Erratum: Learnability of Quantum Neural Networks [PRX QUANTUM 2, 040337 (2021)]

Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Shan You, and Dacheng Tao

(Received 29 March 2022; published 11 July 2022)

DOI: 10.1103/PRXQuantum.3.030901

The original version of this paper contains a minor error which changes some statements of the conclusions. The essence of this error was brought to our attention by Mr Bobak Toussi Kiani. Here we fix this minor error. It is noteworthy that these corrections do not affect the main results and conclusions of the whole paper.

Our paper should be amended as follows: replace the range of the regularization coefficient $\lambda \in [0, 1/3\pi] \cup (1/\pi, \infty)$ in Theorem 1 and elsewhere with $\lambda \in (1/\pi, \infty)$.

Formal issue. The revised range of λ arises from a mistake in the derivation of Lemma 4 (PL condition of the loss function). In particular, to satisfy the inequality in Eqn. (D14), λ should range from $1/\pi$ to ∞. For completeness, we append the correct proof of Lemma 4 below.

Proof of Lemma 4. Recall the definition of Polyak-Lojasiewicz as formulated in Definition 2, it requires that there exists a constant $\mu > 0$ such that the objective function $\mathcal{L}$ satisfies

$$\|\nabla\mathcal{L}(\boldsymbol{\theta})\|^2 \geq 2\mu(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}^*) , \tag{D.12}$$

where $\mathcal{L}^* = \min_{\boldsymbol{\theta}\in\mathcal{C}} \mathcal{L}(\boldsymbol{\theta})$.

We first derive a lower bound of $\|\nabla\mathcal{L}(\boldsymbol{\theta})\|^2$. In particular, we have

$$\|\nabla\mathcal{L}(\boldsymbol{\theta})\|^2 = \sum_{j=1}^{d}(\nabla_j\mathcal{L}(\boldsymbol{\theta}_j))^2 \geq \max_j(\nabla_j\mathcal{L}(\boldsymbol{\theta}))^2 . \tag{D.13}$$

The lower bound of $\max_j(\nabla_j\mathcal{L}(\boldsymbol{\theta}))^2$ as shown in Eqn. (D.13) follows

$$\max_j(\nabla_j\mathcal{L}(\boldsymbol{\theta}))^2 \geq \min_{\boldsymbol{\theta}_j\in[\pi,3\pi]}(-1+\lambda\boldsymbol{\theta}_j)^2 , \tag{D.14}$$

where the inequality is achieved by exploiting the last second line of Eqn. (D.3), the fact that the output of QNNs ranges from 0 to 1 (i.e., $\hat{y}_i, y_i, \hat{y}_i^{(\pm j)} \in [0,1]$), $\boldsymbol{\theta}_j \in [\pi, 3\pi]$ for all j, and $\lambda > 0$, i.e.,

$$\nabla_j\mathcal{L}(\boldsymbol{\theta}) = \frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\frac{\hat{y}_i^{(+j)} - \hat{y}_i^{(-j)}}{2} + \lambda\boldsymbol{\theta}_j \geq -1 + \lambda\boldsymbol{\theta}_j .$$

Combining the assumption $\lambda \in (\frac{1}{\pi}, \infty)$ and the above results, the lower bound of Eqn. (D.13) satisfies

$$\|\nabla\mathcal{L}(\boldsymbol{\theta})\|^2 \geq (-1+\lambda\boldsymbol{\theta}_j)^2 > 0 .$$

We then derive the upper bound of the term $(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}^*)$ in Eqn. (D.12). In particular, we have

$$\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}^* < \mathcal{L}(\boldsymbol{\theta}) + 0 \leq 1 + \lambda d(3\pi)^2 , \tag{D.15}$$

where the first inequality comes from the definition of $\mathcal{L}^*$, i.e.,

$$-\mathcal{L}^* = -\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i^* - y_i)^2 - \frac{\lambda}{2}\|\boldsymbol{\theta}^*\|^2 < 0\,,$$

with $\hat{y}_i^* = \mathrm{Tr}(\Pi U(\boldsymbol{\theta}^*)\rho_i U(\boldsymbol{\theta}^*)^\dagger)$, and the second inequality employs the definition of $\mathcal{L}(\boldsymbol{\theta})$ with

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 < 1 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2\,,$$

and $\frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \le \frac{\lambda}{2}d\|\boldsymbol{\theta}\|_\infty^2 < (3\pi)^2\lambda d$.

