Journal of Physics: Conference Series

**PAPER** • **OPEN ACCESS**

# Reinforcement Learning Based Control of Coherent Transport by Adiabatic Passage of Spin Qubits

To cite this article: Riccardo Porotti *et al* 2019 *J. Phys.: Conf. Ser.* **1275** 012019

View the article online for updates and enhancements.

IOP ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

This content was downloaded from IP address 151.21.79.173 on 05/12/2019 at 21:07

IOP Conf. Series: Journal of Physics: Conf. Series **1275** (2019) 012019 doi:10.1088/1742-6596/1275/1/012019

# Reinforcement Learning Based Control of Coherent Transport by Adiabatic Passage of Spin Qubits

**Riccardo Porotti[1,2,] , Dario Tamascelli[4], Marcello Restelli[5,] and Enrico Prati[1]**

[1] Istituto di Fotonica e Nanotecnologie, Consiglio Nazionale delle Ricerche, Piazza Leonardo da Vinci 32, I-20133 Milano, Italy
[2] Dipartimento di Fisica "Aldo Pontremoli", Università degli Studi di Milano, via Celoria 16, I-20133 Milano, Italy
[4] Quantum Technology Lab, Dipartimento di Fisica "Aldo Pontremoli", Università degli Studi di Milano, via Celoria 16, I-20133 Milano, Italy
[5] Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133 Milano, Italy

E-mail: `enrico.prati@cnr.it`

**Abstract.** Several tasks involving the determination of the time evolution of a system of solid state qubits require stochastic methods in order to identify the best sequence of gates and the time of interaction among the qubits. The major success of deep learning in several scientific disciplines has suggested its application to quantum information as well. Thanks to its capability to identify best strategy in those problems involving a competition between the short term and the long term rewards, reinforcement learning (RL) method has been successfully applied, for instance, to discover sequences of quantum gate operations minimizing the information loss. In order to extend the application of RL to the transfer of quantum information, we focus on Coherent Transport by Adiabatic Passage (CTAP) on a chain of three semiconductor quantum dots (QD). This task is usually performed by the so called counter-intuitive sequence of gate pulses. Such sequence is capable of coherently transfer an electronic population from the first to the last site of an odd chain of QDs, by leaving the central QD unpopulated. We apply a technique to find nearly optimal gate pulse sequence without explicitly give any prior knowledge of the underlying physical system to the RL agent. Using the advantage actor-critic algorithm, with a small neural net as function approximator, we trained a RL agent to choose the best action at every time step of the physical evolution to achieve the same results previously found only by *ansatz* solutions.

## 1. Introduction

We generate coherent adiabatic passage suitable for transferring quantum states by using a reinforcement learning approach, thus recovering counter-intuitive pulse sequences with no prior knowledge of the system. Such result demonstrates that quantum information processing can be optimized and controlled by reinforcement learning technique, to discover non-trivial strategies.

Coherent transport by adiabatic passage [1] has been introduced in semiconductor quantum dots [2] and later discussed for ab-initio simulations and extended to multiple spin qubits as well [3, 4, 5] . Contrary to supervised learning [6, 7], which has been applied to quantum information related systems earlier [8] , reinforcement learning has been introduced only very recently

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

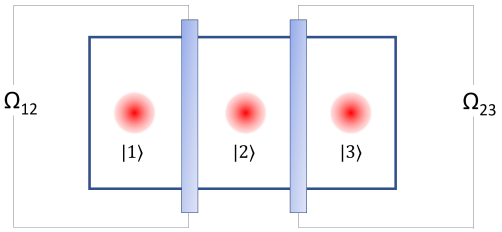


**Figure 1.** The triple quantum dot system. Top control gates of the three quantum dots not shown. The system is occupied by an individual spin at the time, so that it can move along the three quantum dots. The coupling gates can be controlled separately by changing $\Omega_{12}$ and $\Omega_{23}$ respectively.

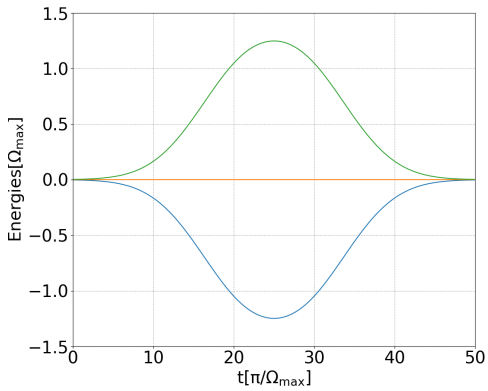


**Figure 2.** The eigenvalues of the triple quantum dot system. The orange value corresponds to the $|D_0\rangle$ eigenvector.

