NEURAL NETWORK SEARCH FOR OPTIMAL PULSE TRAINS FOR QUBIT DYNAMICS CONTROL

Mikhail A. Sergeev,
Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia, ***mas_135@mail.ru***

Marina V. Bastrakova,
Lobachevsky State University of Nizhny Novgorod; Scientific Research Institute of Physics and Technology of Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia, ***bastrakova@phys.unn.ru***

Vsevolod A. Vozhakov,
Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Moscow, Russia, ***sevozh@yandex.ru***

Nikolai V. Klenov,
Lomonosov Moscow State University; Moscow Technical University of Communications and Informatics, Moscow, Russia, ***nvklenov@mail.ru***

Igor I. Soloviev,
Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Moscow, Russia, ***igor.soloviev@gmail.com***

Maxim V. Tereshonok,
Moscow Technical University of Communications and Informatics, Moscow, Russia, ***m.v.tereshonok@mtuci.ru***

DOI: 10.36724/2072-8735-2024-18-7-56-63

Manuscript received 20 May 2024;
Accepted 24 June 2024

Keywords: *quantum computing, quantum hardware control, deep learning, reinforcement learning, AlphaZero, PPO, machine learning, transmon*

Despite its insensitivity to charge noise, the transmon's anharmonicity and control pulse length are limited. Transmon state control traditionally involves a quadrature mixer that mixes microwave signals from a room-temperature oscillator and an arbitrary waveform generator to control the single qubit states. Scaling up quantum processors faces challenges in hardware, management of qubit operations, and read-out procedure due to the large amount of expensive room-temperature equipment required for each qubit. Operating at millikelvin temperatures, these devices introduce thermal noise, reducing qubit lifetime and distorting control signals. An alternative promising control method is based on superconducting digital electronics. In these digital circuits, a bit of information is represented by a short unipolar voltage pulse generated when a single flux quantum (SFQ) pulse passes through a Josephson junction. The qubit states are controlled by the action of a sequence of SFQ pulses, with the pulse-to-pulse timing adjusted to induce a coherent rotation of the state vector in the computational subspace and to minimize leakage to the outside. The paper discusses a method for controlling the states of a transmon qubit using digital superconducting electronics. In this approach, the sequences of picosecond voltage pulses are used to control the state of a quantum computing system. We have considered a control scheme based on a bipolar pulse generator and proposed an algorithm for finding the optimal implementation of a bipolar short pulse control sequence for performing high-precision single-bit operations (with fidelity equal to 99,99 %) using deep learning algorithms with reinforcement: AlphaGo Zero, AlphaZero and Proximal Policy.

Для цитирования:
Сергеев М.А., Бастракова М.В., Вожаков В.А., Кленов Н.В., Соловьев И.И., Терешонок М.В. Нейросетевой поиск оптимальных последовательностей импульсов для управления динамикой кубитов // T-Comm: Телекоммуникации и транспорт. 2024. Том 18. №7. С. 56-63.

For citation:
Sergeev M.A., Bastrakova M.V., Vozhakov V.A., Klenov N.V., Soloviev I.I., Tereshonok M.V. (2024). Neural network search for optimal pulse trains for qubit dynamics control. *T-Comm*, vol. 18, no.7, pp. 56-63.

Introduction

The transmon qubit is currently the most popular solution for superconducting qubits but managing this type of qubit requires complex and expensive equipment [1]. Transmon is an LC-oscillator in a quantum regime with a Josephson contact replacing the inductance, adding anharmonicity to the system and allowing effective separation of the low-lying levels for computational purposes [2]. Despite its insensitivity to charge noise, the transmon's anharmonicity and control pulse length are limited. Transmon state control traditionally involves a quadrature mixer that mixes microwave signals from a room-temperature oscillator and an arbitrary waveform generator to control the single qubit states. Scaling up quantum processors faces challenges in hardware, management of qubit operations, and read-out procedure due to the large amount of expensive room-temperature equipment required for each qubit. Operating at millikelvin temperatures, these devices introduce thermal noise, reducing qubit lifetime and distorting control signals.

An alternative promising control method is based on superconducting digital electronics [3,4]. In these digital circuits, a bit of information is represented by a short unipolar voltage pulse generated when a single flux quantum (SFQ) pulse passes through a Josephson junction. The qubit states are controlled by the action of a sequence of SFQ pulses, with the pulse-to-pulse timing adjusted to induce a coherent rotation of the state vector in the computational subspace and to minimize leakage to the outside [5]. The realization of this approach has already been experimentally demonstrated [6]. In the so-called scalable leakage optimized pulse sequence (SCALLOPS) [7], more than one picosecond pulse is used during the oscillation period of the qubit, which allows to significantly speed up the quantum operations.

Following this SCALLOPS idea, another approach using more complex sequences based on pulses of different polarities has also been proposed [8]. This control approach is mainly reduced to the task of optimizing quantum dynamic functionals by searching for control sequences based on some input data (qubit frequencies, nonlinearities, oscillator frequencies). It is well known that the vast majority of quantum dynamics optimization algorithms are based on heuristic or purely stochastic approaches (e.g., coordinate descent or genetic algorithms). A major limitation of such algorithms is the strong dependence of the final accuracy on the hyperparameters, as well as the need to completely reapply the algorithm even for small changes in the system parameters. In turn, machine learning using neural networks and reinforcement learning algorithms allows to achieve high accuracy of results and at the same time has the ability to change it when changing system parameters even without the need for complete retraining [9].