By combining Eqns. (D.17) and (D.15) with Eqn. (D.12), we obtain the following relation

$$\|\nabla\mathcal{L}(\boldsymbol{\theta})\|^2 \ge (-1+\lambda\pi)^2 \ge 2\mu(1+\lambda d(3\pi)^2) > 2\mu(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}^*)\,. \tag{D.16}$$

The above relation indicates that the objection function $\mathcal{L}(\boldsymbol{\theta})$ satisfies PL condition with

$$\mu = \frac{(-1+\lambda\pi)^2}{2(1+\lambda d(3\pi)^2)}\,.$$

■

We remark that Lemma 4 considers a general result for QNNs. Consequently, to achieve PL condition, the regularization term is much larger than the data-dependent term in the loss function, which leads to optimization parameters that largely depend on the regularization term $\frac{\lambda}{2}\|\boldsymbol{\theta}\|^2$. In the following, we demonstrate that when taking into account to a specific class of QNNs, the PL condition can be effectively satisfied in which the data-dependent term dominates the loss and the utility bound R_2 becomes much more meaningful.

A concrete example is over-parameterized QNNs [1–3] with the lazy training behavior [4–6]. That is, in training over-parameterized QNNs via gradient descent, the parameters may not change significantly and remain close to their initial values. The lazy training phenomenon suggests that the gradient of the data-dependent term is bounded by a small constant, i.e., $|\nabla_j \ell_i(\boldsymbol{\theta})| \equiv |\frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\frac{\hat{y}_i^{(+j)} - \hat{y}_i^{(-j)}}{2}| \le \epsilon$. Using this relation, the lower bound of $\max_j(\nabla_j\mathcal{L}(\boldsymbol{\theta}))^2$ as shown in Eqn. (D.13) can be tightened, i.e.,

$$\max_j(\nabla_j\mathcal{L}(\boldsymbol{\theta}))^2 \ge \min_{\boldsymbol{\theta}_j\in[\pi,3\pi]}(-\epsilon + \lambda\boldsymbol{\theta}_j)^2\,. \tag{D.17}$$

To enable the left hand-side is greater than 0, the range of λ becomes

$$\lambda \in (\epsilon/\pi, \infty], \tag{D.18}$$

and the corresponding parameter to achieve PL condition is $\mu = \frac{(-\epsilon+\lambda\pi)^2}{2(1+\lambda d(3\pi)^2)}$. Notably, when the lazy behavior is evident and ϵ is sufficiently small, e.g., $\epsilon = 10^{-5}$, λ can be small and the data-dependent term dominates the loss function and the achieved utility bound R_2 in Theorem 1 becomes meaningful.

Let us further comment on a debated issue in both classical and quantum optimization theory, i.e., *whether the employment of PL condition makes the achieved results with respect to the global convergence (utility bound R_2) trivial*. As stated in [7], when the target function is non-convex, a general analysis of its convergence to the global optima is extremely difficult or even intractable. To make the problem become tractable, certain assumptions are necessary to (quantum) neural networks. For instance, to analyze the global landscape, several studies have focused on deep *linear* neural networks [8,9] or *two-layer* neural networks [10,11]. Although these networks have little representation power and are not very interesting from a learning perspective, it is a valid problem from an optimization perspective. In the same manner, to analyze the global landscape of quantum neural networks (QNNs), as measured by R_2, we assume that the loss function satisfies PL condition. Lemma 4 exhibits a sufficient condition of λ when the explored loss function meets the PL condition. The employment of PL condition is because it is a general assumption broadly used in the optimization theory of both deep neural networks [12] and QNNs [13]. In particular, Ref. [13] also leverages PL condition to analyze the utility bound R_2 as we did, except that the optimization is completed by stochastic gradient descent method in the ideal setting. In this respect, the achieved results related to R_2 are crucial in understanding the global behavior of QNNs.