[9, 10]. In the past, some of us have developed genetic algorithms to generate universal logic port silicon spin qubits [11] , CMOS silicon logical qubit design for both Steane and surface codes respectively [12] and CTAP for the hybrid qubits. CTAP provides an excellent example of non-trivial strategy to find a gate pulse sequence to achieve optimum by moving in a quantum framework. The coherent transfer is indeed achieved by a counter-intuitive combination of gate pulsing sequence, i. e. the near and the far gates from the initially occupied quantum dot are opened second and first respectively. By adopting reinforcement learning, we show that such a non-trivial pulse sequence can be discovered with no prior knowledge of the system by applying reinforcement learning method by a neural network.

First, we show that our software based on QuTIP library [13] is able to recover all the classical results of prior literature. Next, the reinforcement learning algorithm is sized to the problem in order to have a reasonable neural network consisting of three layers (input, hidden and output layers respectively), in terms of number of neurons and its implementation. Like for the case of the artificial intelligence learning in the classical Atari game environment [14], the reinforcement learning routine here interacts with the QuTIP simulation of the CTAP implementing the Hilbert space and the time evolution of the system, and exploits the information retrieved by the feedback given by such time evolution generated by the time-varying Hamiltonian. By appropriately defining a simple reward function which constrains the system to avoid occupation of the central quantum dot and to reward occupation of the third quantum dot, the neural network progressively learns how to shape the gate pulses in order to prevent the occupation of the central dot to increase and next to maximize the occupation of the third, thanks of the structure of the eigenvectors of the Hamiltonian. We therefore conclude that reinforcement learning can be successfully applied to quantum information processing to discover highly non-trivial sequences of actions.

## 2. Standard CTAP

Figure 1 shows the typical configuration of three semiconductor quantum dots used to investigate CTAP. Each quantum dot is controlled by a control gate (not shown) from the top to set the ground state energy relatively to the Fermi energy of some external reservoir of electrons. Two coupling gates control the coupling between adjacent dots. Ground state are set by $E_1 = E_2 = E_3 = 0$. If we define the coupling between $i^{th}$ QD and $j^{th}$ QD as $\Omega_{ij}$, the Hamiltonian

reads:

$$H = \begin{pmatrix} E_1 & -\Omega_{12} & 0 \\ -\Omega_{12} & E_2 & -\Omega_{23} \\ 0 & -\Omega_{23} & E_3 \end{pmatrix}. \tag{1}$$

The eigenstates of $H$, written in the QD base (with the respective energies), are [1] :

$$\begin{aligned} |D_+\rangle &= \sin\Theta_1 \sin\Theta_2 \,|1\rangle + \cos\Theta_2 \,|2\rangle + \cos\Theta_1 \sin\Theta_2 \,|3\rangle \\ |D_-\rangle &= \sin\Theta_1 \cos\Theta_2 \,|1\rangle - \sin\Theta_2 \,|2\rangle + \cos\Theta_1 \cos\Theta_2 \,|3\rangle \\ |D_0\rangle &= \cos\Theta_1 \qquad |1\rangle + 0 \qquad |2\rangle - \sin\Theta_1 \qquad |3\rangle \end{aligned}$$

with

$$\Theta_1 = \arctan\left(\frac{\Omega_{12}}{\Omega_{23}}\right). \tag{2}$$

The Hamiltonian dynamics is generated by a routine based on QuTIP. This section is dedicated to show that our implementation recovers all the previous findings known from literature, in particular to Ref.[1]. The energies of the eigenstates are plotted in Figure 2. By changing the eigenstate parameters in time, $|D_0\rangle$ can transform from $|1\rangle$ at $t = 0$ to $|3\rangle$ at $t = t_{max}$. If the Hamiltonian is prepared in $|D_0\rangle$ at $t = 0$, it will remain in the same eigenstate if the adiabaticity criterion is met, that is:

$$|\epsilon_0 - \epsilon_\pm| \gg |\left\langle \dot{D}_0 \middle| D_\pm \right\rangle|. \tag{3}$$

