Measurement-based adaptation protocol with quantum reinforcement learning

F. Albarrán-Arriagada,[1,*] J. C. Retamal,[1,2] E. Solano,[3,4,5] and L. Lamata[3]

[1]*Departamento de Física, Universidad de Santiago de Chile (USACH), Avenida Ecuador 3493, 9170124, Santiago, Chile*
[2]*Center for the Development of Nanoscience and Nanotechnology 9170124, Estación Central, Santiago, Chile*
[3]*Department of Physical Chemistry, University of the Basque Country UPV/EHU, Apartado 644, 48080 Bilbao, Spain*
[4]*IKERBASQUE, Basque Foundation for Science, Maria Diaz de Haro 3, 48013 Bilbao, Spain*
[5]*Department of Physics, Shanghai University, 200444 Shanghai, China*

(Dated: May 9, 2019)

Machine learning employs dynamical algorithms that mimic the human capacity to learn, where the reinforcement learning ones are among the most similar to humans in this respect. On the other hand, adaptability is an essential aspect to perform any task efficiently in a changing environment, and it is fundamental for many purposes, such as natural selection. Here, we propose an algorithm based on successive measurements to adapt one quantum state to a reference unknown state, in the sense of achieving maximum overlap. The protocol naturally provides many identical copies of the reference state, such that in each measurement iteration more information about it is obtained. In our protocol, we consider a system composed of three parts, the "environment" system, which provides the reference state copies; the register, which is an auxiliary subsystem that interacts with the environment to acquire information from it; and the agent, which corresponds to the quantum state that is adapted by digital feedback with input corresponding to the outcome of the measurements on the register. With this proposal we can achieve an average fidelity between the environment and the agent of more than 90% with less than 30 iterations of the protocol. In addition, we extend the formalism to d-dimensional states, reaching an average fidelity of around 80% in less than 400 iterations for $d = 11$, for a variety of genuinely quantum as well as semiclassical states. This work paves the way for the development of quantum reinforcement learning protocols using quantum data, and the future deployment of semi-autonomous quantum systems.

I. INTRODUCTION

Machine learning (ML) is an area of artificial intelligence that focuses on the implementation of learning algorithms, and which has undergone great development in recent years [1–3]. ML can be classified into two broad groups, namely, learning by means of big data and through interactions. For the first group we have two classes, supervised learning, which uses previously classified data to train the learning program, inferring the function of relationship to classify new data. This is the case, e.g., of pattern recognition problems [4–7]. The other class is unsupervised learning, which does not require training data, but this paradigm uses the big data distribution to obtain an optimal way to classify it using specific characteristics. An example is the clustering problem [8, 9].

For the second group, learning from interactions, we have the case of reinforcement learning (RL) [10, 11]. RL is the more similar paradigm to the human learning process. Its general framework is as follows: we define two basic systems, an agent A and an environment E, while often it is useful to define a register R as an auxiliary system. The concept consists of A inferring information by direct interaction with E, or indirectly, using as a mediator the system R. With the obtained information, A makes a decision to perform a certain task. If the result of this task is good, then the agent receives a reward, otherwise a punishment. In addition, the RL algorithms can be divided into three basic parts, the policy, the reward function (RF) and the value function (VF). The policy can be subdivided into three stages: first, interaction with the environment. In this stage, the way in which A or R interacts with E is specified. Second, information extraction, which indicates how A obtains information from E. Finally, action, where A takes the decision of what to do with the information of the previous step. RF refers to the criterion to award the reward or punishment A in each iteration. And VF evaluates the utility of A referred to the given task. An example of RL consists in artificial players for go or chess [12, 13].

Other essential aspect of the RL protocols is the exploitation-exploration relation. Exploitation refers to the ability to make good decisions, while exploration is the possibility of making different decisions. For example, if we want to select a gym to do sports, the exploitation is given by the quality of the gym we tested, while the exploration is the size of the search area in which we will choose a new gym to test. In the RL paradigm, a good exploitation-exploration relation can guarantee the convergence of the learning process, and its optimization depends on each algorithm.

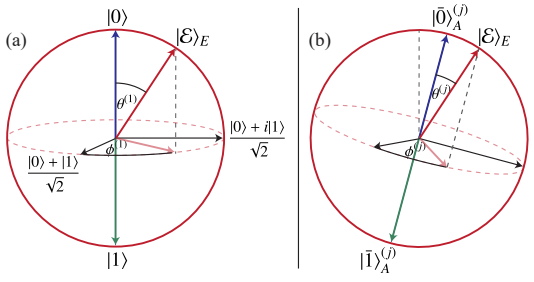

FIG. 1. (a) Bloch representation of the environment state at initial time, with $|0\rangle$ the state of the agent. (b) Bloch representation of the environment in the jth iteration, where $|\bar{0}_j\rangle$ is the state of the agent, which was rotated in the previous iterations.

* F. Albarrán-Arriagada francisco.albarran@usach.cl

On the other hand, quantum mechanics is known to improve computational tasks [14], so a natural question is: how are the learning algorithms modified in the quantum domain? To answer this question the quantum machine learning field (QML) has emerged. In recent years, QML has been a fruitful area [15–22], in which quantum algorithms have been developed [23–26], that show a possible speedup in certain situations in relation with their classical counterparts [27, 28]. However, these novel works focus mainly on learning from classical data encoded in quantum systems, processed with a quantum algorithm and decoded to be read by a classical machine. In this context, the speedup of quantum machine learning is often balanced with the necessary resources to encode and decode the information, which leads to an unclear quantum supremacy. On the other hand, recent works analyze the QML paradigm in a purely quantum way [29–31], in which quantum systems learn quantum data.

