical line), and its effective value $S_{\tilde{\mathscr{E}}}(\boldsymbol{\theta})$ (blue-dashed vertical line); while suboptimal, this value is further fine-tuned by means of a model-free scheme (see main body). Bottom panel. We show score functions $S_{\mathscr{E}_0}$ and $S_{\mathscr{E}_1}$ before and after a change-point occurs in the environment. As a consequence, the device optimally configured under $\mathscr{E}_0$ needs now to be re-calibrated to the new optimal configuration for $\mathscr{E}_1$.

Reinforcement Learning (RL). The setting described above can be framed in the RL language [10, 22, 43-46]. , where an agent repeatedly interacts with an environment in order to maximize a reward function, during different episodes. Here, at i^{th} episode (experiment), the agent selects parameter values $\boldsymbol{\theta}$, observes measurement outcomes $\boldsymbol{n}_i$, and finally post-processes them in order to provide a *claim* for the underlying task the quantum device is used for. Based on the accuracy of this final action, the agent is given a reward signal, which uses to improve its estimate on how valuable the decisions performed were. In RL, this is captured by the so-called *state-action value-function* $Q_\pi(s,a)$ [10], standing for the expected reward when departing from state s and taking action a (*i.e.* either selecting parameters $\boldsymbol{\theta}$ or providing a claim based on the outcomes acquired [10, 43]), and following decision criteria — or policy— π. For an optimal device usage, the agent shall choose configurations $\boldsymbol{\theta}^*$ leading to a maximum score $S_{\mathscr{E}}(\boldsymbol{\theta})$. Nonetheless, since $\mathscr{E}$ is not available to the agent, value functions need to be estimated out of several experiment repetitions. Importantly, the agent's strategy is optimized solely based on the rewards acquired during learning. Here, not only such rewards are a way to estimate value-functions, but also serve as a lighthouse for the agent to navigate the decision landscape, allowing a *model-free* calibration of the device. We provide further details on how model-free calibration works in Appendix A.

As an example, we consider a single-control device, whose score function $S_{\mathscr{E}}(\boldsymbol{\theta})$ is schematized in Fig. 2. A model-free agent would initially set the parameter θ at random and consequently estimate its score function out of repeated experiments. On the contrary, keeping an effective model $\tilde{\mathscr{E}}$ can readily help the agent to improve such initialization strategy. Here, the agent's internal model $S_{\tilde{\mathscr{E}}}(\boldsymbol{\theta})$ serves as an ansatz for the underlying behavior of score $S_{\mathscr{E}}(\boldsymbol{\theta})$ w.r.t. the control θ. Intuitively, the internal model $\tilde{\mathscr{E}}$ is expected to be easier to estimate out of few experiment repetitions.

RL methods have recently been applied to a wide variety of quantum technology scenarios, among them calibrating a quantum communication setting [10, 11, 47-51], optimizing quantum pulses [22, 52-54], quantum gated-circuit layout [55, 56], and even graph-processing applications [57], to name a few. However, little has been investigated in the capabilities of the learning model to *adapt* the calibration to changes in the environment $\mathscr{E}$ happening while the device is being used, *e.g.* in the *deployment stage* [58, 59]. In Fig. 2 (bottom) we exemplify how a change in the environment would affect the score function, requiring a recalibration. In order to detect the new landscape, we can consider that the new observations will be different from the ones predicted by the previous exploration. Thus, having indications about changes in the environment even during off calibration stages.

Decalibration witness. In order to realize that a change occurred in the environment, the agent must rely on an experimentally accessible quantity, which we define as decalibration witness and denote with $\mathscr{W}_d$. By monitoring the behavior of $\mathscr{W}_d$ over different experiments, the agent can readily detect whether a change-point occurs in the environment and thus re-start the calibration routine if anomalies are detected. Examples of potential decalibration witnesses are estimates of outcome probabilities, which the agent can straightforwardly construct from the information acquired during previous experiments.

Automatic re-calibration. The definitions outlined above set a framework to analyze the recurrent calibration of a device. We now turn to describe our automatic re-calibration method, which makes use of effective score functions, RL routines and decalibration witnesses. Here, we picture a scenario where the device is to be initially calibrated and, while it is being deployed, the device enters an off-calibration stage which needs to be compensated. The quality of a given configuration is measured by a score function — which in turn depends on the current experimental conditions—; the maximum of the latter encodes the solution to the problem for which the device is being used for.

For instance, in a communication setting, the device configuration is defined by the encoding-decoding strategy (*e.g.* the quantum measurement performed to decode information out of the incoming signal), and the score function is given by the success probability of the protocol. Alternatively, in variational quantum computing applications [60, 61], *i.e.* the VQE algorithm [62], the device configuration is defined by the free parameters of the parametrized quantum circuit and the score function is given by the energy landscape, which needs to be estimated out of several repetitions of an experiment. The initially-optimal configuration can be attained by model-free RL schemes [10, 52], *i.e.* trial-and-error learning mechanisms. In this approach, the score function is typically estimated from the rewards acquired from each device configuration, e.g. by an empirical estimation of its value functions (see Appendix A). In this work, we depart from this concept by initializing the value function estimates to a surrogate quantity, defined by the effective score function $S_{\tilde{\mathscr{E}}}(\boldsymbol{\theta})$; such quantity is to be estimated out of a few experiment repetitions, and serves as an ansatz for which score value is assigned to a given device configuration (see Fig. 1 for a schematic representation).

The usage of effective score values is motivated by the fact that experimental conditions might not dramatically differ from the ideal case. For example, the VQE energy landscape shall preserve certain similarities between a noiseless scenario and a noisy one, assuming the noise strength is sufficiently low [63]. From this informed initialization of value-function estimates, we then exploit the model-free features of traditional RL algorithms, which allows the agent to fine-tune the device configuration, adapting it to specific deployment conditions.

The mechanism described above constitutes the *calibration stage*, in which the actions performed by the calibrating agents can be rewarded according to their accuracy/correctness. With the initial calibration task accomplished, the device is then deployed, *e.g.* used without the necessity of rewarding the agent. As experiments proceed, it is to be expected that the device undergoes a decalibration, *e.g.* experimental conditions might eventually vary. In order to detect such a change occurs, the agent controls the decalibration witness $\mathscr{W}_d$ —for example measurement outcome probabilities —, which is used by the specific change-point detection protocol the agent keeps. Thus, by monitoring $\mathscr{W}_d$, the agent can detect that the device entered into a decalibration stage, *e.g.* the deployment conditions have changed. As a consequence, the new optimal configuration is a different one, and a re-calibration is carried out. This is done similarly to the initial calibration stage: the effective model configuration landscape is estimated out of few experiments, and the model-free RL algorithm is then used to adjust the configuration to the new optimal one.