Consequently $\Omega_{12}$ and $\Omega_{23}$ pulses of Gaussian form can achieve coherent transport with high fidelity, if $t_{max} \geq \frac{10\pi}{\Omega_{max}}$. The remarking fact is that the two pulses must be applied in the so-called *counter-intuitive* sequence, as shown in Fig. 2. In Fig. 3a, CTAP for a 3-QD system is shown with a $t_{max} \geq \frac{50\pi}{\Omega_{max}}$.
Time evolution is governed by a master equation involving the density matrix $\rho$:

$$\dot{\rho} = \frac{1}{i\hbar}[H, \rho]. \tag{4}$$

Therefore the occupation of the first dot is provided by $\rho_{11}(t)$, that of the central by $\rho_{22}(t)$ and the final by $\rho_{33}(t)$.

Figure 3 shows the transfer of population from the state $|1\rangle$ to the state $|3\rangle$ with no passage in the state $|2\rangle$. This is achieved by a sequence of pulse called counter-intuitive, as the second gate is pulsed before the first one, closest to the $|1\rangle$ quantum dot. Figure 4 reproduces the results of the detuning of the ground state by $\Delta_{12} = E_2 - E_1$ and $\Delta_{13} = E_3 - E_1$ shown in Ref. [1]. In Figure 5 we further exploit our software to show the pulse region for which $\rho_{33}(t_f) = 1$ at the end of the pulse sequence, complemented with the integral of $\rho_{22}(t)$ during the process. The CTAP is successful only when both $\rho_{33} = 1$ and $\rho_{22} = 0$. for the whole time evolution. The pulses are centered at $\alpha_1$ and $\alpha_2$ respectively, being their time evolution:

$$\begin{aligned} \Omega_{12} &= \Omega^{max} e^{[-(t-\frac{t_{max}+\sigma}{2})^2/(2\sigma^2)]} \\ \Omega_{23} &= \Omega^{max} e^{[-(t-\frac{t_{max}-\sigma}{2})^2/(2\sigma^2)]} \\ \sigma &= t_{max}/8. \end{aligned} \tag{5}$$

There is a wide region of counter-intuitive sequence of pulses allowing to achieve $\rho_{33}(t_f) = 1$. This value is achieved also for some combinations for the "intuitive sequence", i.e. when the closest gate is pulsed first. The integral of $\rho_{22}$ during the process rules out those pulse sequences where the left gate is controlled first, as $\rho_{22}(t)$ does not equal zero for the whole time evolution.

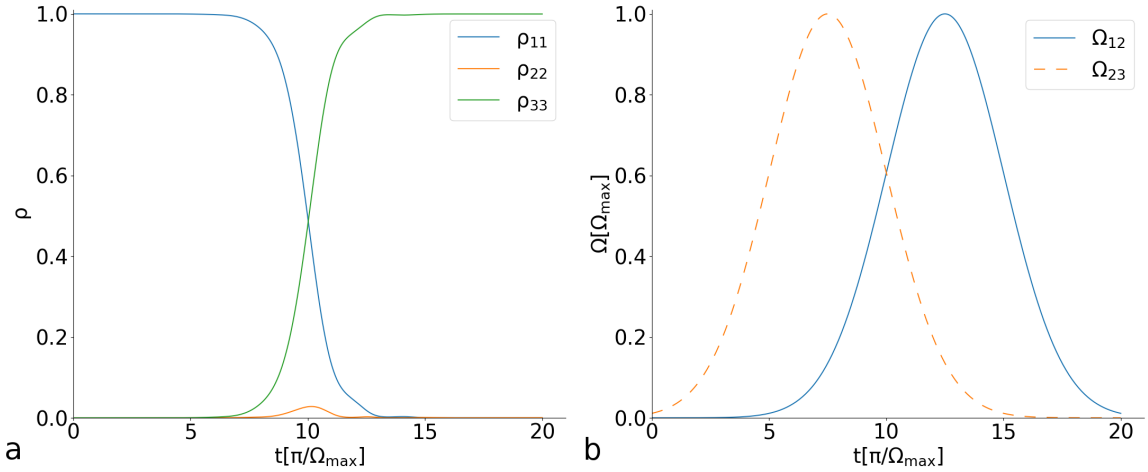


**Figure 3.** a) The CTAP from QD 1 to QD 3 with no occupation of the central QD 2 generated by QuTIP. b) the counter-intuitive pulse sequence of the coupling gates. Blue pulse of the gate voltage between QD 1 and QD 2 opens second.