Other minor issues. There are also some omitted details and minor mistakes in the original version.

(1) For clarity, in Eq. (D1), it should be specified that the output of QNNs ranges from 0 to 1.
(2) There is a missing factor $1/2$ in the loss function of Eqn. (1).
(3) Ref. [32] should be "M. C. Caro and I. Datta, Pseudo-dimension of quantum circuits, Quantum Mach. Intell. 2, 14 (2020)" and the corresponding link is https://doi.org/10.1007/s42484-020-00027-5 [14].
(4) Ref. [46] should be "M. C. Caro, Quantum learning Boolean linear functions w.r.t. product distributions, Quantum Inf Process 19, 172 (2020)" and the corresponding link is https://doi.org/10.1007/s11128-020-02661-1 [15].
(5) Ref. [38] should be "Gyurik, C., van Vreumingen, D., & Dunjko, V. (2021). Structural risk minimization for quantum linear classifiers. arXiv preprint arXiv:2105.05566" [16].

[1] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and M Cerezo, Theory of overparametrization in quantum neural networks. *arXiv preprint arXiv:2109.11676*, 2021.

[2] Junyu Liu, Francesco Tacchino, Jennifer R Glick, Liang Jiang and Antonio Mezzacapo, Representation learning via quantum neural tangent kernels. *arXiv preprint arXiv:2111.04225*, 2021.

[3] Norihito Shirai, Kenji Kubo, Kosuke Mitarai and Keisuke Fujii, Quantum tangent kernel. *arXiv preprint arXiv:2111.02951*, 2021.

[4] Erfan Abedi, Salman Beigi and Leila Taghavi, Quantum lazy training. *arXiv preprint arXiv:2202.08232*, 2022.

[5] Junyu Liu, Khadijeh Najafi, Kunal Sharma, Francesco Tacchino, Liang Jiang and Antonio Mezzacapo, An analytic theory for the dynamics of wide quantum neural networks. *arXiv preprint arXiv:2203.16711*, 2022.

[6] Matthias C Caro, Hsin-Yuan Huang, M Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J Coles, Generalization in quantum machine learning from few training data. *arXiv preprint arXiv:2111.05292*, 2021.

[7] Ruo-Yu Sun, Optimization for deep learning: An overview. Journal of the Operations Research Society of China, **8**(2):249–294, 2020.

[8] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous and Yann LeCun, The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.

[9] Yeonjong Shin, Effects of depth, width, and initialization: A convergence analysis of layer-wise training for deep linear neural networks. Analysis and Applications, **20**(01):73–119, 2022.

[10] C Daniel Freeman, Joan Bruna, Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

[11] Mahdi Soltanolkotabi, Adel Javanmard and Jason D. Lee, Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. IEEE Transactions on Information Theory, **65**(2):742–769, 2018.

[12] Marina Danilova, Pavel Dvurechensky, Alexander Gasnikov, Eduard Gorbunov, Sergey Guminov, Dmitry Kamzolov and Innokentiy Shibaev, Recent theoretical advances in non-convex optimization. *arXiv preprint arXiv:2012.06188*, 2020.

[13] Ryan Sweke, Frederik Wilde, Johannes Meyer, Maria Schuld, Paul K Fährmann, Barthélémy Meynard-Piganeau and Jens Eisert, Stochastic gradient descent for hybrid quantum-classical optimization. Quantum, **4**, 314, 2020.

[14] Matthias C Caro and Ishaun Datta, Pseudo-dimension of quantum circuits. Quantum Machine Intelligence, **2**(2):1–14, 2020.

[15] Matthias C Caro, Quantum learning boolean linear functions wrt product distributions. Quantum Information Processing, **19**(6):1–41, 2020.

[16] Casper Gyurik, Dyon van Vreumingen and Vedran Dunjko, Structural risk minimization for quantum linear classifiers. *arXiv preprint arXiv:2105.05566*, 2021.