Chinese Journal of Electronics
Vol.29, No.6, Nov. 2020



# Quantum Contextuality for Training Neural Networks

ZHANG Junwei[1] and LI Zhao[2]

(1. *School of Computer Science and Technology, Tianjin University, Tianjin 300350, China*)
(2. *Search Division, Alibaba Group, Hangzhou 311100, China*)

**Abstract — In the training process of Neural networks (NNs), the selection of hyper-parameters is crucial, which determines the final training effect of the model. Among them, the Learning rate decay (LRD) can improve the learning speed and accuracy; the Weight decay (WD) improves the over-fitting to varying degrees. However, the decay methods still have problems such as hysteresis and stiffness of parameter adjustment, so that the final model will be inferior. Based on the Quantum contextuality (QC) theory, we propose a Quantum contextuality constraint (QCC) to constrain the weights of nodes in NNs to further improve the training effect. In the simplest classification model, we combine this constraint with different methods of LRD and WD to verify that QCC can further improve the training effect on the decay method. The performance of the experiments shows that QCC can significantly improve the convergence and accuracy of the model.**

**Key words — Quantum contextuality, Non-classical probability constraints, Training neural networks.**

## I. Introduction

Neural networks (NNs) have been widely concerned and studied since it was proposed. In recent years, it has rapidly developed and achieved gratifying results. NNs is highly effective and contributes to significantly improve the state-of-the-art in Speech recognition, Translation, Visual object recognition, Drug discovery, Driverless car technology, *etc.*[1−3]. However, the training process of NNs is a difficult optimization problem, and the training effect is closely related to the choice of its hyper-parameters.

NNs typically use the optimization method of Stochastic gradient descent (SGD) to update weights. In this optimization process, the learning rate plays a very important role. It is well known that too small a learning rate will make a training algorithm converge slowly while too large a learning rate will make the training algorithm diverge[4]. Therefore, the researchers propose various methods to dynamically adjust the learning rate to balance the relationship between the two. Among them, the achievements in recent years include: Smith[5] proposes a method to improve the learning effect of the model by cyclical learning rates within a fixed range; Dauphin *et al.*[6] introduce an adaptive learning rate scheme based on the equilibration pre-conditioner; Gulcehre *et al.*[7] propose an adaptive learning rate algorithm, which utilizes curvature information for automatically tuning the learning rate. Moreover, people also use the method of Weight decay (WD) to limit the update of weights and thus improve the generalization ability of the model. Krogh *et al.*[8] explains why WD can improve the generalization ability of feed-forward NNs, and proves that WD has the effect of suppressing unrelated components in weight vectors and suppressing the influence of some static noise on targets in linear networks. For the update of weights, the method of Learning rate decay (LRD) has a certain hysteresis and inaccuracy of the adjustment of the learning rate value, and then the update of the weight is in a passive position. Moreover, the method of WD lacks some flexibility. The above problems leads directly to a decline in the learning speed and the final effect of the model. This prompt us to find a better way to solve this optimization problem.

Quantum contextuality (QC) is one of the most important principles in Quantum theory[9,10], as well as an important quantum resource[11,12]. It points out that the measurement result of a measurement operator

Manuscript Received Aug. 06, 2019; Accepted Oct. 16, 2020. This work is supported by the National Key R&D Program of China (No.2017YEF0111900), the National Natural Science Foundation of China (No.61876129), the National Natural Science Foundation of China (No.U1636203), the Alibaba Innovation Research Foundation 2017 and the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement (No.721321), Part of the work was performed when Junwei Zhang visited the Alibaba Inc. in 2019.
 DOI:10.1049/cje.2020.10.003

is directly related to the context (*i.e.*, the quantum context) in which it is located. Due to the absence of practical quantum computers, simulating or quantifying QC[13,14] into specific constraints on classical computers is a possible option. In order to make the article structure neat, we put the basic concepts and research progress of QC.