This paper treats the control of the qubit state as a combinatorial optimization problem using reinforcement learning algorithms. However, the problem is still quite hard for these types of algorithms to solve because the accuracy of quantum operations is very sensitive to small changes in the pulse sequences, making the optimization task much harder due to the narrow local optimum. This was the reason why the Zero algorithm family was the first choice for the optimization algorithm [10] (it will be discussed in Ch.2). Nevertheless, the algorithm changed drastically during the training process, and the algorithm updates as well as the reasons for them are given in Sec. 3, including the transition of the reward function from leakage to non-computational qubit basis to quantum fidelity (both terms are introduced in Sec. 1). Finally, in the results section, we present found SFQ pulse sequences for different qubit parameters, all of which satisfy the accuracy criterion of a single qubit operation of 99,99% (fidelity).

1. Model of the syst

For the transmon state control under the action of SFQ pulses simulation, we implement a mathematical model of the system using dc/SFQ-converter as the pulse generator [11]. The transmon is a modified version of a Cooper Pair Box (CPB) [2], containing a superconducting “island” connected to the rest of the circuit only by a Josephson tunneling current, shunted by the capacitance C_Q, see Figure 1. The coupling with the dc/SFQ-converter is capacitive through C_c.

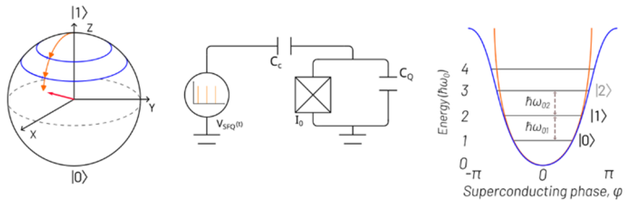

Fig. 1. Schematic representation of the transmon spectrum from [2] (right), capacitively coupled through C_c to the short pulse generator (central region). Schematic representation of the SFQ pulses effect on a qubit using Bloch sphere (left)

In transmons, the capacitive energy $E_C = e^2/2C$ (where $C = C_C + C_Q$) is much smaller than the Josephson energy E_J, namely, the energy stored in a Josephson junction when a supercurrent flows through it, which can also be corrected in situ by applying an external magnetic flux to the circuit. The transmon Hamiltonian including the electrostatic and Josephson contributions can be written in the following form:

$$H = 4E_c(\hat{n} - n_g) - E_J(1 - cos\hat{\phi}), \quad (1)$$

where $\hat{n} = -i\frac{\partial}{\partial\varphi}$ is the number of Cooper pairs in charge units $2e$, φ is the phase operator, and $n_g = C_c V(t)/2e$ is called the effective charge displacement. The commutation relation on the operators is satisfied, considering $\hbar = 1$.

The SFQ control pulse has a picosecond duration τ, while its voltage integral over time is equal to the magnetic flux quantum $\Phi_0 = \int_0^\tau V(t)dt$. Then, carrying out the quantization procedure, we define the creation $\hat{a}^\dagger$ and annihilation operators $\hat{a}$ in terms of fluctuations of charge and phase

$$\hat{n} = \frac{-i}{2}\left(\frac{E_J}{2E_c}\right)^{\frac{-1}{4}}(\hat{a} - \hat{a}^\dagger), \hat{\phi} = \frac{-i}{2}\left(\frac{2E_c}{E_J}\right)^{\frac{-1}{4}}(\hat{a} + \hat{a}^\dagger) \quad (2)$$

Substituting the expressions (2) into Eq. (1) and averaging over the rapidly oscillating summands one can obtain the Hamiltonian of the qubit in the representation of the Bose-Hubbard model:

$$H = \omega_{01}\hat{a}^\dagger\hat{a} + \alpha(\hat{a}^\dagger + \hat{a})^4 - i\varepsilon(t)(\hat{a}^\dagger - \hat{a}), \quad (3)$$

where $\omega_{01} = \sqrt{8E_J E_c}$ is the transition frequency between the ground state $0\rangle$ and the first excited state $1\rangle$, $\alpha = \frac{-E_c}{12}$ is the

nonlinearity parameter, and the control field is defined as $\varepsilon(t) = \frac{C_c V(t)}{2}\sqrt{\frac{\omega_{01}}{2C_Q}}$. Thus, a single SFQ pulse in the sequence induces a discrete small rotation by an angle of

$$\Delta\theta = C_c\Phi_0\sqrt{\frac{\omega_{01}}{2C_Q}} \quad (4)$$

on the Bloch sphere corresponding to the change of the qubit state (see Fig.1).

In this case, the evolution of the qubit states (3) is defined as

$$\psi(t)\rangle = U(t)\psi_0\rangle, U(t) = Pe^{-i\int_0^t H(t')dt'}, \quad (5)$$

where P is the chronological ordering operator, $\psi_0\rangle$ is the initial state of the qubit. The population of levels of the system is defined as:

$$W_m(t) = |\langle\psi(t)\rangle|^2, m = 0,1,2 \ldots \quad (6)$$

to calculate the fidelity of the quantum operation

$$\langle F\rangle = \frac{1}{6}\sum_{\psi_0\rangle}|\langle\psi_0\rangle|^2 \quad (7)$$

averaged over initial states at different poles of the Bloch sphere $|\psi_0\rangle = \{x_\pm\rangle, \vee\, y_\pm\rangle, \vee\, z_\pm\rangle\}$:

$$|x_\pm\rangle = \frac{0\rangle \pm 1\rangle}{\sqrt{2}} \quad (8)$$

In (7) U_g is the gate matrix in the realization of ideal quantum operations. For example, for the operation $Y_{\frac{\pi}{2}}$: $U_g = \frac{1}{\sqrt{2}}$.