In this article, we present a quantum machine learning algorithm based on a reinforcement learning approach, to convert the quantum state of a system (Agent), into an unknown state encoded, via multiple identical copies, in another system (Environment), assisted by measurements on a third system (Register). We propose to use coherent feedback loops, conditioned to the measurements in order to perform the adaptation process without human intervention. In our numerical calculations we obtain average fidelities of more than 90% for qubit states after less than 40 measurements, while for qudits the protocol achieves average fidelities of 80% using 400 iterations with $d = 11$ dimensions, either for genuinely quantum or semiclassical states. This proposal can be useful in the way to implement semi-autonomous quantum devices.

II. THE QUANTUM ADAPTATION ALGORITHM

Our framework is as follows. We assume a known quantum system called agent (A), and many copies of an unknown quantum state provided by a system called environment (E). We also consider an auxiliary system called register (R) which interacts with E. Then, we obtain information about E by measuring R, and employ the result as an input to the reward (RF) function. Finally, we perform a partially-random unitary transformation on A, which depends on the output of the RF. The idea is to improve the fidelity between A and E, without projecting the state of A with measurements.

This protocol differs from quantum state estimation [32–37] in the fact that we propose a semi-autonomous quantum agent, that is, the aim is that in the future a quantum agent will learn the state of the environment without any human intervention. Other authors have considered the inverse problem, an unknown state evolved to a known state assisted by measurements [38], which deviate from the machine learning paradigm. Therefore, an optimal measurement is not performed in each step, but after a certain number of autonomous iterations, the agent converges to a large fidelity with the unknown state.

In the rest of the article we use the following notation: the subscripts A, R and E refer to each subsystem, and the superscripts indicate the iteration. For example, $O_\alpha^{(k)}$ refers to the operator O that acts on the subsystem α during the kth iteration. Moreover, the lack of any of these indices indicates that we are referring to a general object in the iterations and/or in the subsystems.

We start with the case where each subsystem is described by a qubit state. We assume that $A(R)$ is described by $|0\rangle_{A(R)}$, and E by an arbitrary state expressed in the Bloch sphere as $|\mathcal{E}\rangle_E = \cos(\theta^{(1)}/2)|0\rangle_E + e^{-i\phi^{(1)}}\sin(\theta^{(1)}/2)|1\rangle_E$ [see Fig. 1 (a)]. The initial state reads

$$|\psi^{(1)}\rangle = |0\rangle_A|0\rangle_R[\cos(\theta^{(1)}/2)|0\rangle_E + e^{i\phi^{(1)}}\sin(\theta^{(1)}/2)|1\rangle_E]. \quad (1)$$

First of all, we will introduce the general elements of our reinforcement learning protocol, such as policy, the RF and the VF. For the policy, we perform a Controlled-NOT (CNOT) gate ($U_{E,R}^{NOT}$) with E as control and R as target (i.e., the interaction with the environment), in order to copy information of E into R, obtaining

$$\begin{aligned}|\Psi_1\rangle =& U_{E,R}^{NOT}|\psi^{(1)}\rangle \\ =& |0\rangle_A\Big[\cos(\theta^{(1)}/2)|0\rangle_R|0\rangle_E + e^{i\phi^{(1)}}\sin(\theta^{(1)}/2)|1\rangle_R|1\rangle_E\Big]. \quad (2)\end{aligned}$$

We then measure the register qubit in the basis $\{|0\rangle, |1\rangle\}$, with probability $p_0^{(1)} = \cos^2(\theta_1/2)$, or $p_1^{(1)} = \sin^2(\theta_1/2)$, to obtain the state $|0\rangle$ or $|1\rangle$ respectively (i.e., information extraction). If the result is $|0\rangle$, it means that we collapse E into A and do nothing, but if the result is $|1\rangle$, it means that we measure the orthogonal component to A of E, and thus we accordingly modify the agent. As we do not have additional information about the environment, we perform a partially-random unitary operator on A given by $U_A^{(1)}(\alpha^{(1)},\beta^{(1)}) = e^{-iS_A^{z(1)}\alpha^{(1)}}e^{-iS_A^{x(1)}\beta^{(1)}}$ (action), where $\alpha^{(1)}$ and $\beta^{(1)}$ are random angles of the form $\alpha(\beta)^{(1)} = \xi_{\alpha(\beta)}\Delta^{(1)}$, with $\xi_{\alpha(\beta)} \in [-1/2, 1/2]$ a random number, $\Delta^{(1)}$ is the range of random angles, $\alpha(\beta)^{(1)} \in [-\Delta^{(1)}/2, \Delta^{(1)}/2]$, and $S_A^{k(1)} = S^k$ is the kth spin component. Now, we initialize the register qubit state and employ a new copy of E, obtaining the next initial state for the second iteration

$$|\psi^{(2)}\rangle = \mathcal{U}_A^{(1)}|0\rangle_A|0\rangle_R|\mathcal{E}\rangle_E = |\bar{0}\rangle_A^{(2)}|0\rangle_R|\mathcal{E}\rangle_E, \quad (3)$$

with

$$\mathcal{U}_A^{(1)} = \Big[m^{(1)}U_A^{(1)}(\alpha^{(1)},\beta^{(1)}) + (1-m^{(1)})\mathbb{I}_A\Big]. \quad (4)$$

Here, $m^{(1)} = \{0, 1\}$, is the outcome of the measurement, $\mathbb{I}$ is the identity operator, and we define the new agent state as $|\bar{0}\rangle_1^{(2)} = \mathcal{U}_1^{(1)}|0\rangle_1$.