Based on the generalization of KCBS inequality[15], we describe the boundaries given by quantum theory and classical probability (*i.e.*, non-contextuality) theory as the constraints of quantum contextual conditions and mutual exclusive conditions on general probability theory respectively, *i.e.* Quantum contextual constraint (QCC) and Mutual exclusive constraint (MEC). Under the guidance of this idea, we introduce QCC into NNs to constrain the weights. Since we want to verify that QCC can still improve the training effect of the model on the existing LRD and WD, we need the control group and the experimental group, and the two groups are compared to verify the contribution of QCC. We set the model to which only the decay method is added as the control group, and set the model to which the decay method and QCC are added as the experimental group. Tested under the simplest classification model for two standard classification datasets, the expression shows that QCC can significantly improve the convergence speed and generalization ability.

## II. Quantifying and Utilizing QC

Kochen-specker (KS) theorem[16] is most proposed by Kochen and Specker in 1975, which reveals the characteristics of contextuality and nonlocality[17] of quantum theory. QC points out that the measurement results of a measurement operator depend explicitly on the measurement base it is in, and this measurement base constitutes a contextual scenario of the measured quantity. QC is considered to be one of the most essential features of quantum theory, and many other quantum properties can be derived from it. At the same time, QC itself is a very important quantum resource. Quantum nonlocality is only a special case of QC, which reveals that quantum theory violates the most basic locality principle of classical physics, that is, things directly affect only what is in their immediate vicinity[18].

Although contextuality and nonlocality can be regarded as non-classical features and can be applied to information processing tasks[19,20], their research and application are not yet reached a matching degree. In the absence of quantum computer resources, the researchers mainly work in two directions[14]:

- Quantify contextuality and non-locality;
- Find suitable carriers to achieve quantum characteristics under classical conditions[21,22], *e.g.*, NNs.

In terms of quantifying QC, Kleinmann *et al.*[23] investigate the memory cost as the critical resource in a classical simulation of QC and the use of memory is taken as a quantitative index of QC; Svozil and Karl[24] believe that the amount of this violation of non-contextuality can be quantified by the frequency of occurrence of QC; Grudka *et al.*[26] propose a program of quantifying contextuality based on two complementary approaches.

Among them, the family of inequalities proposed by Cabello *et al.*[15] is very meaningful, which clearly indicates the boundary of the classical probability (*i.e.*, non-contextual), the quantum and the generalized probability theory for the high-dimensional system or multi-observable. Based on the work, we will get MEC and QCC. The quantification method and how it works with NNs will be given below.

**1. Quantifying QC**

The Klyachko-Can-Binicioğlu-Shumovsky (KCBS) inequality[26] (that is, one satisfied by any non-contextual hidden variable theory) plays an important role in the quantum theory because it can verify the existence of QC, and also shows that quantum theory cannot be explained simply by hidden variable theory. In known inequalities, the KCBS inequality is the simplest non-contextual inequality, which can be violated by a three-level quantum system or qutrit[15,27].

The KCBS inequality is experimentally validated[28] and contribute to the development of a lot of related research. Cabello *et al.*[10] summarize and generalize the research on QC, and obtain a set of inequalities covering the non-contextual, the quantum and the generalized probabilistic theory, which are denoted by $\mathcal{E}_{CN}$, $\mathcal{E}_Q$ and $\mathcal{E}_{GP}$, respectively. The extreme values over these inequalities are denoted $\beta_{CN}$, $\beta_Q$ and $\beta_{GP}$, respectively, which relationship between them can be expressed as

$$\beta_{CN} \le \beta_Q \le \beta_{GP} \tag{1}$$

by definition[15].

For a given hypergraph $\Gamma$, we can define the adjacency graph $G$ on the vertex set $V$: two vertices $i, j \in V$ are joined by an edge if and only if there exists a context $C \in \Gamma$ such that both $i, j \in C$. Then

$$\beta_{CN}(\Gamma) = \alpha(G),\ \beta_Q(\Gamma) = \vartheta(G),\ \beta_{GP}(\Gamma) = \alpha^*(G) \tag{2}$$

where $\alpha(G) = \dfrac{n-1}{2}$ is the independence number, $\vartheta(G) = \dfrac{ncos(\pi/n)}{1+cos(\pi/n)}$ is the Lovász number[29] and $\alpha^*(G) = \frac{n}{2}$ is the so-called fractional packing number[30].