The effective model used by the agent $S_{\tilde{\mathscr{E}}}(\boldsymbol{\theta})$, along with the decalibration witness $\mathscr{W}_d$ and the RL algorithm (*e.g.* search strategy, value functions, reward definitions), define a re-calibration strategy. However, each of the strategy components requires the agent to pre-set a number of *hyperparameters*. Among them, the number of experiment repetitions needed to estimate effective-model configuration landscape (which we denote as N_{eff}), the number of experiments needed to fine-tune the configuration using a RL method, denoted as N_{rl}, the undecision region for which values that take $\mathscr{W}_d$ will not lead the agent to re-activate the calibration routine, and the parameters defining the behavior of the RL routine, whose nature depends on the particular algorithm used. In order to help with notation, we will comprise all such parameters by $\boldsymbol{\xi}$; in Algorithm 1 a pseudocode of our re-calibration method is provided. In the following, we showcase the re-calibration method introduced above in a canonical example for long-distance classical-quantum communication. We stress that our method can be applied to a wide variety of scenarios, not necessarily constrained to the quantum technology realm.

Algorithm 1: Automatic re-calibration method.

```
  input  : $S_{\tilde{\mathscr{E}}}(\boldsymbol{\theta})$, $\mathscr{W}_d$, $\boldsymbol{\xi}$, $RL$-algorithm
  output: $\boldsymbol{\theta}^*$ (optimal configuration)
1 Calibration stage by $S_{\tilde{\mathscr{E}}}(\boldsymbol{\theta})$
2 Fine-tuning by $RL$
3 Deployment stage
4 while $\mathscr{W}_d$ retrieves normal do
5     deploy device
6     if $\mathscr{W}_d$ retrieves anomaly then
7         return to step 1
```

III. Illustrative example and numerical development

As an application example, we consider the binary coherent-state discrimination, which is a primitive used in long-distance classical-quantum communication. The usage of quantum resources is expected to boost long-distance communication rates [64, 65]and provide unconditional security [66]. Optimally performing quantum-state discrimination is of utmost importance to reach capacity rates [67, 68], and the binary coherent-state discrimination problem currently stands as a canonical problem both from a theoretical point of view [69-72], as well as an experimental one [11-13, 15-17, 73-78]. The interest on this problem lies on the fact that the optimal quantum measurement to be done by the receiver can be implemented sequentially, combining linear optical operations and feedback operations, which constitutes an experimentally friendly setting.

In this setting, the sender encodes a bit $k = 0, 1$ in the phase of a quantum coherent-state $|(-1)^k\alpha\rangle$, which is sent to the receiver; *e.g.* the signal is prepared in an orbital space (satellite) station, travels through the atmosphere and arrives to a receiver, in a ground-earth station. The latter performs a binary-outcome quantum measurement, leading to measurement outcome $n \in \{0, 1\}$. With this information, the receiver provides a guess $\hat{k}$ on the value of the bit transmitted, and the quality of such protocol is given by the success probability. Such a quantity represents the score function $S_{\mathscr{E}}(\boldsymbol{\theta})$ introduced in Sec. II, and depends on the intensity $|\alpha|^2$ of the transmitted states, the quantum channel acting over which the communication takes place, and the specific quantum measurement that is performed by the receiver.

Among all possible quantum measurements that the receiver can implement, we will here focus on the Kennedy receiver [79] , which consists in displacing the incoming signal by a value θ and measuring the resulting state via an on/off photo-detector, as schematized in Fig. 3. While the optimal quantum measurement is given by the Dolinar receiver [70, 76] , which involves complex conditional measurements ultimately leading to difficulties in experimental implementations [73, 74], the Kennedy receiver can readily beat the standard quantum limit [80] and essentially constitutes the main building block of the former one.

In this example, the device configuration is defined by *(i)* the parameter θ in the displacement operation, and *(ii)* a guessing rule which associates the measurement outcome n to the guessed value of the initially transmitted bit k. We note that access to the score value (success probability) is granted only in cases where the transmission channel and device functioning have been perfectly characterized. Such is not often the case, as atmospheric conditions turn to strongly vary unpredictably, a fact that ultimately affects the transmission performance [11-13, 15-17, 77].

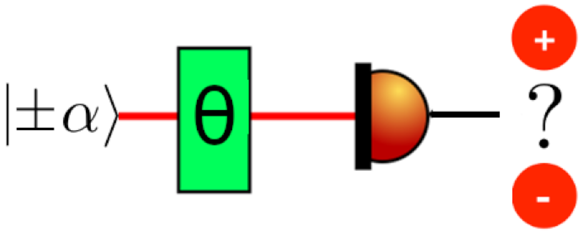

FIG. 3: Diagram of a Kennedy receiver; this consists in applying a displacement θ to the incoming signal and measure it with an on/off photo-detector.

We now revisit the re-calibration framework introduced in Sec.II for the Kennedy receiver. As stated above, the score function $S_{\mathscr{E}}(\boldsymbol{\theta})$ is given by the success probability of the communication protocol, which depends on the displacement value θ, and the guessing rule $\hat{k}(\theta, n)$. Note that if access to the outcome probabilities is granted, then the agent would perform a maximum-likelihood guess. However, in situations where such probabilities are not available, *e.g.* no model of transmission channel, then the agent needs also to learn the optimal guessing rule. Thus, we remark that the score function is dependent on the specific transmission channel acting between sender and receiver, and potentially differs from the noiseless success probability, *i.e.* identity channel acting in between parties. The latter quantity constitutes in our approach the effective score $S_{\tilde{\mathscr{E}}}(\boldsymbol{\theta})$. Here, the intensity $|\alpha|^2$ is initially estimated using N_{eff} experiments, where the displacement value θ is set to zero, and thus the outcomes probabilities can readily be linked to α through the Born rule, $p(n = 0|(-1)^k\alpha) = e^{-|(-1)^k\alpha|^2}$, with $\sum_{i=0,1} p(n = i|(-1)^k\alpha) = 1$. Outcome statistics are used to estimate the signal intensity, which in turn serves as a way to initialize the state-action value functions $\{Q(\theta), Q(\hat{k};n,\theta)\}$ to the success probability of setting displacement θ and conditional probabilities of having $\hat{k}$ given observation n and displacement θ respectively.

The aforementioned quantities are consequently used by a Q-learning agent, which fine-tunes the calibrating strategy to the experimental conditions at hand; this is done by providing a binary reward to the agent according to the correctness of its guess $\hat{k}$, and it can be proven that such scheme converges to the optimal device configuration [10]. The Q-learning method is applied for N_{rl} experiments, and then the receiver is *deployed*. While in deployment stage, the agent monitors the measurement outcome statistics, by keeping track of a running average. This quantity serves as a decalibration witness $\mathscr{W}_d$, and abrupt changes of this quantity indicate that a change-point has occurred. When the system is out of the expected region (specified by the agent), the calibration protocol is re-started.