We assume that the sequences can be represented as short and small amplitude SFQ pulses that can be applied at the clock frequency of a $2\pi/\omega_{(gen)}$ high-frequency oscillator. The spectrum of a single SFQ pulse is broad and constant in the frequency range of interest. Thus, the spectrum of the whole sequence depends only on the structure of the sequence itself, and the shape of the single pulse is not significant [12].

By using SFQ pulses of different polarity and extending the space of possible pulse states to bipolar [13], we can approximately double the pulse density with respect to the unipolar SCALLOP approach [14]. Since the operation time is significantly shorter than the relaxation and dephasing time, the main source of error will be the leakage to the higher qubit states (beyond the basis levels). In this case, an important challenge is to develop algorithms that optimize the control sequence structure to reduce this leakage, and consequently improve the accuracy of the single-qubit operation itself. Unlike the work of [15], we do not restrict the possibility of arranging the pulses solely as time-symmetric pairs. In addition, we envision the possibility of applying a "virtual Z-gate" [16] after the gating sequence to achieve the desired phase of the wave function.

2. AlphaZero neural network algorithm for SFQ search of pulse sequences

First of all, we would like to affirm the choice of AlphaGoZero algorithm [10] as an initial approach for the problem. Since the pulse sequence for qubit control is discrete, and the action space is narrow (only 3 actions for every possible pulse in every sequence index), we can reduce the types of reviewed algorithms to discrete optimization. One of the most distinguishable attributes of AlphaZero algorithm is that it's a model-based algorithm (in contrast to most RL algorithms being model-free), specifically, utilizing the access to training environment to boost its training capabilities. Contrary to that, the vast majority of (model-free) algorithms for quantum dynamics optimization are based on heuristic or purely stochastic approaches (e.g., the method of coordinate descent or genetic algorithms [17]). For such algorithm types, a significant limitation is the strong dependence of the final accuracy on the initial assumptions (hyperparameters), as well as the need for complete reapplication of the algorithm even for the small system parameter changes. In turn, machine learning using neural networks and reinforcement learning algorithms, some of which are AlphaGo Zero and AlphaZero [18], can achieve high accuracy results when the system parameters are changed, even without requiring a complete re-training.

In this paper, we have chosen a kind of convolutional neural network called ResNet (Residual neural Network) [19]. Its distinctive feature is that in some layers, in addition to the output data from the previous layer, output data from several layers backward are also fed to their input (see Fig.2). This approach ensures high accuracy of processing complex patterns and prevents the neural network from losing important data.

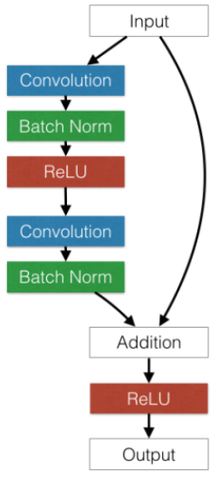

Fig. 2. Schematic representation of ResNet single block used in the AlphaZero algorithm

In the AlphaZero and AlphaGo Zero algorithms, the neural network is required to predict Monte Carlo tree search (MCTS) policies, which requires predicting action probabilities and average empirical rewards for each of the states. Technically, this requirement is expressed as having two outputs (reward/policy) originating from the same set of residual layers.

3. Steps taken in the development and modification of the algorithm

Largely, the development of the algorithm for optimizing SFQ bipolar sequences can be described in several stages:

i. Implementing leakage to non-computational levels as a loss function.
ii. MCTS search alteration.
iii. Transition from AlphaGo Zero to AlphaZero.
iv. Modification of the reward function.
v. Transition to PPO algorithm.

i. Implementing leakage to non-computational levels as a loss function

Selecting a reward metric that aligns with the specific goals and objectives of the reinforcement learning problem can greatly impact the overall success of the algorithm. By focusing on a physics-based reward metric, we are able to capture the intricacies of the system and optimize the learning process better. In this case, prioritizing the reduction of leakage to non-computational levels $W_{n\geq 2}(t) < 10^{-5}$ as the primary reward metric allows for a more effective training process. This strategic decision has resulted in quickly converging algorithm, demonstrating the importance of selecting the right reward metric for optimal performance.

ii. Modification of the MCTS structure

The MCTS structure used in this work has undergone several changes compared to the original method presented in [18]: each Monte-Carlo tree search consists of searching a leaf of the tree, possibly expanding the leaf, and performing a backup. The main bottleneck here is the expansion operation, which requires the use of a neural network to obtain a priori probabilities of actions and the estimated value of the game. To make this expansion more efficient, we use minibatches when searching multiple leaves, but then perform the expansion in a single evaluation execution from the neural network, as shown in Fig.4. This approach has one drawback: since multiple MCTSs are executed in a single batch, the result is not the same as when they are executed sequentially. Indeed, initially, when we have no nodes stored in the MCTS class, our first search will expand the root node, the second will expand some of its child nodes, and so on. Initially, however, a single batch of searches can only expand a single root node. Of course, later individual searches in a batch can follow different paths and expand. However, initially, mini-batch expansion is much less efficient than sequential MCTS. To compensate for this, mini-batches are still used, but now multiple MCTSs are performed. This approach provides more extensive exploration of states at the beginning of training, speeding up the learning process.