If we assume that the interactions between the vertices of $G$ are the same (that is, the vertices have the same status), then we can get the mean of each vertex (that is, the mean of each vertex in the case of the generalized probabilistic, the quantum and the non-contextuality theory), which are

$$\overline{\alpha}(G) = \frac{\alpha(G)}{n},\ \overline{\vartheta}(G) = \frac{\vartheta(G)}{n},\ \overline{\alpha^*}(G) = \frac{\alpha^*(G)}{n} \tag{3}$$

respectively. Note that although we assume that all vertices in $G$ have the same status, the fact is that the premise of the KCBS inequality is based on this assumption. Therefore, averaging for each vertex is normal and natural.

Here we introduce the constraint quantity $\mathcal{C}$, *i.e.*, MEC $\mathcal{C}_{me}$, QCC $\mathcal{C}_{qc}$, let $\alpha(G) = \mathcal{C}_{me} \cdot \alpha^*(G)$, $\vartheta(G) = \mathcal{C}_{qc} \cdot \alpha^*(G)$. This implies that the constraint of the generalized probabilistic theory is 1. In the same way, we can get the expression of $\mathcal{C}$ under the mean condition, which is

$$\overline{\alpha}(G) = \frac{\mathcal{C}_{me} \cdot \alpha^*(G)}{n} = \frac{1}{2} \cdot (1 - \frac{1}{n}) \tag{4}$$

$$\overline{\vartheta}(G) = \frac{\mathcal{C}_{qc} \cdot \alpha^*(G)}{n} = \frac{1}{2} \cdot (1 - \tan^2(\frac{\pi}{2n})) \tag{5}$$

$$\overline{\alpha^*}(G) = \frac{1 \cdot \alpha^*(G)}{n} = \frac{1}{2} \cdot 1 \tag{6}$$

respectively. So far, we obtain the two most important constraints in this paper, *i.e.*, MEC of each vertex under non-contextual (*i.e.*, classical probabilistic) theory,

$$\mathcal{C}_{me}(n) := 1 - \frac{1}{n} \tag{7}$$

and QCC of each vertex under quantum theory,

$$\mathcal{C}_{qc}(n) := 1 - \tan^2(\frac{\pi}{2n}) \tag{8}$$

Their curve graphs are shown in Fig.1.

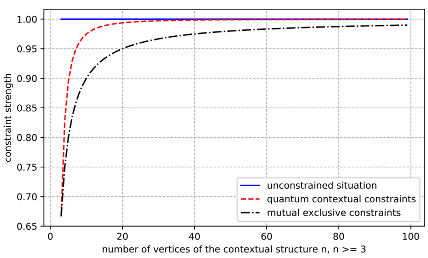



Fig. 1. The yellow line is MEC under the non-contextual theory; The red line is QCC under the quantum theory; The blue line is the unconstrained situation under the generalized probabilistic theory

Since $\mathcal{C}_{qc}$ is the average QCC of each vertex obtained in the case of $n$-cycle graphs, in the other $n$ vertex graphs, $\mathcal{C}_{qc}$ will be larger than this constraint but smaller than MEC. Since the existing research is still unable to determine QC of any graph, we define a degree of freedom $\gamma$ to give a change block for QCC, which is

$$\mathcal{C}_{QCC}(n, \gamma) = \gamma \cdot \mathcal{C}_{qc}(n) + (1 - \gamma) \cdot \mathcal{C}_{me}(n) \tag{9}$$

where $\gamma \in [0, 1]$.

**2. Utilizing QC**

From the derivation method of QCC and MEC, we are inspired that QC can be used as a constraint item and act on the probability item to make it have features of QC. This article is based on such ideas or motivations.

The introduction of QCC can be thought of as a description of the relationship between vertices within multiple probability spaces. The difficulty in introducing this constraint into the classical structure is to find a graphical structure with multiple probability spaces. However, since there is no multiple probability space in the classical world, we can only retreat to the next, and use the independence between the vertices to identify the existence of multiple spaces. Since in the fully connected layer of NNs, the vertices in the same layer are not connected, *i.e.*, independent of each other, and the vertices in the adjacent layer are connected, it constitutes a multi-probability space that meets the basic requirements, as shown in Fig.2. Based on this structure, we introduce QCC into NNs to constrain the weight of the node.