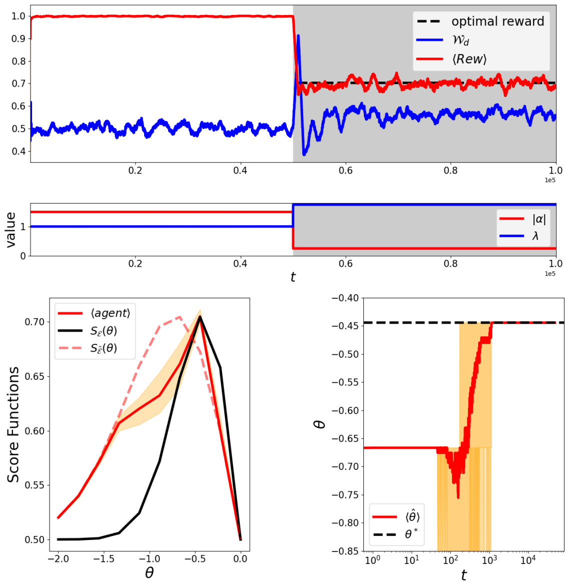

FIG. 4: Recalibration and learning curve. (Top): Learning curve evolution. Running average of the reward acquired using 10^3 *experiments, and the evolution of the decalibration witness* $\mathscr{W}_d$ *estimated by measurement statistics. As can be seen in the change-point, the Witness presents a big fluctuation, starting a recalibration of the system until the agent converges to the optimal reward. (Bottom): Update of the Q-values curve (left) and evolution of agent's greedy strategy, i.e. the configuration the calibrating agent would choose at each experiment (right).*

Malfunctioning device example. We now consider the case in which the Kennedy receiver is initially calibrated to its optimal configuration, with a pre-defined intensity value $|\alpha_0|^2$, deployed to ideal conditions for such initial environment $\mathscr{E}_0$, and incurs into a decalibration. The new environment $\mathscr{E}_1$ consists in a different intensity value $|\alpha_1|^2$ of the signals arriving to the receiver, plus a *faulty* displacement. Here, the value θ the agent fixes, actually displaces the signal by a value $\lambda\theta$, with $\lambda \geqslant 1$ being an unknown parameter. The effect of this faulty behavior is to make displacements bigger than expected, shifting the value of the optimal configuration θ^*. As a consequence, the score-function landscape gets modified. We remark that the malfunctioning behavior is unknown to the agent, who first loads the Q-values for the ideal case using its effective model $S_{\tilde{\mathscr{E}}}(\boldsymbol{\theta})$, and then fine-tunes them by Q-learning; further details on the implementation are provided in Appendix A.

In Fig. 4 (top) we show the learning-curve of the agent in terms of cumulative reward acquired. The decalibration witness $\mathscr{W}_d$ is taken to be an estimate of the outcome probability $\hat{p}(n=1)$, and by monitoring abrupt changes in this quantity, the agent is able to detect the environment shift $\mathscr{E}_0 \rightarrow \mathscr{E}_1$. The change point in which the device enters a malfunctioning stage occurs at experiment 5 10^5, and can readily be seen in the top panel of Fig. 4 by the change of $\mathscr{W}_d$ behavior. Additionally, this can be detected by an abrupt change in the cumulative reward acquired; however, we note that such quantity is potentially not available during deployment stage. As a consequence, the agent uses its change-point detection strategy to re-activate the calibration protocol again, by estimating the new signal intensity and initializing the Q-values in the effective model obtained thereby. Note that in this new scenario, the effective model does not coincide anymore with the underlying truth. This fact is illustrated by the initial and final Q-values obtained by the agent, shown in Fig. 4 (bottom), where we additionally show the optimal configuration $\hat{\theta}^*$ suggested by the agent at a given experiment.

IV. Outlook & future research directions

In this work, we presented a re-calibration framework, accompanied by an automatic re-calibration method, and targeted to quantum technology applications.

We illustrated the proposed method by studying a Kennedy receiver under heavily varying deployment conditions. As in any device, decalibration is a frequent problem that needs to be addressed. This example serves as a test-bed for our automatic re-calibration framework, showing that not only the calibrating agent is able to configure the device in a semi-agnostic way, but also to detect situations in which the device gets off-calibrated. Our mechanism allows for the automation of the re-calibration process and can readily be applied to a wider scope, even beyond quantum technology applications.

Specifically, our technique reduces the number of experiment repetitions needed to (re)-calibrate the device. This is done by making use of an effective model, whose purpose is to capture main features in the configuration landscape, and is complemented by model-free reinforcement-learning techniques. Additionally, we introduce the decalibration witness statistic, which plays a key role in detecting either novelties or anomalies referring to the device's functioning. Such quantity is conceived as a figure of merit to be calculated during device deployment. In this stage, the score function for the quality of the controls that are chosen by the agent is not computable, and the agent can only rely on information available in the experiment, *e.g.* statistics from the measurement outcome.

A plethora of change-point/anomaly detection methods can be used in order to complement our method [30-32, 81-83]. However, let us remark that an alternative to monitoring the decalibration witnesses can also be brought to attention, *i.e.* by presetting a calibration control routine. Such scheme demands balancing between device deploying and guaranteeing that the optimal configuration is being kept, and can potentially be implemented by allowing intermediate calibration stages in between deployment. We remark that while model-free RL techniques could potentially adapt the controls to *smooth* changes in the optimal configuration (without the necessity of an effective model nor a decalibration witness), abrupt changes would in practice corrupt a successful adaptation. Here, an *abruptness* notion is unveiled when it comes to environment changes: on the one hand we identify continual reinforcement learning [84, 85] (where the calibration agent smoothly adapts the configuration as the environment *smoothly* varies), and on the other hand domain adaptation in reinforcement learning [86, 87] (where the calibration needs to be adapted under changes of abrupt nature, as the ones considered in this paper). The setting studied in here might also be tackled from an active learning framework [88], in which the agent may inject prior knowledge on the different conditions in which the device is expected to be deployed, and can potentially be used to further exploit the symbiosis between model-free and model-aware routines considered above.

A straightforward extension of settings where our recalibration framework finds real-world implementations is given by Noisy Intermediate-Scale Quantum (NISQ) devices, where the strong presence of noise severely limits the scope of applications, and developing tools to address such issues is an active area of research.

Furthermore, our work opens the door for several follow-up implementations and enhancements of the re-calibration protocol. Among them, usage of more sophisticated RL methods [43, 89] and inspecting the possibility of a coherent re-calibration by usage of quantum correlations [90, 91].

V. Acknowledgments

M.B. is grateful for useful discussions with John Calsamiglia about enhancing the Q-values initialization. T.C wants to thank F. Tomás B. Perez, for ideas in possible applications and numerical results. T.C. acknowledges the support from the CVC fellowship program. M.B. acknowledges support from AGAUR Grant no. 2023 INV-2 00034 funded by the European Union, Next Generation EU and Grant PID2021-126808OB-I00 funded by MCIN/AEI/ 10.13039/501100011033 and by ERDF A way of making Europe. F.V and M.B acknowledge the support from the Spanish Ministry of Science and Innovation through the project GRAIL, grant no. PID291-1268080B-100. J. M. M. and L.R. acknowledge to Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) and support from ANPCyT Argentina, Préstamo BID, Grant no. PICT 20203490.

REFERENCES

[1] V. Cimini, E. Polino, M. Valeri, I. Gianani, N. Spagnolo, G. Corrielli, A. Crespi, R. Osellame, M. Barbieri y F. Sciarrino. Calibration of Multiparameter Sensors via Machine Learning at the Single-Photon Level. Physical Review Applied **15** (abr. de 2021). http://dx.doi.org/10.1103/physrevapplied.15.044003.