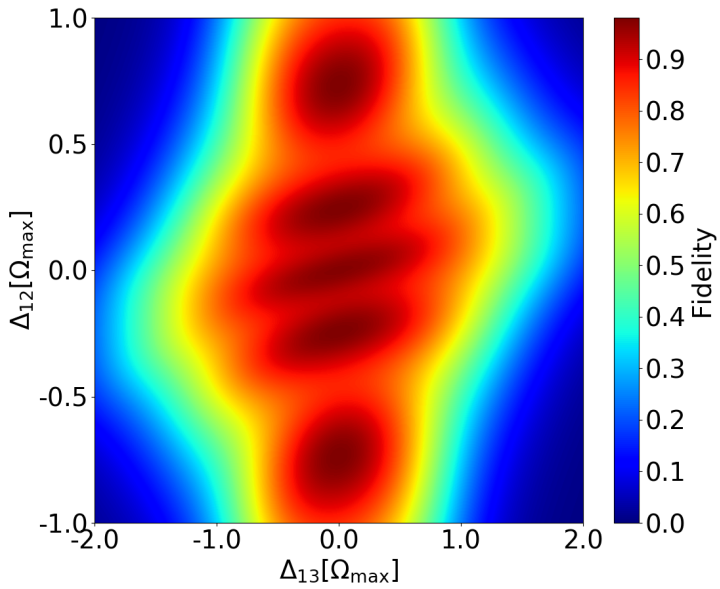


**Figure 4.** CTAP as a function of the detuning of the ground state energy of QD 1 with respect to QD 2 and QD 3 like in [1].

## 3. Description of the Reinforcement Learning algorithm

To rediscover the best pulses for CTAP by exploiting an artificial intelligence agent, we used an Advantage Actor-Critic (A2C) [10, 11] model. The algorithm finds the shape of the pulses of the two control gates without providing any prior knowledge of the system. From now on, we impose the system to behave consistently with episodic Markov Decision Processes (MDPs), so there exists a final state $s_f$ that terminates the episode. A2C mixes two well-known paradigms of RL, namely value-based and policy-based algorithms. In a value-based framework, the expected

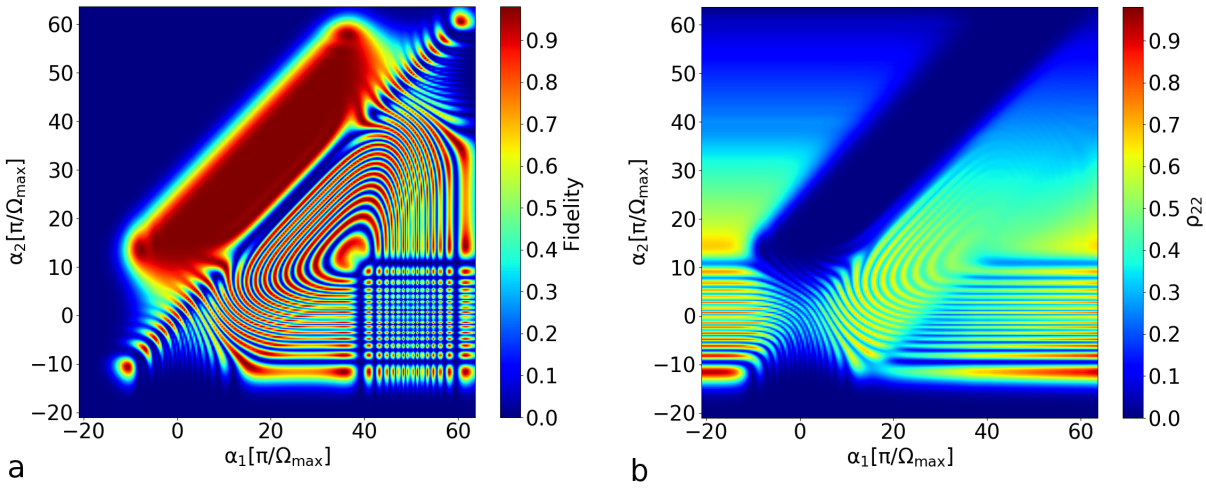


**Figure 5.** a) Value of $\rho_{33}(t_f)$ at the final time $t_f$ after two Gaussian pulses centered at $\alpha_1$ and $\alpha_2$ respectively. b) The integral of $\rho_{22}$ during the process.