In order to avoid the introduction of interference factors and thus affect the experimental results. We use a single hidden layer classification model structure to experiment. In the model, all output layer nodes and each hidden layer node form a context structure, so the number of hidden layer nodes is equal to the number of context structures.

We apply $\mathcal{C}_{QCC}$ to the neuron nodes of the output layer to achieve the purpose of limiting the weight of the node. Its form is:

$$evidence_i = \sum_j (\mathcal{C}_{QCC})_i W_{i,j} x_j + b_i \tag{10}$$

where $W$ is the weight matrix, $b$ is the bias, $x$ is the value entered, $j$ represents the attributes of the sample and $i$ represents the number of the samples. It should be noted here that adding constraint $\mathcal{C}_{QCC}$ is the main difference between our model and the classic model, and the other parts are basically the same. In order to meet the needs of the operation, $\mathcal{C}_{QCC}$ here is a diagonal matrix. The *evidence* is then converted to probability

$$y = softmax(evidence) \tag{11}$$

using the softmax function. The softmax function can be defined as

$$softmax(x) = normalize(exp(x)) \tag{12}$$

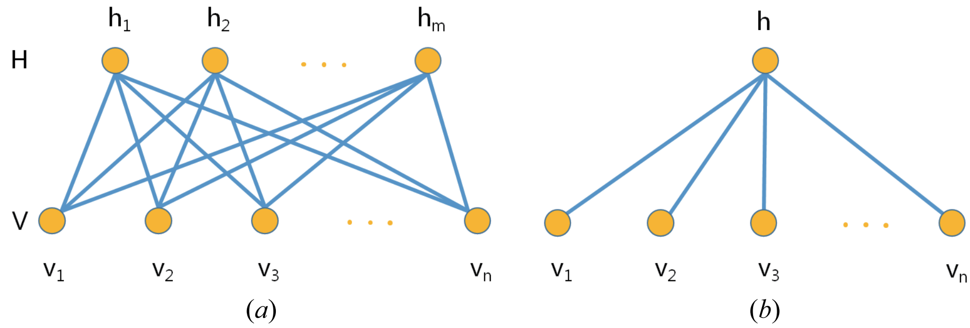



Fig. 2. $(a)$ is a fully connected layer common in NNs. $(b)$ is a basic component of $(a)$, that is, one of the upper layers connected to all nodes of the lower layer together constitutes a basic component. This basic component is considered to be a basic component that can produce QC with $n+1$ vertices (nodes of neurons)

where *normalize* represents a normalization function. Note that $\mathcal{C}_{QCC}$ here is fundamentally different from the L2 regularity. The L2 regularity acts on the weight update to reduce the weight, *i.e.*, on the cost function. And $\mathcal{C}_{QCC}$ directly acts on the weight to constrain the weight, *i.e.*, on the objective function, and it needs to learn the specific parameters $\gamma$.

## III. Experiments

The purpose of this section is to demonstrate the effectiveness of $\mathcal{C}_{QCC}$ in the simplest classification model. We add $\mathcal{C}_{QCC}$ to the model that already uses the method of LRD to verify whether $\mathcal{C}_{QCC}$ will further improve the convergence of the loss function. Similarly, we add $\mathcal{C}_{QCC}$ to the model that already uses WD, *i.e.*, L2 regularity, to verify that $\mathcal{C}_{QCC}$ can further improve the generalization ability of the model.

### 1. Model structure and datasets

We use a single hidden layer classifier as an experimental model to verify the effectiveness of $\mathcal{C}_{QCC}$. The experimental model of the control group is a classical single-layer classification model without adding $\mathcal{C}_{QCC}$; the experimental model of the experimental group is a single-layer classification model with $\mathcal{C}_{QCC}$ added. The model structure is shown in Fig.3. The number of hidden layer nodes is the same as the number of input layer nodes, and the number of output layer nodes is the number of categories of samples. The final output nodes of the model use the softmax function to adjust its output value to a relative probability.