Now, we define the RF to modify the exploration range of the kth iteration $\Delta^{(k)}$ as

$$\Delta^{(k)} = \Big[(1-m^{(k-1)})\mathcal{R} + m^{(k-1)}\mathcal{P}\Big]\Delta^{(k-1)}, \quad (5)$$

where $m^{(k-1)}$ is the outcome of the $(k-1)$th iteration, while $\mathcal{R}$ and $\mathcal{P}$ are the reward and punishment ratios, respectively. Equation (5) means, that the value of Δ is modified by $\mathcal{R}\Delta$ for the next iteration when the previous outcome is $m = 0$, and by $\mathcal{P}\Delta$ when the outcome is $m = 1$. In our protocol, we

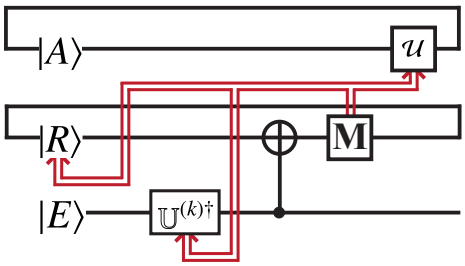

FIG. 2. Quantum circuit diagram for the measurement-based adaptation protocol. The box labelled with M indicates the projective measurement process, and the red lines denote feedback loops.

choose for simplicity $\mathcal{R} = \epsilon < 1$ and $\mathcal{P} = 1/\epsilon > 1$, such that, every time that the state $|0\rangle$ is measured, the value of Δ is reduced, and increased in the other case. Also, the fact that $\mathcal{R}\cdot\mathcal{P} = 1$ means that the punishment and the reward have the same strength, or in other words, if the protocol yields the same number of outcomes 0 and 1, the exploration range does not change. Finally, the VF is defined as the value of $\Delta^{(n)}$ after all iterations. Therefore, $\Delta^{(n)} \to 0$ if the protocol improves the fidelity between A and E.

To illustrate the behaviour of the protocol, we consider the kth iteration. The initial state is given by

$$|\psi\rangle^{(k)} = |\bar{0}\rangle_A^{(k)}|0\rangle_R|\mathcal{E}\rangle_E, \tag{6}$$

where $|\bar{0}\rangle_A^{(k)} = \mathbb{U}_A^{(k)}|0\rangle_A$, $\mathbb{U}_A^{(k)} = \mathcal{U}_A^{(k-1)}\mathbb{U}_A^{(k-1)}$, with $\mathbb{U}_A^{(1)} = \mathbb{I}_A$, and $\mathcal{U}_A^{(j)}$ given by Eq. (4). Also $U_A^{(j)} = e^{-iS_A^{z(j)}\alpha^{(j)}}e^{-iS_A^{x(j)}\beta^{(j)}}$, where we define

$$\begin{aligned} S_A^{z(j)} &= \frac{1}{2}(|\bar{0}\rangle_A^{(j)}\langle\bar{0}| - |\bar{1}\rangle_A^{(j)}\langle\bar{1}|) = \mathcal{U}_A^{(j-1)\dagger}S_A^{z(j-1)}\mathcal{U}_A^{(j-1)}, \\ S_A^{x(j)} &= \frac{1}{2}(|\bar{0}\rangle_A^{(j)}\langle\bar{1}| + |\bar{1}\rangle_A^{(j)}\langle\bar{0}|) = \mathcal{U}_A^{(j-1)\dagger}S_A^{x(j-1)}\mathcal{U}_A^{(j-1)}, \end{aligned} \tag{7}$$

with ${}_A^{(j)}\langle\bar{0}|\bar{1}\rangle_A^{(j)} = 0$. We can write the state of E in the Bloch representation using $|\bar{0}_j\rangle$ as a reference axis (see Fig. 1 (b)), and apply the operator $\mathbb{U}_E^{(k)\dagger}$, obtaining for E,

$$\begin{aligned} \mathbb{U}_E^{(k)\dagger}|\mathcal{E}\rangle_E =& \mathbb{U}_E^{(k)\dagger}\left[\cos(\theta^{(k)}/2)|\bar{0}\rangle_E^{(k)} + e^{i\phi^{(k)}}\sin(\theta^{(k)}/2)|\bar{1}\rangle_E^{(k)}\right] \\ =& \cos(\theta^{(k)}/2)|0\rangle_E + e^{i\phi^{(k)}}\sin(\theta^{(k)}/2)|1\rangle_E = |\bar{\mathcal{E}}\rangle_E^{(k)}. \end{aligned} \tag{8}$$

We can write the states $|\bar{0}^{(k)}\rangle$ and $|\bar{1}^{(k)}\rangle$ in terms of the initial logical states $|0\rangle$ and $|1\rangle$, and the unknown angles $\theta^{(k)}$, $\theta^{(1)}$, $\phi^{(k)}$ and $\phi^{(1)}$ as follows

$$\begin{aligned} |\bar{0}^{(k)}\rangle =& \cos\left(\frac{\theta^{(1)}-\theta^{(k)}}{2}\right)|0\rangle + e^{i\phi^{(1)}}\sin\left(\frac{\theta^{(1)}-\theta^{(k)}}{2}\right)|1\rangle \\ |\bar{1}^{(k)}\rangle =& -e^{-i\phi^{(k)}}\sin\left(\frac{\theta^{(1)}-\theta^{(k)}}{2}\right)|0\rangle + e^{i(\phi^{(1)}-\phi^{(k)})}\sin\left(\frac{\theta^{(1)}-\theta^{(k)}}{2}\right)|1\rangle. \end{aligned} \tag{9}$$