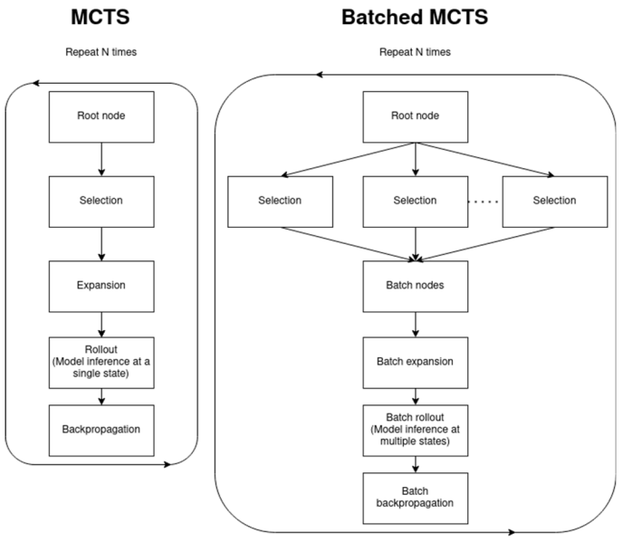

Fig. 3. Conventional MCTS structure (left) and modified (batched) MCTS structure (right)

iii. Transition from AlphaGo Zero to AlphaZero

The next step was to transition the structure of the algorithm from AlphaGo Zero to AlphaZero. The need for this change was due to the fact that AlphaGo Zero has encountered several issues during the training process. First of all, it turned out that the algorithm was excessively data-demanding, requiring a huge amount of training steps just to collect enough training data from the environment, losing its computational efficiency and taking more time to produce pulse sequences. In addition, this iteration of the training algorithm is sensitive to hyperparameters since it demands two neural networks (and two different policies) to train, making the training a difficult process, minimizing the advantage in comparison with other gradient-free and metaheuristic algorithms. To avoid these issues, a transition to a more adaptive version of this algorithm (AlphaZero) had to be done.

Several differences between AlphaGo Zero and AlphaGo should be noted:

– Supervised learning is not used to initialize policies.

– A single policy is used.

– A single network is used for both the policy and the value function.

– Raw state configurations are fed into the network instead of manually created functions.

In an effort to remove a priori probabilities (the assumption of random variable distribution) from the algorithm, Google DeepMind released AlphaZero [18]. AlphaZero retains the same model and overall learning process as AlphaGo Zero, but removes some components that don't transfer well to other games. Notable changes include:

– Removal of data augmentation (increasing the number of states in memory by modifying current states), which AlphaGo Zero did because it would have created impossible configurations for many environments.

– The course of the independent game now involves a single neural network instead of two (best player and student).

– AlphaZero reuses the same hyperparameters except for one (the study depends on the number of moves allowed) for all games.

In addition to switching to a newer version of the algorithm, parallelization of the process of collecting information about the environment during training (self-play) was also implemented, which allowed for faster training. Also, for some of the hyperparameters we introduced their dynamic variation: for the constant of the neural network learning rate, we introduced exponential fading with a variable coefficient.

iv. Modification of the reward function

Despite the fact that leakage can be considered a well-fitting physics-based metric for this problem, it does not fully reflect all the aspects of the quantum operation (e.g., it does not include the phase of the state and its changes). We have switched to the operation fidelity <F> as a reward value measure as it better represents the chosen metric as the control pulse quality measure. In addition, we also implemented a reward shaping procedure. It involved adjusting the reward potential in the form of a logarithmic value $R = \log(1 - F)$ to provide a smoother transition in the region where $(1 – F) = 10^{-4}$. This adjustment allowed the algorithm to incrementally increase the final result to an average of 99,99%, a slight improvement from the 99,95% achieved over the previous iteration.

Additionally, a reward discount was introduced based on each pulse index inside the sequence. However, this led to the agent experiencing overtraining early in the sequence construction process. This caused the agent to pay too much attention to the amplitude of the state and to not to take state phase into account in the beginning of the sequence, computing additional phases at the end of the operation, negatively impacting the overall efficiency and effectiveness of the algorithm.

v. ansition to PPO algorithm

The transition from AlphaZero algorithm was conditioned by several reasons. The algorithmic simplicity of AlphaZero is sacrificed for the power of state exploring and better system dynamic prediction provided by the MCTS search. In addition, due to the fidelity value sensitivity to the even slight changes in pulse sequence, training stability is meandering, negatively affecting steadiness of network inference. And, finally, the overall complexity of AlphaZero algorithm makes it prone to overfitting under the circumstances of this work, namely, the relative simplicity of the agent environment. On the other hand, Proximal Policy Optimization (PPO) is a recent advancement in reinforcement learning that provides an improvement in trust region policy optimization (TRPO). This algorithm was proposed in 2017 and has shown remarkable performance [20].