return value of a state given an action is determined following a given policy $\pi$, called $Q^{\pi}(s,a)$. Then, if $Q^{\pi^*}(s,a)$, where $\pi^*$ is the optimal policy:

$$\pi^*(s) = \underset{a}{argmax}\ [Q^{\pi}(s,a)]\,. \tag{6}$$

In a policy-based framework, instead, one is interested in directly finding the optimal policy, by minimizing:

$$\boldsymbol{\nabla}_{\theta} J(\theta) = \mathbb{E}_{\pi}\left[G_t \boldsymbol{\nabla}_{\theta} \log \pi_{\theta}\right] \tag{7}$$

where $J$ is a proper loss function and $G_t$ is:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{8}$$

where $\gamma$ is the discount factor and $R_t$ is the reward at time $t$.

In A2C, the so-called advantage function $Q^{\pi^*}(s,a)$ replaces $R_t$. The latter method reduces the variance in the results and it could be also used to upgrade value-based techniques to non-episodic problems. The algorithm operates as in the following. The time evolution of $H$ is divided in $N$ timesteps, with $t \in [0, t_{max}]$. At each time step, the A2C agent can choose 9 discrete actions corresponding to 3 possible actions of both the control gates, namely increase, hold and decrease by an arbitrary quantity $\Delta\Omega$:

$$\Omega_{ij}(t+1) = \begin{cases} \Omega_{ij}(t) - \Delta\Omega \\ \Omega_{ij}(t) \\ \Omega_{ij}(t) + \Delta\Omega \end{cases} . \tag{9}$$

As the next step, the master equation solver implemented in QuTiP is used to evolve the quantum state at time t with the Hamiltonian calculated at step $t+1$, by providing as input the updated pulse values as decided by the agent. During each time step, the Hamiltonian is constant. As both actor and critic, we chose a neural network as a function approximator. The actor network is trained with policy gradient (PG) and it involves a Softmax output layer

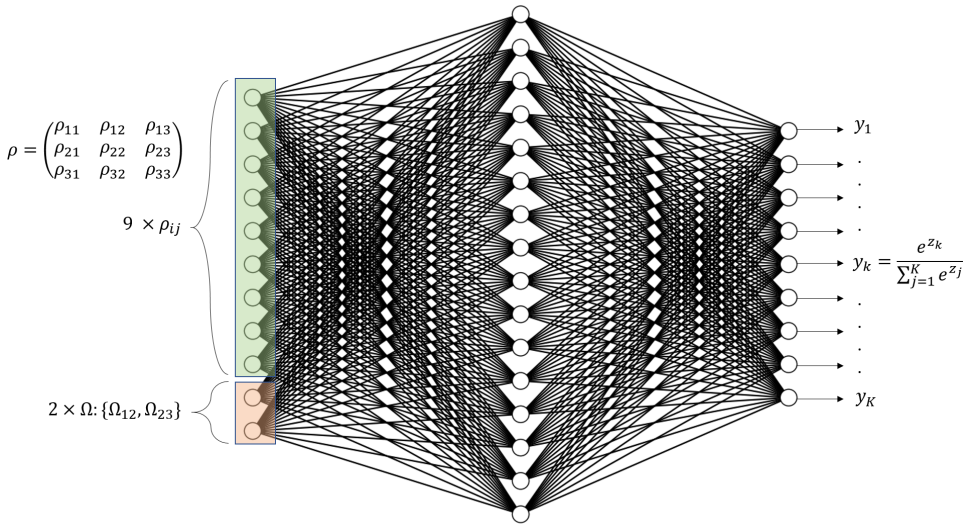


**Figure 6.** The network used to implement the advantage actor-critic algorithm. One hidden layer ($hl$) of 16 neurons is used. The 9 output neurons generate the combinations of possible actions (upward, hold, downward) of the two neurons.