Therefore, the operator $\mathbb{U}^{(k)\dagger}$ performs the necessary rotation to transform $|\bar{0}^{(k)}\rangle \to |0\rangle$ and $|\bar{1}^{(k)}\rangle \to |1\rangle$. Then, we perform the gate $U_{E,R}^{NOT}$

$$\begin{aligned} |\Phi^{(k)}\rangle &= U_{E,R}^{NOT}|\bar{0}\rangle_A^{(k)}|0\rangle_R|\bar{\mathcal{E}}\rangle_E \\ &= |\bar{0}\rangle_A^{(k)}\left[\cos(\theta^{(k)}/2)|0\rangle_R|0\rangle_E + e^{i\phi^{(k)}}\sin(\theta^{(k)}/2)|1\rangle_R|1\rangle_E\right], \end{aligned} \tag{10}$$

and we measure R, with probabilities $p_0^{(k)} = \cos^2(\theta^{(k)}/2)$, and $p_1^{(k)} = \sin^2(\theta^{(k)}/2)$, for the outcomes $m^{(k)} = 0$ and $m^{(k)} = 1$, respectively. Finally, we apply the RF given by Eq. (5). We point out that, probabilistically, when $p_0^{(k)} \to 1$, $\Delta \to 0$, and when $p_1^{(k)} \to 1$, $\Delta \to 4\pi$. In terms of exploitation-exploration relation this means that when the exploitation decreases (we measure $|1\rangle$ often), we increase the exploration (we increase the value of Δ) to increase the probability of making a beneficial change, and when the exploitation improves (we measure $|0\rangle$ many times), we reduce the exploration to allow only small changes in the following iterations. The diagram of this protocol is shown in Fig. (2).

Fig. 3 (a) shows the numerical calculation of mean fidelity between A and E for the single-qubit case. For this computation we use 2000 random initial states with $\epsilon = 0.1$ (blue line), $\epsilon = 0.3$ (red line), $\epsilon = 0.5$ (yellow line), $\epsilon = 0.7$ (purple line), and $\epsilon = 0.9$ (green line). We can see that the protocol can reach fidelities over 90% in less than 30 iterations. Fig. 3 (b) depicts the evolution of the exploration parameter Δ for each iteration for the same values of constant ϵ. We can see from Fig. 3, that when the parameter ϵ is small, the fidelity between A and E increases quickly (the learning speed increases), requiring less iterations to reach high fidelities, however, the maximum value of the average fidelity (maximum learning) is smaller than when ϵ increases. This means that small changes in the scan parameter Δ (large ϵ) result in a higher but slower learning.

III. MULTILEVEL PROTOCOL

In this section, we extend the previous protocol to the case where A, R, and E are described by one d-dimensional qudit state. One of the ingredients in the qubit case is the CNOT gate. Here, we use the extension of the CNOT gate to multilevel states, also known as the XOR gate [39] ($U_{a,b}^{XOR}$). The action of this gate is given by

$$U_{a,b}^{XOR}|j\rangle_a|k\rangle_b = |j\rangle_a|j\ominus k\rangle_b, \tag{11}$$

where the index $a(b)$ refers to control(target) state, respectively, and $\ominus$ denotes the difference modulo d, with d the dimension of each subsystem. The CNOT gate has two important properties, namely, (i) $U_{a,b}^{NOT}$ is Hermitian, and (ii) $U_{a,b}^{NOT}|j\rangle_a|k\rangle_b = |j\rangle_a|0\rangle_b$ if and only if $j = k$. These two properties are maintained in the XOR gate defined in Eq. (11). The Policy and VF are essentially the same than in previous case, but now we consider the multiple outcomes ($m^{(j)} \in \{0, 1, \ldots, d-1\}$) that result of measuring R. First, we introduce $|\bar{1}_j\rangle_A = |m^{(j)}\rangle_A$ for the definition of $S_A^{z(j)}$ and $S_A^{x(j)}$ in Eq. (7). As in the previous case, we assume the initial state of A to be $|0\rangle_A$, while R is initialized in $|0\rangle_R$. Moreover, the state of

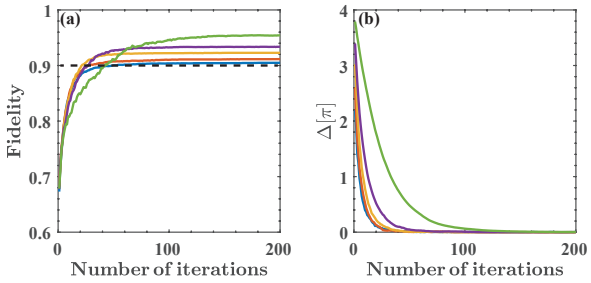

FIG. 3. Performance of the measurement-based adaptation protocol for the qubit case. (a) shows the mean fidelity for 2000 initial random states, and (b) the value of $\Delta^{(j)}$ in each iteration. For both figures, we have $\epsilon = 0.1$ (blue line), $\epsilon = 0.3$ (red line), $\epsilon = 0.5$ (yellow line), $\epsilon = 0.7$ (purple line),and $\epsilon = 0.9$ (green line). The black dashed line indicates an average fidelity of 90%.