The magnitude of policy update will be limited to a small region to avoid huge updates that could potentially be detrimental to the learning process. In other words, PPO behaves exactly like other policy gradient methods, in the sense that it also involves calculating probabilities of forward pass outcomes based on various parameters and computing gradients to improve these decisions or backward pass probabilities. It involves using an important sampling factor like its predecessor, TRPO. However, it also ensures that the old policy and the new policy are at least at a certain proximity (denoted by ε), and very large updates are not allowed. The reliability and computational efficiency of this algorithm, including massive parallelization options, greatly fits in the conditions of this work.

Besides, the ResNet neural network architecture was replaced by multi-layer perceptron (MLP), which has significantly improved training speed without noticeable loss in results due to extensive hyperparameter calculation.

The hyperparameters were adjusted both for the AlphaGo Zero, AlphaZero and PPO algorithm implementation. Since PPO shows better learning dynamics, this work presents a visualization of the dynamics of the reward function in response to hyperparameter tuning, as shown in Fig.4, and the best run for a specific set of parameters is shown in Fig.5. The hyperparameters for PPO algorithm [20] varied in the following ranges:

- *decreasing learning rate (LRann)* [True, False] – enable/disable learning rate decrease over training steps;
- *entropy coef (Entropy)* [0,001; 0,001; 0,01] –entropy coefficient in PPO loss function;
- *GAE lambda (GAElam)* [0,900; 0,905;… 0,990]– lambda coefficient in general advantage estimation algorithm;
- *gamma (Gamma)* [0,900; 0,905;… 0,990] – discount factor;
- *layer size (LS)* [64; 128; 256] – number of neurons inside each of two layers of MLP;
- *total steps (Steps)* [3×10^5; 35×10^4;… 15×10^5] – total amount of training steps (learning rate decrease rate depends on this parameter);
- *epochs* [3; 4;... 10] – the number of backpropagation iterations during each training step;
- *value coef* [0,25; 0,30;... 0,75] – value coefficient in PPO loss function;
- Adam gradient descent algorithm parameters were fixed according to [21].

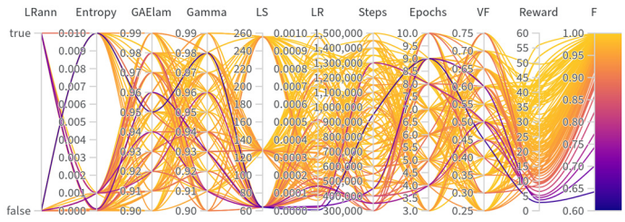

Fig. 4. Hyperparameter search for PPO algorithm. Each search run is represented with a curve with its own distinctive color. Brighter colors correspond to better fidelity values at the end of the training

As shown in Figure 4, by systematically varying the values of key hyperparameters and observing how they affect the overall performance of the algorithm, one can gain valuable insight into how to optimize models for better results. This process involves running multiple experiments and recording the corresponding performance metrics, in this case fidelity. Overall, strong hyperparameter dependence is a crucial aspect of deep reinforcement learning, since even small changes in hyperparameters can significantly improve or degrade training stability.

4. Results

In our analysis of the developed neural network algorithms (AlphaGo Zero, AlphaZero, and PPO) for tackling the optimization problem of calculating the quantum dynamics of a qubit, we have found interesting insights that are showcased in Figure 5. The comparison of their performance is crucial in understanding how well each algorithm fares in achieving the desired outcomes.

The central performance metric used in this comparison is infidelity (1 - <F>) from the training time, a convenient measure to compare different learning algorithms. It becomes apparent from our analysis that both AlphaGo Zero and AlphaZero algorithms fall short of the required accuracy level, with (1 - <F>) staying above 0,0001. In contrast, PPO emerges as the standout performer, consistently achieving the desired precision level.

This disparity in performance highlights the varying capabilities of these algorithms when tasked with the complexity of quantum dynamics calculations. While AlphaGo Zero and AlphaZero show promise, it is evident that PPO excels in delivering accurate results. This comparative analysis serves as a valuable reference point for researchers and practitioners looking to leverage neural network algorithms in quantum computing applications.

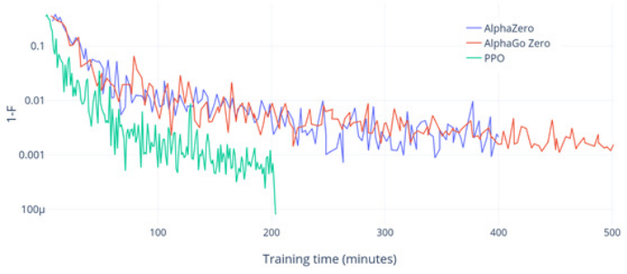

Fig. 5. Performance comparison for AlphaGo Zero (blue), AlphaZero (red) and PPO (green) algorithms, measured in fidelity change over averaged training time (in minutes). Qubit parameters correspond to that of row 3 in Figure 6

Figure 6 and Table 1 present the results of search of control sequences for the parameters of the system (2) under consideration, for example, an operation $Y_{\frac{\pi}{2}}$.