[2] V. Cimini, M. Valeri, S. Piacentini, F. Ceccarelli, G. Corrielli, R. Osellame, N. Spagnolo y F. Sciarrino. Variational quantum algorithm for experimental photonic multiparameter estimation. npj Quantum Information **10** (feb. de 2024). http://dx.doi.org/10.1038/s41534-024-00821-0.

[3] H. Ren, J. Yang, X. Liu, P. Huang y L. Guo. Sensor Modeling and Calibration Method Based on Extinction Ratio Error for Camera-Based Polarization Navigation Sensor. Sensors **20** (jul. de 2020). http://dx.doi.org/10.3390/s20133779.

[4] F. Vernuccio, A. Bresci, V. Cimini, A. Giuseppi, G. Cerullo, D. Polli y C. M. Valensise. Artificial Intelligence in Classical and Quantum Photonics. Laser & Photonics Reviews **16** (mar. de 2022). http://dx.doi.org/10.1002/lpor.202100399.

[5] K. Ono. Calibration Methods of Acoustic Emission Sensors. Materials **9** (jun. de 2016). http://dx.doi.org/10.3390/ma9070508.

[6] S. Zhao, J. Liu e Y. Li. Online Calibration Method for Current Sensors Based on GPS. Energies **12** (mayo de 2019). http://dx.doi.org/10.3390/en12101923.

[7] H. Haitjema. The Calibration of Displacement Sensors. Sensors **20** (ene. de 2020). http://dx.doi.org/10.3390/s20030584.

[8] V. Cimini, I. Gianani, N. Spagnolo, F. Leccese, F. Sciarrino y M. Barbieri. Calibration of Quantum Sensors by Neural Networks. Physical Review Letters **123** (dic. de 2019). http://dx.doi.org/10.1103/physrevlett.123.230502.

[9] Y. Zhang, L. O. H. Wijeratne, S. Talebi y D. J. Lary. Machine Learning for Light Sensor Calibration. Sensors **21** (sep. de 2021). http://dx.doi.org/10.3390/s21186259.

[10] M. Bilkis, M. Rosati, R. M. Yepes y J. Calsamiglia. Real-time calibration of coherent-state receivers: Learning by trial and error. Phys. Rev. Res. **2,** 033295 (ago. de 2020). https://link.aps.org/doi/10.1103/PhysRevResearch.2.033295.

[11] M. Bilkis, M. Rosati y J. Calsamiglia. *Reinforcement-learning calibration of coherent-state receivers on variable-loss optical channels* en *2021 IEEE Information Theory Workshop (ITW)* (2021), 1-6.

[12] D. Dequal, L. T. Vidarte, V. R. Rodriguez, A. Leverrier, G. Vallone, P. Villoresi y E. Diamanti. *Feasibility of satellite quantum key distribution with continuous variable* en *Quantum Inf. Meas. V Quantum Technol.* **Part F165-** (OSA, Washington, D.C., feb. de 2019), T5A.89. ISBN: 978-1-943580-56-9. arXiv: 2002.02002. http://arxiv.org/abs/2002.02002%20https://www.osapublishing.org/abstract.cfm?URI=QIM-2019-T5A.89.

[13] L. C. Andrews y R. L. Phillips. *Laser Beam Propagation through Random Media* ISBN: 9780819459480. http://ebooks.spiedigitallibrary.org/book.aspx?doi=10.1117/3.626196 (SPIE, 1000 20th Street, Bellingham, WA 98227-0010 USA, sep. de 2005).

[14] S. Pirandola. Satellite quantum communications: Fundamental bounds and practical security. Phys. Rev. Res. **3,** 023130 (mayo de 2021). ISSN: 2643-1564. https://link.aps.org/doi/10.1103/PhysRevResearch.3.023130.

[15] S. Pirandola. Limits and security of free-space quantum communications. Phys. Rev. Res. **3,** 013279 (mar. de 2021). ISSN: 2643-1564. https://link.aps.org/doi/10.1103/PhysRevResearch.3.013279.

[16] D. Y. Vasylyev, A. A. Semenov y W. Vogel. Toward Global Quantum Communication: Beam Wandering Preserves Nonclassicality. Phys. Rev. Lett. **108,** 220501 (jun. de 2012). https://link.aps.org/doi/10.1103/PhysRevLett.108.220501.

[17] D. Vasylyev, A. A. Semenov, W. Vogel, K. Günthner, A. Thurn, Ö. Bayraktar y C. Marquardt. Free-space quantum links under diverse weather conditions. Phys. Rev. A **96,** 043856 (oct. de 2017). https://link.aps.org/doi/10.1103/PhysRevA.96.043856.

[18] D. Jung, Y. Choi y J. Kim. Multiobjective Automatic Parameter Calibration of a Hydrological Model. Water **9** (mar. de 2017). http://dx.doi.org/10.3390/w9030187.

[19] D. Kavetski, G. Kuczera y S. W. Franks. Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts. Journal of Hydrology **320.** The model parameter estimation experiment, 173-186 (2006). ISSN: 0022-1694. https://www.sciencedirect.com/science/article/pii/S0022169405003379.

[20] K. Papafotis, D. Nikitas y P. P. Sotiriadis. Magnetic Field Sensors' Calibration: Algorithms' Overview and Comparison. Sensors **21** (ago. de 2021). http://dx.doi.org/10.3390/s21165288.

[21] G. Cao, X. Xu y D. Xu. Real-Time Calibration of Magnetometers Using the RLS/ML Algorithm. Sensors **20** (ene. de 2020). http://dx.doi.org/10.3390/s20020535.

[22] A. Fallani, M. A. C. Rossi, D. Tamascelli y M. G. Genoni. Learning Feedback Control Strategies for Quantum Metrology. PRX Quantum **3,** 020310 (abr. de 2022). https://link.aps.org/doi/10.1103/PRXQuantum.3.020310.

[23] L. J. Fiderer, J. Schuff y D. Braun. Neural-Network Heuristics for Adaptive Bayesian Quantum Estimation. PRX Quantum **2,** 020303 (abr. de 2021). https://link.aps.org/doi/10.1103/PRXQuantum.2.020303.

[24] L. J. Fiderer y D. Braun. Quantum metrology with quantum-chaotic sensors. Nature Communications **9,** 1351 (abr. de 2018). ISSN: 2041-1723. https://doi.org/10.1038/s41467-018-03623-z.

[25] C. Lee, B. Lawrie, R. Pooser, K.-G. Lee, C. Rockstuhl y M. Tame. Quantum Plasmonic Sensors. Chemical Reviews **121** (mar. de 2021). http://dx.doi.org/10.1021/acs.chemrev.0c01028.

[26] S. Nolan, A. Smerzi y L. Pezzè. A machine learning approach to Bayesian parameter estimation. npj Quantum Information **7** (dic. de 2021). http://dx.doi.org/10.1038/s41534-021-00497-w.

[27] Y. Ban, J. Echanobe, Y. Ding, R. Puebla y J. Casanova. Neural-network-based parameter estimation for quantum detection. Quantum Science and Technology **6** (ago. de 2021). http://dx.doi.org/10.1088/2058-9565/ac16ed.

[28] Y. Chen, Y. Ban, R. He, J.-M. Cui, Y.-F. Huang, C.-F. Li, G.-C. Guo y J. Casanova. A neural network assisted 171Yb+ quantum magnetometer. npj Quantum Information **8** (dic. de 2022). http://dx.doi.org/10.1038/s41534-022-00669-2.