Our intention to use the simplest classification model is to minimize the impact of other variables on the experimental results. Moreover, it is known from the nature of QC that the more complex the model structure, the smaller the average contextual strength of the substructure. This is why we chose a simple model.

We train the model with two multi-category datasets, MNIST and Fashion-MNIST. The details of the datasets can be viewed separately in Refs.[31,32].

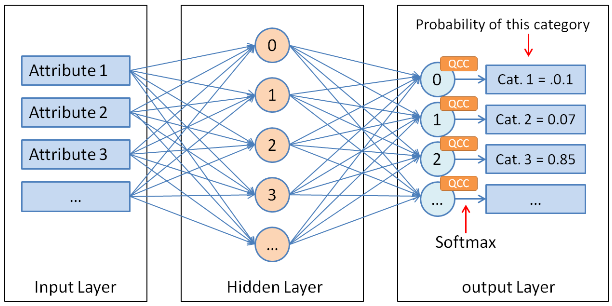



Fig. 3. Model structure of the experimental object

### 2. Learning rate decay with QCC

In order to verify that the addition of $\mathcal{C}_{QCC}$ will improve the convergence of the loss function and further improve the training effect of the model, we combine the current mainstream method of LRD, such as Exponential decay, Staircase decay, Polynomial decay, Polynomial decay (cycle), Inverse Time decay and Natural Exponential decay, to verify the performance of $\mathcal{C}_{QCC}$. We train the model under the MNIST and Fashion-MNIST dataset, respectively. The experimental results of the model are shown in Fig.4 and Fig.5, respectively.

In order to objectively verify the contribution of $\mathcal{C}_{QCC}$, we first tune the model that add different methods of LRD, and train to obtain experimental data of the control group. Without modifying any parameters, we only add $\mathcal{C}_{QCC}$ and train the model to get the data of the experimental group. The advantage of setting the experimental procedure in this way is that the contribution of $\mathcal{C}_{QCC}$ can be clearly determined by comparing the data of the experimental group with the data of the control group, and the same parameters can be used to reduce the influence of the parameters on the experiment. The parameters of LRD can be selected according to the loss function curve, and the appropriate parameters can be freely selected. Because we will use the

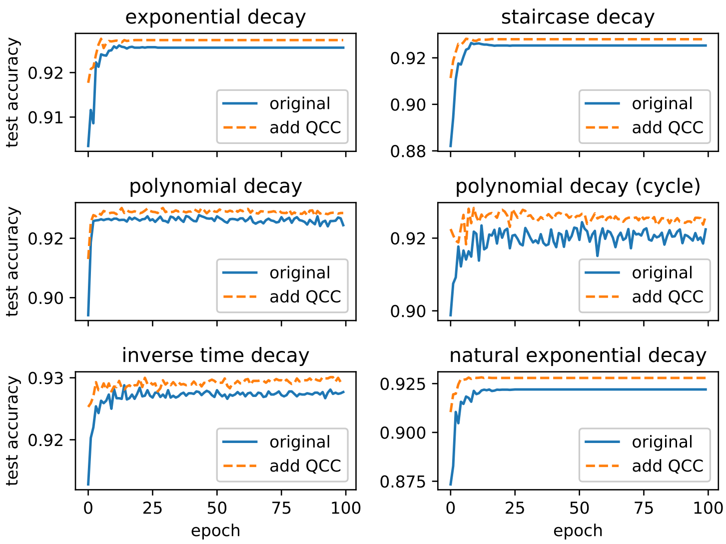



Fig. 4. Under the MNIST dataset, $\mathcal{C}_{QCC}$ is added into the model that already uses Exponential decay, Staircase decay, Polynomial decay, Polynomial decay (cycle), Inverse Time decay and Natural Exponential decay, respectively, to verify the validity of $\mathcal{C}_{QCC}$. The test accuracy of the model is shown in the figure, where "original" means that only the decay method is added