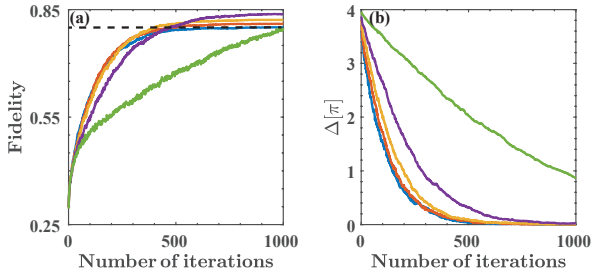

FIG. 4. Performance of the measurement-based adaptation protocol for a total random state given by Eq. (15). (a) shows the mean fidelity for 2000 initial random states, and (b) the value of Δ_j in each iteration. For both figures, we have $\epsilon = 0.1$ (blue line), $\epsilon = 0.3$ (red line), $\epsilon = 0.5$ (yellow line), $\epsilon = 0.7$ (purple line),and $\epsilon = 0.9$ (green line). The black dashed line indicates an average fidelity of 80%.

E is arbitrary, and expressed as $|\mathcal{E}\rangle_E = \sum_{j=0}^{d-1} c_j|j\rangle_E$, where $\sum_{j=0}^{d-1}|c_j|^2 = 1$, and d is the dimension of E. We can rewrite E in a more convenient way as

$$|\mathcal{E}\rangle_E = \cos(\theta^{(1)}/2)|\bar{0}^{(1)}\rangle_E + e^{i\phi^{(1)}}\sin(\theta^{(1)}/2)|\bar{0}^{(1)}_{\perp}\rangle_E, \qquad (12)$$

where $|\bar{0}^{(1)}\rangle_E = |0\rangle_E$, $|\bar{0}^{(1)}_{\perp}\rangle_E = (1/\mathcal{N})\sum_{j=1}^{d-1} c_j|j\rangle_E$ is the orthogonal component to $|0\rangle_E$, and $\mathcal{N}^2 = \sum_{j=1}^{d-1}|c_j|^2$. Subsequently, we perform the XOR gate $U^{XOR}_{E,R}$ obtaining

$$\begin{aligned}|\Phi_0\rangle =& U^{XOR}_{E,R}|0\rangle_A|0\rangle_R|\mathcal{E}\rangle_E \\ =& |0\rangle_A\Big[\cos(\theta^{(1)}/2)|0\rangle_R|0\rangle_E + e^{i\phi^{(1)}}\sin(\theta^{(1)}/2)|\varkappa\rangle_{R,E}\Big], \qquad (13)\end{aligned}$$

with $|\varkappa\rangle_{R,E} = \sum_{j=1}^{d-1}(1/\mathcal{N})c_j|d-j\rangle_R|j\rangle_E$. As in the previous case, we measure R, but now we have multiple outcomes. Therefore, we separated them in two groups. First, the outcome $|0\rangle$ with probability $p_0^{(1)} = \cos^2(\theta_1/2)$. Second, outcomes $|j\rangle$ with $j \neq 0$, and probability to obtain any of them of $p_{\perp}^{(1)} = \sin^2(\theta^{(1)}/2)$. As in the previous case, this means that either we measure in the state of A or in the orthogonal subspace. With this information, we perform a partially-random unitary operation on the agent $\mathcal{U}_A^{(1)} = e^{-iS_A^{z(1)}\alpha^{(1)}}e^{-iS_A^{x(1)}\beta^{(1)}}$, using the definition (7) with $|\bar{1}_A\rangle = |m^{(1)}\rangle_A$, where $m^{(1)} = j$ is the outcome of the measurement. If $m^{(1)} = 0$, then $\mathcal{U}_A^{(1)} = \mathbb{I}_A$. The random angles $\alpha^{(1)}$ and $\beta^{(1)}$ are defined as in the qubit case. Now, the RF changes slightly and is given by

$$\Delta^{(j)} = \Big[\delta_{m^{(j-1)},0}\mathcal{R} + (1-\delta_{m^{(j-1)},0})\mathcal{P}\Big]\Delta^{(j-1)}, \qquad (14)$$

where $\delta_{j,k}$ is the delta function. Equation (14) means that if we measure $|0\rangle$ in R, the value of Δ decreases for the next iteration, and if we measure $|j\rangle$ with $j \neq 0$, Δ increases. Remember that $\mathcal{R} = \epsilon < 1$ and $\mathcal{P} = 1/\epsilon > 1$. As in the qubit case, the RF is binary, since all the results $|j\rangle$ with $j \neq 0$ are equally non-beneficial, so we give the same punishment to the agent. For this reason we use the same policy than in the qubit protocol for the case of multiple levels. As in the case of a single qubit state, the parameter ϵ plays a fundamental role in the learning process by handling the speed of learning and the maximum learning as we will understand in what follows.

Consider the protocol for three different multilevel examples for the E state. First, we consider a total random state with $d = 11$ of the form

$$|\mathcal{E}\rangle_E = \frac{1}{N}\sum_{k=0}^{10} c_k|k\rangle_E, \qquad c_k = a + ib, \qquad (15)$$

where $a, b \in [0, 1)$ are random numbers, and N is a normalization factor. Figure 4 shows the numerical calculations for this case, where (a) gives the average fidelity for 2000 initial states given by Eq. (15), and (b) the evolution of Δ in each iteration. It also shows how this exploration parameter is reduced when the fidelity between E and A grows (increasing the exploitation). We can see from Fig. 4 (a) that the protocol can reach mean fidelities of 80% with about 400 iterations, or, equivalently, the protocol increases the mean fidelity between A and E in about 0.5 using 400 iterations.