Table 1

Qubit parameters configurations and corresponding 1-F for the best pulse sequence found for each parameter set. All of the sequences were produced by the PPO algorithm

№	$\omega_{01}/2\pi$, GHz	$\Omega_{gen}/2\pi$, GHz	$\mu/2\pi$, GHz	$\Delta\theta$, rad	$1-F$, 10^{-5}
1	5	25	0,35	0,024	4
2	4	25	0,25	0,024	8
3	3	25	0,25	0,024	8
4	5	25	0,25	0,032	7
5	5	25	0,5	0,024	5
6	5	25	0,25	0,032	8
7	5	25	0,3	0,024	8
8	7	25	0,25	0,024	9
9	5	30	0,25	0,024	6
10	5	25	0,25	0,024	9
11	5	25	0,4	0,024	7
12	5	25	0,45	0,024	8
13	5	45	0,25	0,024	6
14	5	35	0,25	0,024	7
15	5	25	0,25	0,021	6
16	5	25	0,25	0,03	7
17	6	25	0,25	0,024	7
18	5	40	0,25	0,024	9

Figure 6 presents the results of search of control sequences for the parameters of the system (2) under consideration, for example, an operation $Y_{\frac{\pi}{2}}$.

Conclusion

To summarize the impact on model performance, the key reasons for the outcome should be reviewed sequentially. While the most important contribution to model performance may not be clearly defined, in this case, changes in the reward metric had the most impact on the result. The main problem of using regular fidelity as a reward is that it has a noticeable drawback in reward responsiveness to small changes in fidelity. Since our goal was to achieve a high fidelity value, using plain fidelity as a reward value generates only small weight updates, which leads to instability of the training and makes the algorithms particularly susceptible to hyperparameter changes.

This is exacerbated by the fact that the fidelity value has only narrow local optima with small changes in the value itself. Conversely, the design of the reward function can take into account these peculiarities of the given system and improve the training process of the algorithm. Nevertheless, the algorithm itself had to be changed from AlphaGo Zero to AlphaZero to PPO because our initial approach to solving this problem was too complex and therefore more data dependent and prone to overfitting. Each step in this process was intended to maintain enough algorithmic complexity to handle this task, while being as simple as possible for effective computation and generalization. While the comparison of all three algorithms seems irrelevant due to the fact that their performance on any task can mostly be investigated empirically, in this work the PPO algorithm showed the most balanced approach to tackle the transmon SFQ control problem.

Thus, by comparing different algorithms and approaches to this task, we could state that model-free and on-policy RL algorithms are suitable for optimization tasks with narrow local optima and complex behavioral dynamics, which is proven on an example of our task. Model-based algorithms can cover more complex dynamics in the system, but it turned out that this problem did not require additional planning to solve, so model-free approach was considered the most useful in this case.

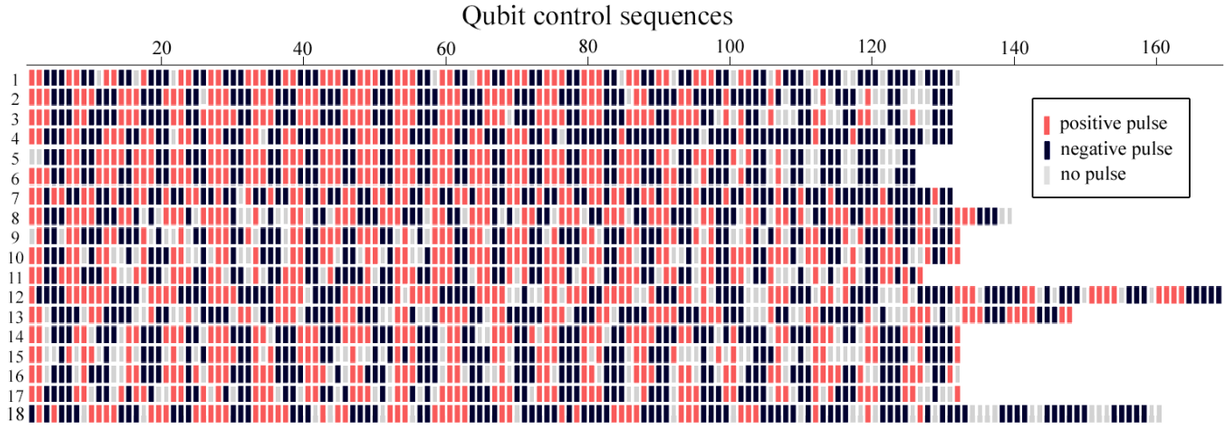

Fig. 6. Pulse sequence visualizations for different qubit parameters. All sequences satisfy the criterion of F > 99,99%

The on-policy approach, on the other hand, allowed stable policy updates to be maintained in order to reach a narrow local optimum without detrimental policy changes.

However, there is still some room for improvement. The accuracy of the algorithm can be increased by additional tweaks (e.g. setting the number of layers in the MLP used in the algorithm as a hyperparameter) and by optimizing the training process both algorithmically and computationally. Additionally, some augmentation can be added to increase the training data diversity and therefore to achieve the high level of robustness [22]. In the future, the algorithm will be trained to generalize over multiple values of qubit parameters. It will help to achieve sustainable results for a wide variety of experimental setups without the need to train separate networks for each parameter set.

Funding

The work was carried out with the support of the Grant of the Russian Science Foundation No. 22-72-10075.

References

1. *Vozhakov V.A., Bastrakova M.V., Klenov N.V., Soloviev I.I., Pogosov W.V., Babukhin D.V., Zhukov A.A., Satanin A.M.* State 378 control in superconducting quantum processors // Phys.-Uspekhi, no. 65, pp. 457-476. 2022.