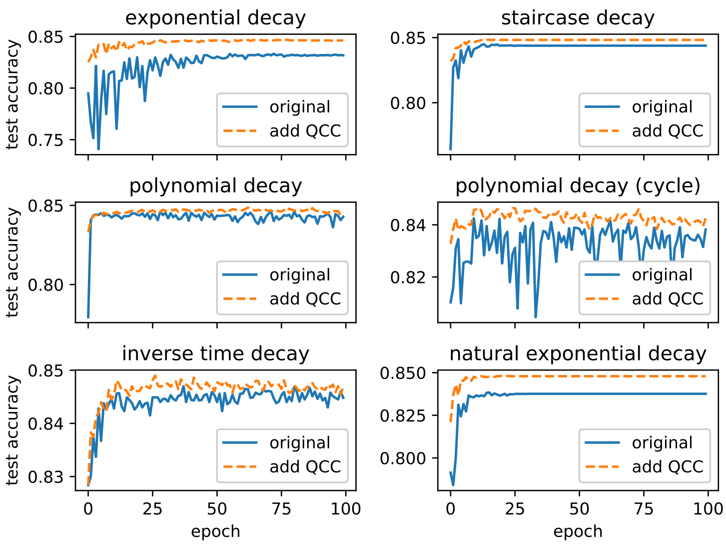



Fig. 5. Under the Fashion-MNIST dataset, $\mathcal{C}_{QCC}$ is added to the model that already uses Exponential decay, Staircase decay, Polynomial decay, Polynomial decay (cycle), Inverse Time decay and Natural Exponential decay, respectively, to verify the validity of $\mathcal{C}_{QCC}$. The test accuracy of the model is shown in the figure, where "original" means that only the decay method is added

same parameters for both the control and the experimental group, the choice of parameters will not affect the comparison results. This experiment uses the parameters that show the best results in the control group. And the other two super-parameters, mini-batch sizes and epochs, are 128 and 100 respectively.

From the comparison of experimental data, we can clearly see that $\mathcal{C}_{QCC}$ plays a certain role in promoting the convergence of the model. Moreover, $\mathcal{C}_{QCC}$ also made a certain contribution to improve the test accuracy of the model. From these results, we can basically determine that $\mathcal{C}_{QCC}$ is effective for improving the training effect of the model. We also selected other kinds of optimization algorithms for experiments, such as Adam, Adgrad and AdaDelta, and obtained similar results.

For the degree (or size) of contribution, the contribution of $\mathcal{C}_{QCC}$ to the final effect of the model is about one percentage point, not a big step forward. There are two main reasons for this: 1. This model reaches the limit that the model can learn, and the space that can be improved is not too large; 2. As mentioned above, as the complexity of the model increases, the constraint strength of quantum contextuality decreases. This is why we do not use complex models for experiments.

**3. Weight decay with QCC**

In order to verify that the addition of $\mathcal{C}_{QCC}$ can still improve the generalization ability for the model that add methods of WD (*e.g.*, L2 regularity). We set the model with the L2 regularity as the control group, and set the model with the L2 regularity and $\mathcal{C}_{QCC}$ as the experimental group. The parameter setting method of L2 regularity is the same. Moreover, the learning rate of both is set to a fixed value and the same. In addition, the learning rate, mini-batch size and epoch of the two groups are set to 0.01, 128 and 100, respectively. Two models were trained to obtain the data of the control group and the experimental group. The results are shown in Fig.6.

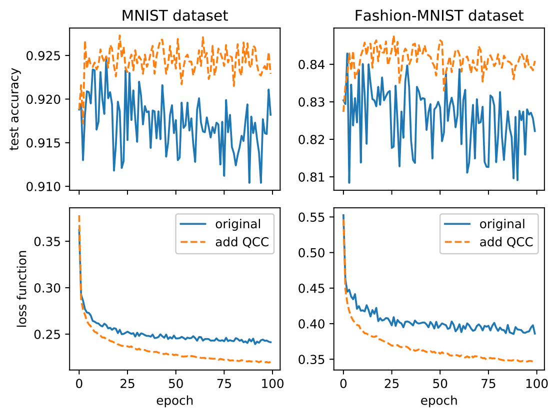



Fig. 6. Under the MNIST and Fashion-MNIST dataset, verify the contribution of $\mathcal{C}_{QCC}$ to the model that has added the L2 regularity. The experimental results are shown in this figure

It can be seen from the experimental results that $\mathcal{C}_{QCC}$ has a constraining effect, that is, the loss function vibration is significantly reduced, and also causes the loss function to converge to a smaller value. From the perspective of test accuracy, $\mathcal{C}_{QCC}$ can significantly improve the generalization ability of the model. It proves that $\mathcal{C}_{QCC}$ can still play a certain role under the premise of L2 regularity.