[29] K. Rambhatla, S. E. D'Aurelio, M. Valeri, E. Polino, N. Spagnolo y F. Sciarrino. Adaptive phase estimation through a genetic algorithm. Physical Review Research **2** (jul. de 2020). http://dx.doi.org/10.1103/physrevresearch.2.033078.

[30] G. Sentís, E. Bagan, J. Calsamiglia, G. Chiribella y R. Muñoz-Tapia. Quantum Change Point. Phys. Rev. Lett. **117,** 150502 (oct. de 2016). https://link.aps.org/doi/10.1103/PhysRevLett.117.150502.

[31] M. Fanizza, C. Hirche y J. Calsamiglia. Ultimate Limits for Quickest Quantum Change-Point Detection. Phys. Rev. Lett. **131,** 020602 (jul. de 2023). https://link.aps.org/doi/10.1103/PhysRevLett.131.020602.

[32] G. Sentís, J. Calsamiglia y R. Muñoz-Tapia. Exact Identification of a Quantum Change Point. Phys. Rev. Lett. **119,** 140506 (oct. de 2017). https://link.aps.org/doi/10.1103/PhysRevLett.119.140506.

[33] K. A. Woźniak, V. Belis, E. Puljak, P. Barkoutsos, G. Dissertori, M. Grossi, M. Pierini, F. Reiter, I. Tavernelli y S. Vallecorsa. *Quantum anomaly detection in the latent space of proton collision events at the LHC* 2023. eprint: arXiv:2301.10780.

[34] J. S. Baker, H. Horowitz, S. K. Radha, S. Fernandes, C. Jones, N. Noorani, V. Skavysh, P. Lamontangne y B. C. Sanders. *Quantum Variational Rewinding for Time Series Anomaly Detection* 2022. eprint: arXiv:2210.16438.

[35] M. Guo, S. Pan, W. Li, F. Gao, S. Qin, X. Yu, X. Zhang y Q. Wen. Quantum algorithm for unsupervised anomaly detection. Physica A: Statistical Mechanics and its Applications **625,** 129018 (2023). ISSN: 0378-4371. https://www.sciencedirect.com/science/article/pii/S0378437123005733.

[36] S. Llorens, G. Sentís y R. Muñoz-Tapia. *Quantum multianomaly detection* 2023. eprint: arXiv:2312.13020.

[37] M. Skotiniotis, R. Hotz, J. Calsamiglia y R. Muñoz-Tapia. *Identification of malfunctioning quantum devices* 2018. eprint: arXiv:1808.02729.

[38] N. Liu y P. Rebentrost. Quantum machine learning for quantum anomaly detection. Phys. Rev. A **97,** 042315 (abr. de 2018). https://link.aps.org/doi/10.1103/PhysRevA.97.042315.

[39] M. A. Nielsen e I. L. Chuang. *Quantum Computation and Quantum Information* (Cambridge University Press, 2000).

[40] M. P. A. Branderhorst, J. Nunn, I. A. Walmsley y R. L. Kosut. Simplified quantum process tomography. New Journal of Physics **11,** 115010 (nov. de 2009). https://dx.doi.org/10.1088/1367-2630/11/11/115010.

[41] A. Shabani, R. L. Kosut, M. Mohseni, H. Rabitz, M. A. Broome, M. P. Almeida, A. Fedrizzi y A. G. White. Efficient Measurement of Quantum Dynamics via Compressive Sensing. Phys. Rev. Lett. **106,** 100401 (mar. de 2011). https://link.aps.org/doi/10.1103/PhysRevLett.106.100401.

[42] M. T. DiMario y F. E. Becerra. Channel-noise tracking for sub-shot-noise-limited receivers with neural networks. Physical Review Research **3** (mar. de 2021). http://dx.doi.org/10.1103/physrevresearch.3.013200.

[43] R. Sutton y A. G. Barto. *Reinforcement Learning Sutton* ISBN: 9780262039246 (MIT Press, 2018).

[44] A. Dawid, J. Arnold, B. Requena, A. Gresch, M. Płodzień, K. Donatella, K. A. Nicoli, P. Stornati, R. Koch, M. Büttner, R. Okuła, G. Muñoz-Gil, R. A. Vargas-Hernández, A. Cervera-Lierta, J. Carrasquilla, V. Dunjko, M. Gabrié, P. Huembeli, E. van Nieuwenburg, F. Vicentini, L. Wang, S. J. Wetzel, G. Carleo, E. Greplová, R. Krems, F. Marquardt, M. Tomza, M. Lewenstein y A. Dauphin. *Modern applications of machine learning in quantum sciences* 2022. eprint: arXiv:2204.04198.

[45] S. Borah, B. Sarma, M. Kewming, G. J. Milburn y J. Twamley. Measurement-Based Feedback Quantum Control with Deep Reinforcement Learning for a Double-Well Nonlinear Potential. Phys. Rev. Lett. **127,** 190403 (nov. de 2021). https://link.aps.org/doi/10.1103/PhysRevLett.127.190403.

[46] H. J. Briegel y G. De las Cuevas. Projective simulation for artificial intelligence. Scientific Reports **2,** 400 (mayo de 2012). ISSN: 2045-2322. https://doi.org/10.1038/srep00400.

[47] J. Wallnöfer, A. A. Melnikov, W. Dür y H. J. Briegel. Machine Learning for Long-Distance Quantum Communication. PRX Quantum **1,** 010301 (sep. de 2020). https://link.aps.org/doi/10.1103/PRXQuantum.1.010301.

[48] C. Cui, W. Horrocks, S. Hao, S. Guha, N. Peyghambarian, Q. Zhuang y Z. Zhang. Quantum receiver enhanced by adaptive learning. Light: Science & Applications **11,** 344 (dic. de 2022). ISSN: 2047-7538. https://doi.org/10.1038/s41377-022-01039-5.

[49] N. Rengaswamy, K. P. Seshadreesan, S. Guha y H. D. Pfister. Belief propagation with quantum messages for quantum-enhanced classical communications. npj Quantum Information **7,** 97 (jun. de 2021). ISSN: 2056-6387. https://doi.org/10.1038/s41534-021-00422-1.

[50] C. Piveteau y J. M. Renes. Quantum message-passing algorithm for optimal and efficient decoding. Quantum **6,** 784 (ago. de 2022). ISSN: 2521-327X. https://doi.org/10.22331/q-2022-08-23-784.

[51] C. L. Cortes, P. Lefebvre, N. Lauk, M. J. Davis, N. Sinclair, S. K. Gray y D. Oblak. Sample-efficient adaptive calibration of quantum networks using Bayesian optimization. Phys. Rev. Applied (mar. de 2022). journals.aps.org/prapplied/abstract/10.1103/PhysRevApplied.17.034067.

[52] V. V. Sivak, A. Eickbusch, H. Liu, B. Royer, I. Tsioutsios y M. H. Devoret. Model-Free Quantum Control with Reinforcement Learning. Phys. Rev. X **12,** 011059 (mar. de 2022). https://link.aps.org/doi/10.1103/PhysRevX.12.011059.