Second, consider the protocol for the coherent state defined by

$$|\alpha\rangle = e^{-|\alpha|^2/2}\sum_{n=0}^{\infty}\frac{\alpha^n}{\sqrt{n!}}|n\rangle \qquad (16)$$

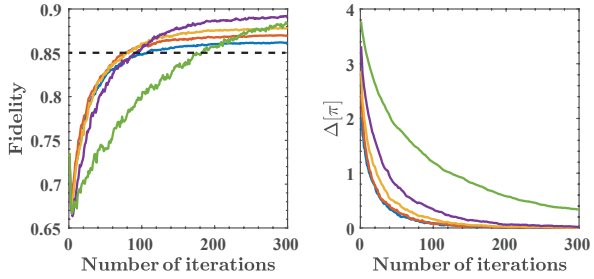

FIG. 5. Performance of the measurement-based adaptation protocol for a coherent state of the form in Eq. (16). (a) shows the mean fidelity for 2000 random pairs $\{a, b\} \in [0, 1)$, where $\alpha = a + ib$; and (b) the value of Δ_j in each iteration. For both figures, we have $\epsilon = 0.1$ (blue line), $\epsilon = 0.3$ (red line), $\epsilon = 0.5$ (yellow line), $\epsilon = 0.7$ (purple line), and $\epsilon = 0.9$ (green line). The black dashed line indicates an average fidelity of 85%.

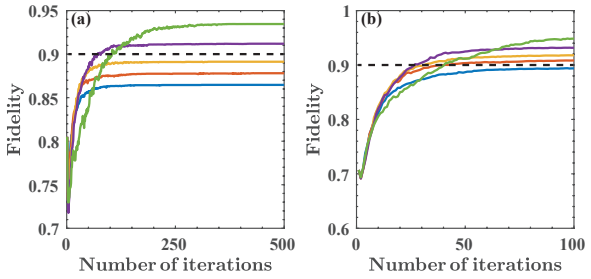

FIG. 6. Performance of the measurement-based adaptation protocol for genuinely quantum states. (a) shows the mean fidelity for 2000 cat states of the form in Eq. (17), and (b) the mean fidelity for 2000 repetitions of the protocol using the superposition $(|0\rangle_E + |10\rangle_E)/\sqrt{2}$ for the environment. In both figures, we use $\epsilon = 0.1$ (blue line), $\epsilon = 0.3$ (red line), $c = 0.5$ (yellow line), $\epsilon = 0.7$ (purple line), and $\epsilon = 0.9$ (green line). The black dashed line indicates an average fidelity of 85%.

for this case we use $\alpha = a + ib$, with a and b positive real random numbers smaller than 1. As $|\alpha|^2 \leq 2$, we can truncate the sum (16) to $n = 10$, since the probabilities to obtain $|n\rangle$ with $n > 10$ are bounded by $|e^{-|\alpha|^2/2}\alpha^{10}/\sqrt{10!}|^2 \leq e^{-1}\sqrt{2^{10}}/\sqrt{10!} \approxeq 0.0062$. Figure 5 (a) shows the fidelity between A and E for each iteration, reaching values of 85% in less than 100 iterations. Figure 5 (b) depicts the value of Δ in this process. We can also observe that the exploration is reduced when A approaches E (increasing the exploitation), as in the previous case.

Finally, we consider two quantum states, a cat state of the form

$$|\mathcal{E}\rangle_E = \sqrt{\frac{1}{\mathcal{N}_\alpha}}\Big(|\alpha\rangle + |-\alpha\rangle\Big) \tag{17}$$

where $|\alpha\rangle$ is given by Eq. (16) and $\mathcal{N}_\alpha$ is a normalization factor. Additionally, we study the superposition

$$|\mathcal{E}\rangle_E = \sqrt{\frac{1}{2}}\Big(|0\rangle + |n\rangle\Big) \tag{18}$$

with $n = 10$. Figure 6 (a) shows the calculation for cat states (17). In this case, we reach fidelities over 90% in about 60 measurements. Moreover, Fig. 6 (b) shows results similar to the qubit case given by Fig. (3), surpassing fidelities of 90% in less than 40 iterations. The last figure reflects the fact that for the state in Eq. (18), the protocol is reduced to the qubit case, given that only two states are involved in the superposition. Thus, all states of the form in Eq. (18) have the same performance as the qubit case.

We can see from Figs. 4, 5 and 6, that the learning speed is inversely proportional to the parameter ϵ, which means that a small value of ϵ implies a rapid increase in fidelity between R and E, that is, it increases the speed of learning. On the other hand, the maximum learning is also directly proportional to ϵ, in other words, a small value of ϵ means lower maximum fidelities between R and E. It is pertinent to emphasize that our protocol for qubit and multilevel cases employs two-level operators $\mathcal{U}_A^{(k)}$, and each iteration only needs to calculate the operator $\mathbb{U}^{(k)} = \mathcal{U}^{(k-1)}\mathbb{U}^{(k-1)}$. Hence, the protocol does not need to store the complete agent history, which is an advantage in terms of the required resources.

This protocol can be implemented in any platform that enables the logical operator $U_{a,b}^{NOT}$ for qubits, or $U_{a,b}^{XOR}$ for qudits, and digital feedback loops, as is the case of circuit quantum electrodynamics (cQEDs). This platform takes particular relevance due to its fast development in quantum computation [40–46]. Current technology in cQEDs allows for digital quantum feedback loops with elapsed times about 2[μs] and fidelities around 99% [47, 48], well-controlled one and two-qubits gates with fidelities over 99% in less than 1[μs] [49], with qubits with coherence times about 100[μs] [50, 51]. This allows for more than 20 iterations of our protocol, a sufficient number for a feasible implementation. Additionally, in the last decade, multilevel gates have been theoretically proposed [52–54], as well as efficient multiqubit gates have recently been proposed using a ML approach [55, 56] providing all the necessary elements for the experimental implementation of the general framework of this learning protocol.