2. *Koch J., Terri M.Y., Gambetta J., Houck A.A., Schuster D.I., Majer J., Blais A., Devoret M.H., Girvin S.M., Schoelkopf R.J.* Charge-insensitive qubit design derived from the cooper pair box // Physical Review A 76, 042319. 2007.

3. *Soloviev I.I., Klenov N.V., Bakurskiy S.V., Kupriyanov M.Y., Gudkov A.L., Sidorenko A.S, Beilstein J.* Beyond Moore's technologies: operation principles of a superconductor alternative // Nanotechnol, no. 8, pp. 2689-2710. 2017.

4. *Cuthbert M., DeBenedictis E., Fagaly R.L., Fagas G., Febvre P., Fourie C., Frank M., Gupta D., Herr A., Holmes D.S., Humble T., de Escobar A.L., Mueller P., Mukhanov O., Nemoto K., Rao S.P., Pelucchi E.., Plourde B, Soloviev I., Tzimpragos G., Vogelsang T., Yoshikawa N., You L.* 2022 IRDS. 2022.

5. *McDermott R., Vavilov M.G.* Accurate Qubit Control with Single Flux Quantum Pulses // Physical Review Applied. 2, 1. 2014.

6. *Opremcak A., Pechenezhskiy I.V., Howington C., Christensen B.G., Beck M.A., Leonard E., Suttle J., Wilen C., Nesterov K.N., Ribeill G.J., Thorbeck T., Schlenker F., Vavilov M.G., Plourde B.L., McDermott R.* Measurement of a Superconducting Qubit with a Microwave Photon Counter, // Science 1242, 1239. 2018; C.H. Liu et al. PRX QUANTUM 4, 030310. 2023.

7. *Li K., McDermott R., Vavilov M.G.* // Phys. Rev. Appl. 12 014044. 2019.

8. *Vsevolod Vozhakov* et al. Speeding up qubit control with bipolar single-flux-quantum pulse sequences // Quantum Sci. Technol. 8 035024. 2023.

9. *Dalgaard M., Motzoi F., Sørensen, J.J., Sherson J.* Global optimization of quantum dynamics with AlphaZero deep exploration // NPJ quantum information, no. 6(1), 6. 2020.

10. *Silver D., Schrittwieser J., Simonyan K. et al.* Mastering the game of Go without human knowledge // Nature, no. 550, pp. 354-359. 2017.

11. *Likharev K.K., Semenov V.K.* 1991 IEEE Trans. Appl. Supercond. 1, pp. 3-28.

12. *McDermott R., Vavilov M.G.* Accurate qubit control with single flux quantum pulses // Physical Review Applied, no. 2(1), 014007. 2014.

13. *Vsevolod Vozhakov* et al. 2023 Quantum Sci. Technol. 8 035024.

14. *Li K., McDermott R., Vavilov M.G.* 2019 Phys. Rev. Appl. 12 014044.

15. *Bowdrey M.D., Oi D.K., Short A.J., Banaszek K., Jones J.A.* Fidelity of single qubit maps // Physics Letters A, no. 294(5-6), pp. 258-260. 2002.

16. *Likharev K.K., Semenov V.K.* RSFQ logic/memory family: A new Josephson-junction technology for sub-terahertz-clock-frequency digital systems // IEEE Transactions on Applied Superconductivity, no. 1(1), pp. 3-28. 1991.

17. *Dalgaard M., Motzoi F., Sørensen J.J., Sherson J.* Global optimization of quantum dynamics with AlphaZero deep exploration // NPJ quantum information, no. 6(1), 6. 2020.

18. *Silver David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, L. Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, Demis Hassabis*. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. ArXiv abs/1712.01815. 2017.

19. *He K., Zhang X., Ren S., Sun J.* Identity mappings in deep residual networks // Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV 14, pp. 630-645. Springer International Publishing.

20. *Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O.* Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347. 2017.

21. *Kingma D.P., Ba J.* Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014.

22. *Ziyadinov V., Tereshonok M.* Low-Pass Image Filtering to Achieve Adversarial Robustness // *Sensors*. 2023. Vol. 23, no. 22, pp. 9032. DOI 10.3390/s23229032.

НЕЙРОСЕТЕВОЙ ПОИСК ОПТИМАЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ИМПУЛЬСОВ ДЛЯ УПРАВЛЕНИЯ ДИНАМИКОЙ КУБИТОВ

Сергеев Михаил Александрович, *Национальный исследовательский нижегородский государственный университет им. Н. И. Лобачевского, Нижний Новгород, Россия,* ***mas_135@mail.ru***

Бастракова Марина Валерьевна, *Национальный исследовательский нижегородский государственный университет им. Н. И. Лобачевского, Нижний Новгород, Россия,* ***bastrakova@phys.unn.ru***

Вожаков Всеволод Андреевич, *Научно-исследовательский институт ядерной физики имени Д.В. Скобельцына Московского государственного университета имени М.В. Ломоносова, Москва, Россия,* ***sevozh@yandex.ru***

Кленов Николай Викторович, *Физический факультет Московского государственного университета имени М.В. Ломоносова, Москва, Россия,* ***nvklenov@mail.ru***

Соловьев Игорь Игоревич, *Научно-исследовательский институт ядерной физики имени Д.В. Скобельцына Московского государственного университета имени М.В. Ломоносова, Москва, Россия,* ***igor.soloviev@gmail.com***

Терешонок Максим Валерьевич, *Московский технический университет связи и информатики, Москва, Россия,* ***m.v.tereshonok@mtuci.ru***

Аннотация

В статье обсуждается метод управления состояниями кубита-трансмона с использованием цифровой сверхпроводниковой электроники. В этом подходе последовательности пикосекундных импульсов напряжения используются для управления состоянием квантовой вычислительной системы. Рассмотрена схема управления на основе генератора биполярных импульсов; предложен алгоритм поиска оптимальной реализации управляющей последовательности биполярных коротких импульсов для выполнения высокоточных однобитовых операций (с точностью 99,99 %) с использованием алгоритмов глубокого обучения с подкреплением AlphaGo Zero, AlphaZero и Proximal Policy.