[53] M. Y. Niu, S. Boixo, V. N. Smelyanskiy y H. Neven. Universal quantum control through deep reinforcement learning. npj Quantum Information **5,** 33 (abr. de 2019). ISSN: 2056-6387. https://doi.org/10.1038/s41534-019-0141-3.

[54] T. Fösel, P. Tighineanu, T. Weiss y F. Marquardt. Reinforcement Learning with Neural Networks for Quantum Feedback. Phys. Rev. X **8,** 031084 (sep. de 2018). https://link.aps.org/doi/10.1103/PhysRevX.8.031084.

[55] P. Altmann, J. Stein, M. Kölle, A. Bärligea, T. Gabor, T. Phan, S. Feld y C. Linnhoff-Popien. *Challenges for Reinforcement Learning in Quantum Circuit Design* 2023. eprint: arXiv:2312.11337.

[56] M. Nägele y F. Marquardt. *Optimizing ZX-Diagrams with Deep Reinforcement Learning* 2023. eprint: arXiv:2311.18588.

[57] A. Skolik, M. Cattelan, S. Yarkoni, T. Bäck y V. Dunjko. *Equivariant quantum circuits for learning on weighted graphs* 2022. eprint: arXiv:2205.06109.

[58] A. Khandelwal y S. DiAdamo. *Enhancing Protocol Privacy with Blind Calibration of Quantum Devices* 2022. eprint: arXiv:2209.05634.

[59] B. Zhou, C. Lu, B.-M. Mao, H.-y. Tam y S. He. Magnetic field sensor of enhanced sensitivity and temperature self-calibration based on silica fiber Fabry-Perot resonator with silicone cavity. Opt. Express **25,** 8108-8114 (abr. de 2017). https://opg.optica.org/oe/abstract.cfm?URI=oe-25-7-8108.

[60] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio y P. J. Coles. Variational quantum algorithms. Nature Reviews Physics **3,** 625-644 (sep. de 2021). ISSN: 2522-5820. https://doi.org/10.1038/s42254-021-00348-9.

[61] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek y A. Aspuru-Guzik. Noisy intermediate-scale quantum algorithms. Rev. Mod. Phys. **94,** 015004 (feb. de 2022). https://link.aps.org/doi/10.1103/RevModPhys.94.015004.

[62] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik y J. L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. Nature Communications **5,** 4213 (jul. de 2014). ISSN: 2041-1723. https://doi.org/10.1038/ncomms5213.

[63] E. Fontana, M. Cerezo, A. Arrasmith, I. Rungger y P. J. Coles1. Non-trivial symmetries in quantum landscapes and their resilience to quantum noise. Quantum **6** (sep. de 2022). https://quantum-journal.org/papers/q-2022-09-15-804/#.

[64] K. Banaszek, L. Kunz, M. Jachura y M. Jarzyna. Quantum Limits in Optical Communications. J. Light. Technol. **38,** 2741-2754 (mayo de 2020). ISSN: 0733-8724. arXiv: 2002.05766. https://ieeexplore.ieee.org/document/8998224/.

[65] M. Rosati y V. Giovannetti. Achieving the Holevo bound via a bisection decoding protocol. J. Math. Phys. **57,** 062204 (jun. de 2015). ISSN: 00222488. arXiv: 1506.04999. http://aip.scitation.org/doi/10.1063/1.4953690%20http://arxiv.org/abs/1506.04999%20http://dx.doi.org/10.1063/1.4953690.

[66] S. Pirandola, U. L. Andersen, L. Banchi, M. Berta, D. Bunandar, R. Colbeck, D. Englund, T. Gehring, C. Lupo, C. Ottaviani, J. L. Pereira, M. Razavi, J. Shamsul Shaari, M. Tomamichel, V. C. Usenko, G. Vallone, P. Villoresi y P. Wallden. Advances in quantum cryptography. Adv. Opt. Photonics **12,** 1012 (dic. de 2020). ISSN: 1943-8206. arXiv: 1906.01645. http://arxiv.org/abs/1906.01645%20http://dx.doi.org/10.1364/AOP.361502%20https://www.osapublishing.org/abstract.cfm?URI=aop-12-4-1012.

[67] M. M. Wilde. *Quantum Information Theory* (Cambridge University Press, 2013).

[68] R. Nasser y J. M. Renes. *Polar codes for arbitrary classical-quantum channels and arbitrary cq-MACs* en *2017 IEEE International Symposium on Information Theory (ISIT)* (2017), 281-285.

[69] C. W. Helstrom. *Quantum Detection and Estimation Theory* 309. ISBN: 9780123400505 (Academic press, New York, 1976).

[70] A. Assalini, N. Dalla Pozza y G. Pierobon. Revisiting the Dolinar receiver through multiple-copy state discrimination theory. Phys. Rev. A **84,** 022342 (ago. de 2011). https://link.aps.org/doi/10.1103/PhysRevA.84.022342.

[71] F. Zoratti, N. Dalla Pozza, M. Fanizza y V. Giovannetti. Agnostic Dolinar receiver for coherent-state classification. Phys. Rev. A **104,** 042606 (oct. de 2021). https://link.aps.org/doi/10.1103/PhysRevA.104.042606.

[72] M. Takeoka, M. Sasaki, P. van Loock y N. Lütkenhaus. Implementation of projective measurements with linear optics and continuous photon counting. Phys. Rev. A **71,** 022318 (feb. de 2005). https://link.aps.org/doi/10.1103/PhysRevA.71.022318.

[73] R. L. Cook, P. J. Martin, J. M. Geremia, B. A. Chase y J. M. Geremia. Optical coherent state discrimination using a closed-loop quantum measurement. Nature **446,** 774-777 (abr. de 2007). ISSN: 14764687. http://www.nature.com/doifinder/10.1038/nature05655%20http://www.nature.com/articles/nature05655.

[74] D. Sych y G. Leuchs. Practical Receiver for Optimal Discrimination of Binary Coherent Signals. Phys. Rev. Lett. **117,** 200501 (nov. de 2016). https://link.aps.org/doi/10.1103/PhysRevLett.117.200501.

[75] F. E. Becerra, J. Fan, G. Baumgartner, S. V. Polyakov, J. Goldhar, J. T. Kosloski y A. Migdall. *M*-ary-state phase-shift-keying discrimination below the homodyne limit. Phys. Rev. A **84,** 062324 (dic. de 2011). https://link.aps.org/doi/10.1103/PhysRevA.84.062324.

[76] S. J. Dolinar. Communication and sciences engineering. Q. Prog. Rep. (Research Lab. Electron. **111,** 115 (1973). https://dspace.mit.edu/handle/1721.1/56414.

[77] V. C. Usenko, B. Heim, C. Peuntinger, C. Wittmann, C. Marquardt, G. Leuchs y R. Filip. Entanglement of Gaussian states and the applicability to quantum key distribution over fading channels. New J. Phys. **14,** 093048 (sep. de 2012). ISSN: 1367-2630. https://iopscience.iop.org/article/10.1088/1367-2630/14/9/093048.