IV. CONCLUSIONS

We propose and analyse a quantum reinforcement learning protocol to adapt a quantum state (the agent) to another, unknown, quantum state (the environment), in the context where several identical copies of the unknown state are available. The main goal of our proposal is for the agent to acquire information about the environment in a semi-autonomous way, namely, in the reinforcement learning spirit. We show that the fidelity increases rapidly with the number of iterations, reaching for qubit states average fidelities over 90% with less than 30 measurements. Also, for states with dimension $d > 2$, we obtain average fidelity over 80% for $d = 11$, with about 400 measurements. The performance is improved for special cases such as coherent states (average fidelities of 85% with less than 100 iterations), cat states (average fidelities of 90% with about 60 iterations) and states of the form $(|0\rangle + |n\rangle)/\sqrt{2}$ (average fidelities of 90% with less than 40 iterations).

The performance of the protocol is handled by the value of the parameter ϵ and by the number of states involved in the superposition of the environment state, E, in the measurement basis. For a small ϵ we get a high learning speed and a reduced maximum learning. Moreover, the number of states in the superposition is related to the overall performance of the protocol, that is, a superposition of fewer terms provides better performance, which increases learning speed as well as maximum learning, requiring less iterations to obtain high fidelity. These two facts imply that a possible improvement of the protocol can be achieved by using a dynamic parameter ϵ and a measurement device that can change its measurement basis throughout the protocol to reduce the number of states involved in the overlap of the state of E. Besides, since our protocol increases the fidelity with a small number of iterations, it is useful even when the number of copies of E is limited. Finally, this protocol opens up the door to the implementation of semi-autonomous quantum reinforcement learning, a

next step for achieving quantum artificial life.

The authors acknowledge support from CONICYT Doctorado Nacional 21140432, Dirección de Postgrado USACH, FONDECYT Grant No. 1140194, Ramón y Cajal Grant RYC-2012-11391, MINECO/FEDER FIS2015-69983-P and Basque Government IT986-16.