consisting of 9 neurons. The input states of both networks are two 11-dimension vectors (Figure 6), in which the first 9 elements are populated by the density matrix elements, and the last 2 are the values of the two gate pulses during the last time step. By choosing an appropriate reward and by looking at the 3x3 density matrix of the system, the agent can learn which is the next best action. The pseudocode is shown in Table 1.

| **Indent** | **Functions** |
|---|---|
| 1 | for each episode e: |
| 2 | initialize s0=[1.,0.,0.]<br>total reward=0<br>for each time step t: |
| 3 | choose action a with prob. p<br>evolve the quantum state<br>total reward+=reward<br>if done: |
| 4 | train act. & critic nets |
| 5 | break |

**Table 1.** Pseudocode

## 4. Results

We developed an OpenAI-like environment [12] to perform the simulations. For the RL part of the problem, we used Keras with Tensorflow backend [13]. The parameters of the neural networks and the optimizers are listed in Table 2.

The simulations were run on an Intel Core i7-8700K @ 3.7 GHz and a Nvidia Titan Xp 12

| **Architecture** and Settings | **Neural Network** | |
|---|---|---|
| | ***Actor*** | ***Critic*** |
| Layers | (11, 16, 9), Dense | (11,16,1), Dense |
| Activation function | ReLU for the *hl*, Softmax for the output | ReLU for the *hl*, Linear for the output |
| Optimizer | Adam (LR[a] =0.001, default [15] elsewhere) | Adam (LR[a] =0.001, default [15] elsewhere) |
| Loss Function | Categorical Crossentropy | Mean Squared Error (MSE) |
| Initialization | Xavier uniform | He uniform |

[a] LR stands for Learning Rate

**Table 2.** Actor Critic Neural Networks

GB. The reward function used for the results shown in Figure 7 consists of 5 parts:

$$\begin{aligned} R_{t+1} = & (\rho_{33,t} + \rho_{11,t})\Theta(\rho_{33,t} - \rho_{33,t-2}) \\ & -\Theta(\rho_{22,t} - 0.1) \\ & -\Theta(\Omega_{12,t} - \Omega_{12,t-4})\Theta(t - \tfrac{3}{4}t_f) \\ & -\Theta(\Omega_{23,t} - \Omega_{23,t-4})\Theta(t - \tfrac{3}{4}t_f) \\ & +100\,\Theta(\rho_{33,t} - 0.9)\delta_{t,t_f} \end{aligned} \tag{10}$$

where $\Theta$ is the Heaviside function and $\delta$ is Kronecker delta.

The agent gets a positive reward if $\rho_{33}$ rises and if it reaches a fidelity higher than 0.9 in the final step of the evolution. Instead, it gets a punishment when the two pulses keep getting higher near the end of the simulation. We chose $N = 300$ steps for the time evolution, where the last step corresponds to $t = t_f$. When an episode ends, the network's weights are updated. In Figure 7 the artificial intelligence-based analogue of the Figure 3 is shown, consisting of the CTAP of the population and the pulse shape obtained by the neural network. There, we chose a $t_{max} = \frac{20\pi}{\Omega_{max}}$ and $\Delta\Omega = 5 \cdot 10^{-4}\Omega_{max}$. After training of 60000 simulations $\rho_{22}$ remains below 0.035 and $\rho_{33}$ reaches 0.99 with approximately 10 hours of computation by a single thread. Figure 8 shows the results during the training after 30, 300, 20000 and 57000 epochs respectively. The artificial intelligence-based method discovers two pulses obeying the counter-intuitive sequence, like those found by human intuition in the seminal paper of Vitanov et al. [14]. In other words, if no one guessed such method for coherent adiabatic transfer of quantum states, the RL would be able to discover it from scratch by exploring the Hilbert space.

## 5. Conclusion

We have shown how to apply advantage actor-critic (A2C) reinforcement learning methods to achieve control of the coherent transport by adiabatic passage of qubits. We have built an environment consisting of a software simulating the CTAP, which returns known results if the pulses are Gaussian and follows the counter-intuitive sequence, as from literature. Next, we use such an environment to train a network by reinforcement learning, giving no prior knowledge to the network. The A2C method exploits a neural network with a hidden layer of 16 neurons