Ключевые слова: *квантовые вычисления, управление квантовыми устройствами, глубокое обучение, обучение с подкреплением, AlphaZero, PPO, машинное обучение, трансмон.*

Литература

1. *Vozhakov V.A., Bastrakova M.V., Klenov N.V., Soloviev I.I. Pogosov W.V., Babukhin D.V., Zhukov A.A., Satanin A.M.* State 378 control in superconducting quantum processors // Phys.-Uspekhi, no. 65, pp. 457-476. 2022.
2. *Koch J., Terri M.Y., Gambetta J., Houck A.A., Schuster D.I., Majer J., Blais A., Devoret M.H., Girvin S.M., Schoelkopf R.J.* Charge-insensitive qubit design derived from the cooper pair box // Physical Review, no. A 76, 042319. 2007.
3. *Soloviev I.I., Klenov N.V., Bakurskiy S.V., Kupriyanov M. ., Gudkov A.L. and Sidorenko A.S. Beilstein J.* Beyond Moore's technologies: operation principles of a superconductor alternative // Nanotechnol, no. 8, pp. 2689-2710. 2017.
4. *Cuthbert M., DeBenedictis E., Fagaly R.L., Fagas G., Febvre P., Fourie C., Frank M., Gupta D., Herr A., Holmes D.S., Humble T., de Escobar A.L., Mueller P., Mukhanov O., Nemoto K., Rao S.P., Pelucchi E., Plourde B., Soloviev I., Tzimpragos G., Vogelsang T., Yoshikawa N., You L.* 2022 IRDS. 2022.
5. *McDermott R., Vavilov M.G.* Accurate Qubit Control with Single Flux Quantum Pulses // Physical Review Applied 2, 1. 2014.
6. *Opremcak A., Pechenezhskiy I.V., Howington C., Christensen B.G., Beck M.A., Leonard E., Suttle J., Wilen C., Nesterov K.N., Ribeill G.J., Thorbeck, F. Schlenker, M.G. Vavilov, B.L. Plourde, R. McDermott.* Measurement of a Superconducting Qubit with a Microwave Photon Counter T. Science 1242, 1239. 2018.; C.H. Liu et al. PRX QUANTUM 4, 030310. 2023.
7. *Li K, McDermott R., Vavilov M.G.* 2019 Phys. Rev. Appl. 12 014044.
8. *Vsevolod Vozhakov* et al. Speeding up qubit control with bipolar single-flux-quantum pulse sequences // Quantum Sci. Technol. 8 035024. 2023.
9. *Dalgaard M., Motzoi F., Sorensen J.J., & Sherson J.* Global optimization of quantum dynamics with AlphaZero deep exploration // NPJ quantum information, no. 6(1), 6. 2020.
10. *Silver D., Schrittwieser J., Simonyan K.* et al. Mastering the game of Go without human knowledge // Nature 550, pp. 354-359. 2017.
11. *Likharev K.K., Semenov V.K.* 1991 IEEE Trans. Appl. Supercond. 1, pp. 3-28.
12. *McDermott R., Vavilov M.G.* Accurate qubit control with single flux quantum pulses // Physical Review Applied, no. 2(1), 014007. 2014.
13. *Vsevolod Vozhakov* et al. 2023 Quantum Sci. Technol. 8 035024.
14. *Li K., McDermott R., Vavilov M.G.* 2019 Phys. Rev. Appl. 12 014044.
15. *Bowdrey M.D., Oi D. K., Short A.J., Banaszek K., Jones J.A.* Fidelity of single qubit maps // Physics Letters A, 294(5-6), pp. 258-260. 2002.
16. *Likharev K.K., Semenov V.K.* RSFQ logic/memory family: A new Josephson-junction technology for sub-terahertz-clock-frequency digital systems // IEEE Transactions on Applied Superconductivity, no. 1(1), pp. 3-28. 1991.
17. *Dalgaard M., Motzoi, F., Sorensen J.J., Sherson, J.* Global optimization of quantum dynamics with AlphaZero deep exploration .. NPJ quantum information, no. 6(1), 6. 2020.
18. *Silver David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, L. Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, Demis Hassabis.* Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. ArXiv abs/1712.01815. 2017.
19. *He K., Zhang X., Ren S., Sun J.* Identity mappings in deep residual networks. In Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV 14 (pp. 630-645). Springer International Publishing.
20. *Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O.* Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347. 2017.
21. *Kingma D. P., Ba J.* Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014.
22. *Ziyadinov V., Tereshonok M.* Low-Pass Image Filtering to Achieve Adversarial Robustness // Sensors. 2023. Vol. 23, No. 22. P. 9032. DOI 10.3390/s23229032.