[1] S. Russell and P. Norvig, *Artificial Intelligence: A modern approach* (Prentice hall, 1995).
[2] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach* (Springer Science & Business Media, 2013).
[3] M. I. Jordan and T. M. Mitchell, Science **349**, 255 (2015).
[4] M. Kawagoe and A. Tojo, Pattern Recognition **17**, 295 (1984).
[5] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, Science **270**, 303 (1995).
[6] A. K. Jain, Nature **449**, 38 (2007).
[7] J. Carrasquilla, and R. G. Melko, Nat. Phys. **13**, 431 (2017).
[8] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, IEEE Transactions on Emerging Topics in Computing **2**, 267 (2014).
[9] P. Baldi, P. Sadowski, and D. Whiteson, Nat. Comm. **5**, 4308 (2014).
[10] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT press Cambridge, 1998).
[11] M. L. Littman, Nature **521**, 445 (2015).
[12] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, Nature **550**, 354 (2017).
[13] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, *et al.*, arXiv preprint arXiv:1712.01815 (2017).
[14] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information* (Cambridge University Press, Cambridge, UK, 2010).
[15] M. Schuld, I. Sinayskiy, and F. Petruccione, Contemporary Physics **56**, 172 (2015).
[16] J. Adcock, E. Allen, M. Day, S. Frick, J. Hinchliff, M. Johnson, S. Morley-Short, S. Pallister, A. Price, and S. Stanisic, arXiv preprint arXiv:1512.02900 (2015).
[17] V. Dunjko, J. M. Taylor, and H. J. Briegel, Phys. Rev. Lett. **117**, 130501 (2014).
[18] V. Dunjko and H. J. Briegel, arXiv preprint arXiv:1709.02779 (2017).
[19] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Nature **549**, 195 (2017).
[20] R. Biswas, Z. Jiang, K. Kechezhi, S. Knysh, S. Mandrà, B. O'Gorman, A. Perdomo-Ortiz, A. Petukhov, J. Realpe-Gómez, E. Rieffel, D. Venturelli, F. Vasko, and Z. Wang, Parallel Computing **64**, 81 (2017), high-End Computing for Next-Generation Scientific Discovery.
[21] A. Perdomo-Ortiz, M. Benedetti, J. Realpe-Gómez, and R. Biswas, arXiv preprint arXiv:1708.09757 (2017).
[22] A. Perdomo-Ortiz, A. Feldman, A. Ozaeta, S. V. Isakov, Z. Zhu, B. O'Gorman, H. G. Katzgraber, A. Diedrich, H. Neven, J. de Kleer, *et al.*, arXiv preprint arXiv:1708.09780 (2017).
[23] M. Sasaki and A. Carlini, Phys. Rev. A **66**, 022303 (2002).
[24] S. Lloyd, M. Mohseni, and P. Rebentrost, arXiv preprint arXiv:1307.0411 (2013).
[25] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz, Phys. Rev. X **7**, 041052 (2017).
[26] M. Benedetti, J. Realpe-Gómez, and A. Perdomo-Ortiz, arXiv preprint arXiv:1708.09784 (2017).
[27] E. Aïmeur, G. Brassard, and S. Gambs, Machine Learning **90**, 261 (2013).
[28] G. D. Paparo, V. Dunjko, A. Makmal, M. A. Martin-Delgado, and H. J. Briegel, Phys. Rev. X **4**, 031002 (2014).
[29] L. Lamata, Scientific Reports **7**, 1609 (2017).
[30] F. Cárdenas-López, L. Lamata, J. C. Retamal, and E. Solano, arXiv preprint arXiv:1709.07848 (2017).
[31] U. Alvarez-Rodriguez, L. Lamata, P. Escandell-Montero, J. D. Martín-Guerrero, and E. Solano, Scientific Reports **7**, 13645 (2017).
[32] R. B. A. Adamson and A. M. Steinberg, Phys. Rev. Lett. **105**, 030406 (2010).
[33] H. Sosa-Martinez, N. K. Lysne, C. H. Baldwin, A. Kalev, I. H. Deutsch, and P. S. Jessen, Phys. Rev. Lett. **119**, 150401 (2017).
[34] A. Lumino, E. Polino, A. S. Rab, G. Milani, N. Spagnolo, N. Wiebe, and F. Sciarrino, arXiv preprint arXiv:1712.07570v1 (2017).
[35] A. Rocchetto, S. Aaronson, S. Severini, G. Carvacho, D. Poderini, I. Agresti, M. Bentivegna, and F. Sciarrino, arXiv preprint arXiv:1712.00127 (2017).
[36] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, arXiv preprint arXiv:1703.05334 (2017).
[37] L.-L. Sun, Y. Mao, F.-L. Xiong, S. Yu, and Z.-B. Chen, arXiv preprint arXiv:1802.00140 (2018).
[38] L. Roa, A. Delgado, M. L. Ladrón de Guevara, and A. B. Klimov, Phys. Rev. A **73**, 012322 (2006).
[39] G. Alber, A. Delgado, N. Gisin, and I. Jex, arXiv preprint quant-ph/0008022 (2000).
[40] A. Blais, R.-S. Huang, A. Wallraff, S. M. Girvin, and R. J. Schoelkopf, Phys. Rev. A **69**, 062320 (2004).
[41] M. H. Devoret, A. Wallraff, and J. M. Martinis, arXiv preprint cond-mat/0411174 (2004).
[42] M. Hofheinz, E. M. Weig, M. Ansmann, R. C. Bialczak, E. Lucero, M. Neeley, A. D. O'Connell, H. Wang, J. M. Martinis, and A. N. Cleland, Nature **454**, 310 (2008).
[43] M. Hofheinz, H. Wang, M. Ansmann, R. C. Bialczak, E. Lucero, M. Neeley, A. D. O'Connell, D. Sank, J. Wenner, J. M. Martinis, and A. N. Cleland, Nature **459**, 546 (2009).
[44] L. DiCarlo, J. M. Chow, J. M. Gambetta, L. S. Bishop, B. R. Johnson, D. I. Schuster, J. Majer, A. Blais, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, Nature **460**, 240 (2009).
[45] M. H. Devoret and R. J. Schoelkopf, Science **339**, 1169 (2013).
[46] J. Otterbach, R. Manenti, N. Alidoust, A. Bestwick, M. Block, B. Bloom, S. Caldwell, N. Didier, E. S. Fried, S. Hong, *et al.*, arXiv preprint arXiv:1712.05771 (2017).
[47] D. Ristè, J. G. van Leeuwen, H.-S. Ku, K. W. Lehnert, and L. DiCarlo, Phys. Rev. Lett. **109**, 050507 (2012).
[48] D. Ristè, C. C. Bultink, K. W. Lehnert, and L. DiCarlo, Phys. Rev. Lett. **109**, 240502 (2012).
[49] R. Barends, A. Shabani, L. Lamata, J. Kelly, A. Mezzacapo, U. Las Heras, R. Babbush, A. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, E. Jeffrey, E. Lucero, A. Megrant, J. Y. Mutus, M. Neeley, C. Neil, P. J. J. O'Malley, C. Quintana, P. Roushan, D. Sank, A. Vainsencher, J. Wenner, T. C. White, E. Solano, H. Neven, and J. M. Martinis, Nature

534, 222 (2016).
[50] H. Paik, D. I. Schuster, L. S. Bishop, G. Kirchmair, G. Catelani, A. P. Sears, B. R. Johnson, M. J. Reagor, L. Frunzio, L. I. Glazman, S. M. Girvin, M. H. Devoret, and R. J. Schoelkopf, Phys. Rev. Lett. **107**, 240501 (2011).
[51] C. Rigetti, J. M. Gambetta, S. Poletto, B. L. T. Plourde, J. M. Chow, A. D. Córcoles, J. A. Smolin, S. T. Merkel, J. R. Rozen, G. A. Keefe, M. B. Rothwell, M. B. Ketchen, and M. Steffen, Phys. Rev. B **86**, 100506 (2012).
[52] F. W. Strauch, Phys. Rev. A **84**, 052313 (2011).
[53] B. Mischuck and K. Mølmer, Phys. Rev. A **87**, 022341 (2013).
[54] E. O. Kiktenko, A. K. Fedorov, O. V. Man'ko, and V. I. Man'ko, Phys. Rev. A **91**, 042312 (2015).
[55] E. Zahedinejad, J. Ghosh, and B. C. Sanders, Phys. Rev. Lett. **114**, 200502 (2015).
[56] E. Zahedinejad, J. Ghosh, and B. C. Sanders, Phys. Rev. Applied **6**, 054005 (2016).