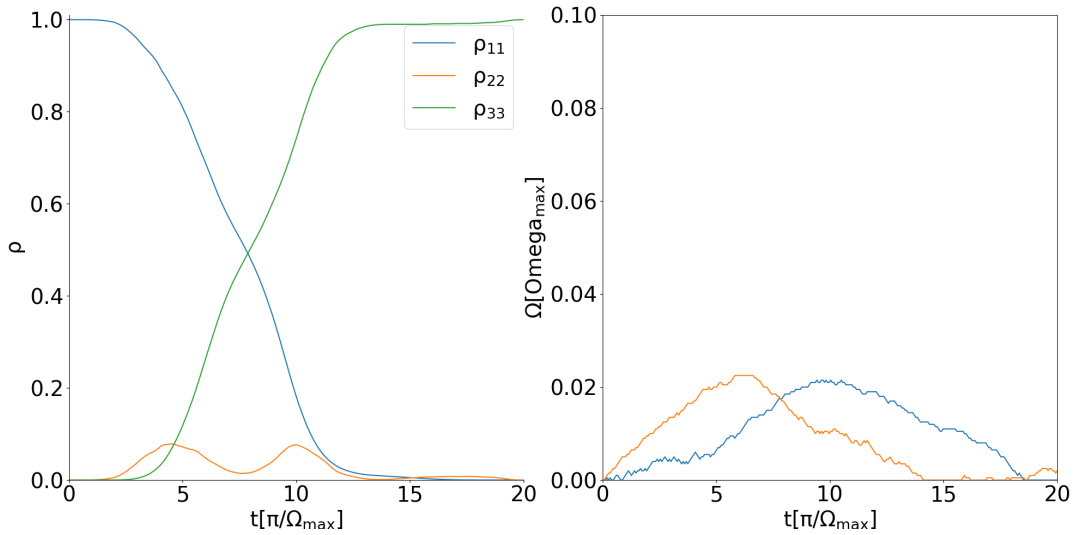


**Figure 7.** a) The CTAP found by reinforcement learning after 60000 epochs. b) The gate pulses found by the actor-critic RL algorithm to achieve the CTAP. The artificial intelligence discovers that the time evolution of the gates must follow a Gaussian-like shape and that the sequence is counter-intuitive.

to maximize the reward consisting of achieving maximum population transfer by keeping null occupation of the central dot in the meanwhile. We therefore demonstrated that RL is capable to discover non-trivial sequences of control gates in quantum information processing, opening the way of calibrating complex and non-ideal systems of qubits.

## Acknowledgments

EP gratefully acknowledges the support of Nvidia Corporation with the donation of the Titan Xp GPU used for this research.

## References

[1] Greentree A D, Cole J H, Hamilton A R, Hollenberg L C L 2004 *Phys. Rev. B* **70** 235317
[2] Maurand R, Jehl X, Kotekar-Patil D, Corna A, Bohuslavskyi H, Laviéville R, Hutin L, Barraud S, Vinet M, Sanquer M, De Franceschi S 2016 *Nat. Comm.* **7** 13575
[3] Ferraro E, De Michielis M, Fanciulli M, Prati E 2015 *Phys. Rev. B* **91** 075435
[4] Prati E, Rotta D, Sebastiano F, Charbon E 2017 *In 2017 IEEE International Conference on Rebooting Computing (ICRC)* 1-4IEEE
[5] Rotta D, De Michielis M, Ferraro E, Fanciulli M, Prati E 2016 *Quant. Inf. Proc.* **15**(6) 2253-2274.
[6] Prati E 2016 *Int. J. of Nanotech.* **13** 7 509
[7] Prati E 2017 *J. of Phys.: Conference Series* **880** 1 012018
[8] Sentís G, Calsamiglia, Munoz-Tapia R, Bagan E 2012 *Sci. Rep. 2* **708**
[9] Porotti R, Tamascelli D, Restelli M, Prati E 2019 arXiv:1901.06603
[10] Fösel T, Tighineau P, Weiss T, Marquardt F 2018 *Phys. Rev. X* **8** 031084
[11] De Michielis M, Ferraro E, Fanciulli M, Prati E 2015 *Journal of Physics A: Math. Theor.* **48** 065304
[12] Rotta D, Sebastiano F, Charbon E, Prati E 2017 *npj Quantum Information* **3** 42
[13] Johansson J R, Nation P D, Nori F 2012 *Comp. Phys. Comm.* **183** 1760
[14] Mnih V *et al* 2015 *Nature* **518** 7540
[15] Sutton R, Barto A 2017 *Reinforcement Learning: An Introduction* (MIT Press)
[16] Szepesvari C 2010 *Algorithms for Reinforcement Learning* (Morgan and Claypool Publishers)
[17] Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W 2016 OpenAI Gym (*Preprint* cs.LG/1606.01540)
[18] Chollet F *et al* 2015 *Keras* https://keras.io