[78] M. T. DiMario y F. E. Becerra. Demonstration of optimal non-projective measurement of binary coherent states with photon counting. npj Quantum Information **8,** 84 (jul. de 2022). ISSN: 2056-6387. https://doi.org/10.1038/s41534-022-00595-3.

[79] R. S. Kennedy. Near-Optimum Receiver for the Binary Coherent State Quantum Channel. MIT Res. Lab. Electron. Q. Prog. Rep. **108,** 219 (1973). https://dspace.mit.edu/handle/1721.1/56346.

[80] A. Ferraro, S. Olivares y M. G. A. Paris. *Gaussian states in continuous variable quantum information* ISBN: ISBN 88-7088-483-X (Bibliopolis, Napoli, 2005).

[81] E. S. PAGE. A test for a change in a parameter occurring at an unknown point. Biometrika **42,** 523-527 (dic. de 1955). ISSN: 0006-3444. eprint: https://academic.oup.com/biomet/article-pdf/42/3-4/523/838813/42-3-4-523.pdf. https://doi.org/10.1093/biomet/42.3-4.523.

[82] E. Brodsky y B. Darkhovsky. *Non-Parametric Statistical Diagnosis: Problems and Methods* ISBN: 9789048154654. https://books.google.es/books?id=Ar56cgAACAAJ (Springer Netherlands, 2010).

[83] M. Basseville e I. Nikiforov. *Detection of Abrupt Change Theory and Application* ISBN: 0-13-126780-9 (Prentice Hall, abr. de 1993).

[84] D. Abel, A. Barreto, B. V. Roy, D. Precup, H. van Hasselt y S. Singh. *A Definition of Continual Reinforcement Learning* 2023. eprint: arXiv:2307.11046.

[85] K. Khetarpal, M. Riemer, I. Rish y D. Precup. *Towards Continual Reinforcement Learning: A Review and Perspectives* 2020. eprint: arXiv:2012.13490.

[86] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba y P. Abbeel. *Domain randomization for transferring deep neural networks from simulation to the real world* en *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2017), 23-30.

[87] J. Xing, T. Nagata, K. Chen, X. Zou, E. Neftci y J. L. Krichmar. *Domain Adaptation In Reinforcement Learning Via Latent Unified State Representation* 2021. eprint: arXiv:2102.05714.

[88] P. Radeva, M. Drozdzal, S. Segui, L. Igual, C. Malagelada, F. Azpiroz y J. Vitria. *Active labeling: Application to wireless endoscopy analysis* en *2012 International Conference on High Performance Computing & Simulation (HPCS)* (2012), 174-181.

[89] Y. Baum, M. Amico, S. Howell, M. Hush, M. Liuzzi, P. Mundada, T. Merkh, A. R. Carvalho y M. J. Biercuk. Experimental Deep Reinforcement Learning for Error-Robust Gate-Set Design on a Superconducting Quantum Computer. PRX Quantum **2,** 040324 (nov. de 2021). https://link.aps.org/doi/10.1103/PRXQuantum.2.040324.

[90] Y. Liao, M.-H. Hsieh y C. Ferrie. *Quantum Optimization for Training Quantum Neural Networks* 2021. eprint: arXiv:2103.17047.

[91] A. A. et. al. *Quantum Optimization: Potential, Challenges, and the Path Forward* 2023. eprint: arXiv:2312.02279.

[92] R. BELLMAN y S. Dreyfus. *Dynamic Programming* ISBN: 9780691146683. http://www.jstor.org/stable/j.ctv1nxcw0f (2024) (Princeton University Press, 2010).

[93] C. Szepesvári. *Algorithms for Reinforcement Learning* 2010. http://dx.doi.org/10.2200/S00268ED1V01Y201005AIM009.

[94] C. Watkins. *"Learning from delayed rewards". PhD thesis* Tesis doct. (Cambridge, 1989). http://www.cs.rhul.ac.uk/%7B~%7Dchrisw/thesis.html.

[95] T. Lattimore y C. Szepesvári. *Bandit Algorithms* (Cambridge University Press, 2020).

[96] T. Crosta, M. Matera y M. Bilkis. *Repository* https://github.com/dmtomas/qrec (GitHub, 2024).

A. Additional details in the RL implementation

In the following, we briefly present the Reinforcement Learning (RL) framework and provide further details on the illustrative example considered Sec. III. Moreover, we briefly analyze an alternative noisy scenario where a change of priors occur, and benchmark our techniques with a standard Q-learning method.

Reinforcement Learning is based on the sequential interaction between an agent and the environment during several episodes [43]. Each episode E consists on steps $t = 1,\ldots,T$ (where T is potentially of stochastic nature). At step t, the agent observes a *state* s_t, and follows a policy $\pi(a_t|s_t) \equiv \pi$ in order to choose an *action* a_t. As a consequence, the agent receives a *reward* r_{t+1} and transitions to the next state s_{t+1}. The goal of the agent is to maximize the reward acquired during episodes, which is accomplished by performing the *optimal policy*. To do this, the agent has to exploit valuable actions but also explore possibly advantageous configurations, leading to an exploration-exploitation trade-off. The framework allows for intermediate rewards appearing during the episode, and hence the *return* is defined as $G_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k+1}$, with $\gamma \leqslant 1$ being a weighting factor. The latter is a quantity that depends on the sequence of state and actions visited during each episode, and its average value — the so-called state-action value function— $Q_\pi(s,a) = \mathbb{E}_\pi G_t|s_t = s, a_t = a$ indicates how valuable action a is by departing from state s and following policy π thereafter. In this setting, the optimal policy π^* is obtained by finding the maximum Q-value for each given state, a problem that is reflected by the so-called *optimal Bellman equation* [43, 92].

In this regard, the *Q-learning algorithm* is a model-free method that exploits the structure of Bellman equations by linking them to contractive operations and shifting the policy towards the fixed point associated to the optimal Bellman operator [93, 94]. In order to find the optimal Q-values $Q^*(s,a) = Q_{\pi^*}(s,a)$, the algorithm updates the Q-estimate as

$$\hat{Q}(s_t,a_t) \leftarrow (1-\lambda_E)\hat{Q}(s_t,a_t) + \lambda_E\left(r_{t+1} + \gamma \underset{a'}{\text{máx}}\,\hat{Q}(s_{t+1},a')\right) \tag{1}$$

with λ_E an episode-dependent learning-rate; in Alg.2 we sketch the Q-learning pseudo-code. Here, the agent explores the state-action space by committing to a ε-greedy policy π_ε, defined as selecting a random action with probability ε, and the one maximizing the current state-action value estimate $\hat{Q}(s,a)$ otherwise. Note that *(i)* such *greedy* action might potentially be suboptimal option, and *(ii)* a schedule for ε is set in practice, in order to balance between exploration and exploitation [10, 43, 95].

We now turn to provide additional details on the numerical implementation for the Kennedy receiver considered in Sec. III. Our code is open-sourced and can be found in Ref. [96]. All hyperparameter values used in this implementation are given in Table 1.