## IV. Conclusions

This paper considers the possibility of further improving the model effect from the perspective of optimization of NNs. Starting from an important principle of quantum theory, *i.e.*, QC, we extract QCC based on the work of Cabello et al.[15] and apply it to the optimization process of NNs. On top of the existing optimization methods, we add QCC to verify whether QCC further improve the convergence effect and generalization ability of the model. The experimental results show that the QCC improves the test accuracy and significantly promotes the stability of the accuracy curve. From the experimental performance, we can basically verify that QCC is effective.

In the absence of a quantum computer that can be used, the use of quantum properties can only be compensated by simulating the basic structure required to produce quantum properties or by quantifying quantum properties[14]. In this paper, we use the method of quantifying quantum contextuality to introduce QC as a constraint into the training process of NNs. Although this method cannot clearly indicate the structural features that produce quantum properties, and can not clearly distinguish the specific type of quantum properties, that is, there is a large ambiguity, it provides the possibility of using quantum properties from the perspective of quantifying quantum properties.

In the later work, we will try to introduce the quantized value of QC into the analog structure that can produce QC, which is combined and complement each other. Moreover, for the problem that QCC can only be applied to the minimum contextual structure and that the constraint boundary is broad, we will further study it in the next step.

## References

[1] I. Goodfellow, Y. Bengio, A. Courville, *et al.*, “Deep learning”, *Cambridge: MIT press*, 2016.

[2] Y. Bengio, “Learning deep architectures for AI”, *Now Publishers Inc*, 2009.

[3] W.U. Yujia, L.I. Jing, S. Chengfang, *et al.*, “Words in Pairs Neural Networks for Text Classification”, *Chinese Journal of Electronics*, Vol.29, No.3, pp.491–500, 2020.

[4] Y. Bengio, “Practical recommendations for gradient-based training of deep architectures”, *Neural networks: Tricks of the trade*, Berlin, Heidelberg, pp.437–478, 2012.

[5] L.N. Smith, “Cyclical learning rates for training neural networks”, *IEEE Winter Conference on Applications of Computer Vision*, pp.464–472, 2017.

[6] Y. Dauphin, H.D. Vries and Y. Bengio, “Equilibrated adaptive learning rates for non-convex optimization”, *Advances in neural information processing systems*, pp.1504–1512, 2015.

[7] C. Gulcehre, M. Moczulski and Y. Bengio, “Adasecant: robust adaptive secant method for stochastic gradient”, *arXiv preprint* arXiv:1412.7419, 2014.

[8] A. Krogh and J.A. Hertz, “A simple weight decay can improve generalization”, *Advances in neural information processing systems*. pp.950–957, 1992.

[9] P. Kurzyński, R. Ramanathan and D. Kaszlikowski, “Entropic test of quantum contextuality”, *Physical review letters*, Vol.109, No.2, pp.020–404, 2012.

[10] A. Cabello, S. Severini and A. Winter, “Graph-theoretic approach to quantum correlations”, *Physical review letters*, Vol.112, No.4, pp.040–401, 2014.

[11] J.B. Vega, N. Delfosse, D.E. Browne, *et al.*, “Contextuality as a resource for models of quantum computation with qubits”, *Physical review letters*, Vol.119, No.12, pp.120–505, 2017.

[12] A.W. Simmons, “How (Maximally) Contextual Is Quantum

Mechanics?”, *Quantum, Probability, Logic. Springer*, Cham, pp.505–519, 2020.

[13] A. Grudka, K. Horodecki, M. Horodecki, *et al.*, “Quantifying contextuality”, *Physical review letters*, Vol.112, No.12, pp.120–401, 2014.