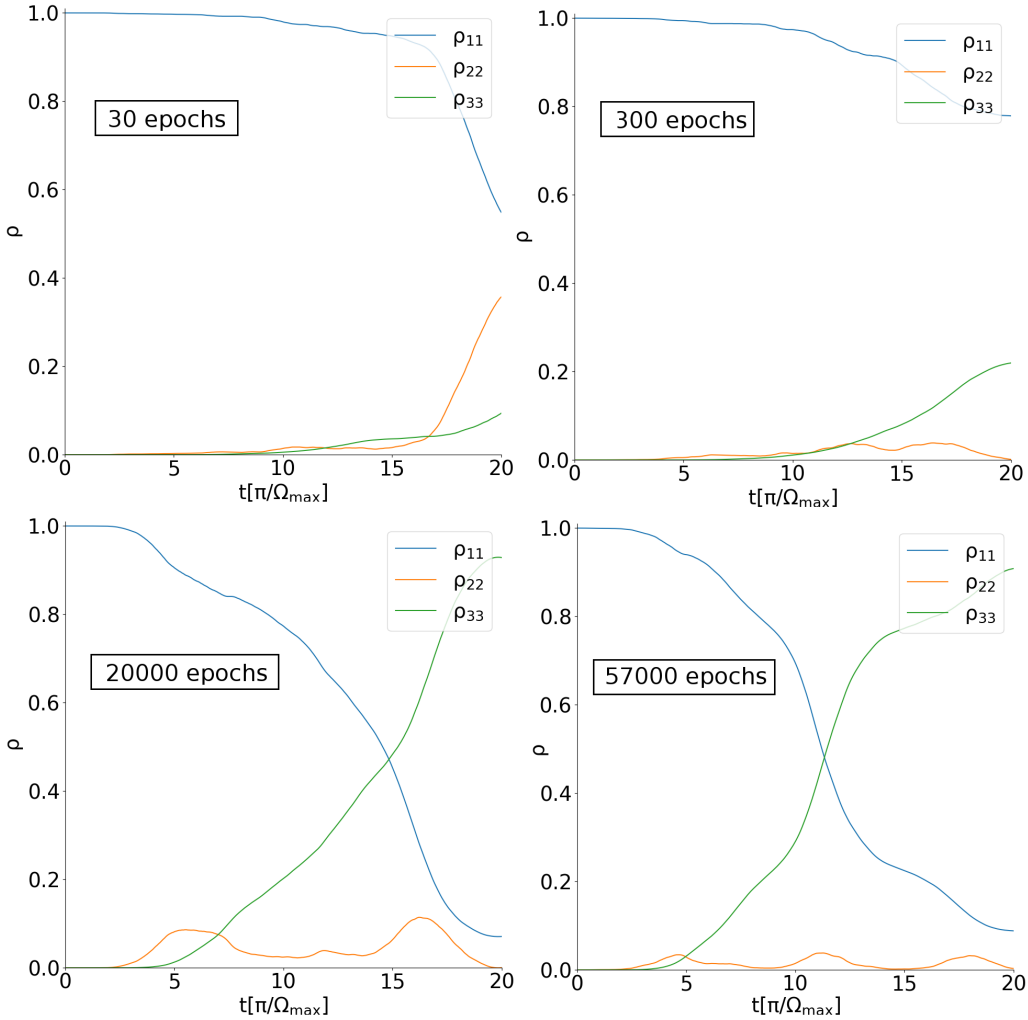


**Figure 8.** Time evolution of the trace elements of the density matrix during the training, after 30, 300, 20000 and 57000 epochs respectively. The training ended at 60000 epochs.

[19] Vitanov N V, Halfmann T, Shore B W, Bergmann K 2001 *Ann. Rev. of Phys. Chem.* **52**
[20] Kingma D P, Ba J 2015 Adam: A Method for Stochastic Optimization (*Preprint* cs.LG/1412.6980)