Parameters	Meaning	Proposed method	Q-learning
check jump threshold	How much repeated selection of the maximum is considered conversion.	3000	3000
δ	How much the change in $\mathscr{W}_d$ has to be in order to recalibrate.	0.1	0.1
ε_0	Minimum exploration of the agent.	0.05	0.1
Δ_ε	Rate of change for ε.	0.9	0.9999
Δ	How much do we want to deviate from a uniform distribution.	50	0
Δ_l	Step from which the learning rate starts	150	1

TABLA 1: Hyperparameters used in the numerical examples with a description of its interpretations.

Algorithm 2: Q-learning pseudocode.

```
for episode E = 1, ... do
    initialize $s_0$
    for step t in episode E do
        choose $a_t \sim \pi_\varepsilon$
        get $r_t, s_{t+1}$
        update $\hat{Q}(s_t,a_t)$ using Eq.1.
```

Decalibration witness. In order to detect changes in the environment during an off-calibration stage, the running average output of the detector is computed across experiments $E = 1,\ldots,N_{\text{eff}}$, where we set $N_{\text{eff}} = 1000$, *e.g.* $\mathscr{W}_d{}^{(E)} = \frac{1}{N_{\text{eff}}}\sum_{i=0}^{N_{\text{eff}}} n_{t-i}$. At each experiment, the difference between the current average and the previous one is computed $|\mathscr{W}_d{}^{E} - \mathscr{W}_d{}^{E-1}|$. Here, if this difference is bigger than the (hyper)parameter δ — and assuming the device is being deployed — the re-calibration process is restarted.

Effective model. When the (re-)calibration is initiated, we consider an effective model given by the success probability of a noiseless Kennedy receiver, computed by first estimating the signal's intensity $|\alpha|^2$. Such success probability can be linked to the optimal Q-values for an ideal environment $\mathscr{E}_0$ in which the device functions correctly [10]. To this end, the displacement value in the Kennedy receiver is set to zero during $N_{\text{eff}} = 1000$ experiments, and the intensity $|\alpha|^2$ is estimated as per $|\alpha|^2 = -\ln(p(n = 0\big|\alpha)) \pm \frac{1}{N_{\text{eff}}}$. Consequently, the Q-learning agent fine-tunes the device configuration, which is potentially deployed under an environment whose score function differs from the noiseless effective-model here considered. Importantly, the Q-values are initialized to $Q_0^\alpha(\theta)$ and $Q_1^\alpha(\theta,\hat{k})$ as per $Q_0^\alpha(\theta) = \sum_{n=\{0,1\}} \text{máx}_{\hat{k}=0,1}\, Q_1^\alpha(\theta,\hat{k})$ and $Q_1^\alpha(\theta,\hat{k}) = \frac{1}{2}e^{-|(-1)^{\hat{k}}\alpha+\theta|^2} + \frac{1}{2}\left(1 - e^{-|(-1)^{\hat{k}+1}\alpha+\theta|^2}\right)$

Q-learning hyperparameters. We scheduled the exploration rate of π_ε as per $\varepsilon_E = \text{máx}\,(\varepsilon_0, \varepsilon_E\Delta_\varepsilon)$, with $\varepsilon_0, \Delta_\varepsilon \in$

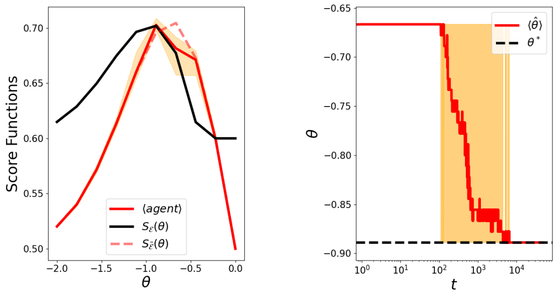

FIG. 5: We show the average internal strategy of 25 calibrating agents to fine-tune the device configuration under a change-of-prior scenario. Specifically, we depict the Q-values (left panel) from the effective model, and the ones obtained after RL fine-tuning. In the right panel, we show the evolution of agent's greedy strategy.

$[0,1)$ and $\varepsilon_{t=0} = 1$, *e.g.* it is reduced over different episodes. Here, ε_0 provides the minimum exploration level, while Δ_ε gives a rate of change in the exploration. On a different note, we modify the uniform sampling in π_ε by $p(\hat{a}|\hat{Q}) = \frac{1}{\mathcal{N}} \exp\left(-\Delta\left|\hat{Q}(\hat{a}^*) - Q(\hat{a})\right|^2\right)$; where $\frac{1}{\mathcal{N}}$ is a normalization factor, $\hat{Q}(a)$ is the current Q-value estimate, $\hat{a}^*$ is the associated greedy action, and Δ an importance-sampling parameter ($\Delta = 0$ returns a uniformly random distribution) [43]. Finally, the Q-learning learning-rate λ_E in Eq.(1) is set to decay as $\sim 1/E$, where we recall that E is the episode number. Note that because of the initial information obtained by setting the effective model, we allow the learning-rate to take smaller values as per $\frac{1}{t+\Delta_l}$, where Δ_l can be understood as how much the effective model is trusted by the RL agent.

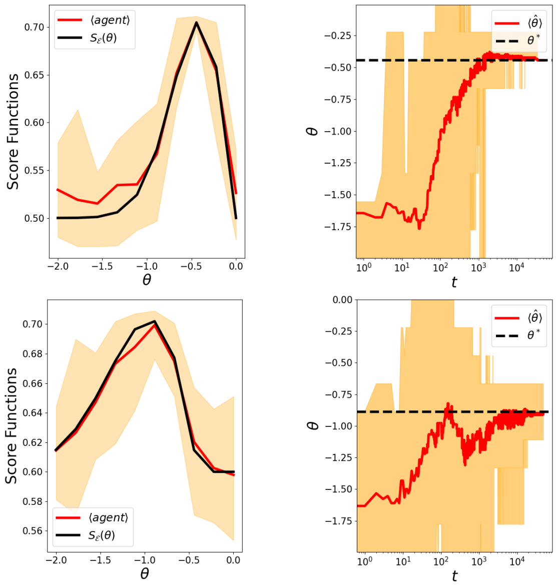

FIG. 6: Re-calibration learning curve. We show the mean internal strategy of 25 reinforcement-learning agent's, to optimize the device configuration, (top) using a faulty displacement and (bottom) a change of the prior probability. Specifically, we depict the Q-values (left panel) obtained by trial and error and the evolution of the greedy strategy (right panel).

Change of priors example. To test the resilience of the method to a different noise source, we here analyze the situation in which a change of priors occurs. We recall the prior constitutes the probability p_k of sending the classical bit k, and in this example it gets modified as per $p_k \to p_k(\lambda_2) = \frac{1}{2} + (-1)^k \lambda_2$, where λ_2 stands for the noise parameter, unknown to the agent. The results of our automatic re-calibration method for this scenario are show in Fig. 5. Here, we average the results over 25 instances obtained through random initializations, taking 10^3 experiments to find the optimal configuration in %90 of the runs and 7×10^3 to finish.

Comparison with Q-learning: Finally, we compare the technique introduced with the *standard* Q-learning method [10]. The results are benchmarked in Fig 6, under the noisy scenarios previously considered. As shown in the figure, traditional Q-learning presents higher fluctuations over the Q-value estimates, requiring $10x$ the amount of experiments than the method introduced in this paper, which exploits the usage of an effective model.