[14] S. Mansfield, “The Mathematical Structure of Non-locality and Contextuality”, 2013.

[15] A. Cabello, S. Severini and A. Winter, “(Non-) contextuality of physical theories as an axiom”, *arXiv preprint* arXiv:1010.2163, 2010.

[16] S. Kochen, and E.P. Specker, “The problem of hidden variables in quantum mechanics”, *The logico-algebraic approach to quantum mechanics. Springer*, Dordrecht, pp.293–328, 1975.

[17] S. Popescu and D. Rohrlich, “Quantum nonlocality as an axiom”, *Foundations of Physics*, Vol.24, No.3, pp.379–385, 1994.

[18] H.W. Lin, M. Tegmark and D. Rolnick, “Why does deep and cheap learning work so well?”, *Journal of Statistical Physics*, Vol.168, No.6, pp.1223–1247, 2017.

[19] K. Horodecki, M. Horodecki, P. Horodecki, *et al.*, “Contextuality offers device-independent security”, *arXiv preprint* arXiv:1006.0468, 2010.

[20] L.I. Panchi and Z. Ya, “Model and algorithm of sequence-based quantum-inspired neural networks”, *Chinese Journal of Electronics*, Vol.27, No.1, pp.9-18, 2018.

[21] B.R. La Cour, “Quantum contextuality in the Mermin-Peres square: A hidden-variable perspective”, *Physical Review A*, Vol.79, No.1, pp.012–102, 2009.

[22] A.Y. Khrennikov, “Contextual approach to quantum formalism”, *Springer Science Business Media*, 2009.

[23] M. Kleinmann, O. Gühne, J.R. Portillo, *et al.*, “Memory cost of quantum contextuality”, *New Journal of Physics*, Vol.13, No.11, pp.113–011, 2011.

[24] K. Svozil, “How much contextuality?”, *Natural Computing*, Vol.11, No.2, pp.261–265, 2012.

[25] A. Grudka, K. Horodecki, M. Horodecki, *et al.*, “Quantifying contextuality”, *Physical review letters*, Vol.1121, No.2, pp.120–401, 2014.

[26] A.A. Klyachko, M.A. Can, S. Binicioğlu, *et al.*, “Simple test for hidden variables in spin-1 systems”, *Physical review letters*, Vol.101, No.2, pp.020–403, 2008.

[27] S. Kochen and E.P. Specker, “The problem of hidden variables in quantum mechanics”, *The logico-algebraic approach to quantum mechanics. Springer*, Dordrecht, pp.293–328, 1975.

[28] S. Yu and C.H. Oh, “State-independent proof of Kochen-Specker theorem with 13 rays”, *Physical review letters*, Vol.108, No.3, pp.030–402, 2012.

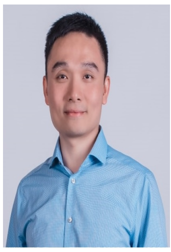
[29] L. Lovász, “On the Shannon capacity of a graph”, *IEEE Transactions on Information theory*, Vol.25, No.1, pp.1–7, 1979.

[30] M. Grötschel, L. Lovász and A. Schrijver, “The ellipsoid method and its consequences in combinatorial optimization”, *Combinatorica*, Vol.1, No.2, pp.169–197, 1981.

[31] Y. LeCun, C. Cortes and C.J. Burges, “MNIST handwritten digit database”, *http://yann. lecun. com/exdb/mnist*, 2010-7-23.

[32] H. Xiao, K. Rasul and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms”, *arXiv preprint* arXiv:1708.07747, 2017.

**ZHANG Junwei** was born in 1989. He is a Ph.D. candidate of Tianjin University. His research interests include Quantum Cognitive, Quantum Artificial Intelligence and Quantum Machine Learning.
(Email: junwei@tju.edu.cn)



**LI Zhao** received the Ph.D. degree from the Computer Science Department, University of Vermont. His research interests include adversarial machine learning, network representation learning, knowledge graphs, multi-agent reinforcement learning, and big data-driven security.
(Email: lizhao.lz@alibaba-inc.com)