Expressibility of the alternating layered ansatz for quantum computation

Kouhei Nakaji and Naoki Yamamoto

Department of Applied Physics and Physico-Informatics & Quantum Computing Center, Keio University, Hiyoshi 3-14-1, Kohoku, Yokohama, 223-8522, Japan

May 27, 2020

Abstract

The hybrid quantum-classical algorithm is actively examined as a technique applicable even to intermediate-scale quantum computers. To execute this algorithm, the hardware efficient ansatz is often used, thanks to its implementability and expressibility; however, this ansatz has a critical issue in its trainability in the sense that it generically suffers from the so-called gradient vanishing problem. This issue can be resolved by limiting the circuit to the class of shallow alternating layered ansatz. However, even though the high trainability of this ansatz is proved, it is still unclear whether it has rich expressibility in state generation. In this paper, with a proper definition of the expressibility found in the literature, we show that the shallow alternating layered ansatz has almost the same level of expressibility as that of hardware efficient ansatz. Hence the expressibility and the trainability can coexist, giving a new designing method for quantum circuits in the intermediate-scale quantum computing era.

1 Introduction

Recent rapid progress in the hardwares for quantum computing stimulates researchers to develop new techniques that utilize even noisy intermediate-scale quantum (NISQ) device for real applications such as machine learning and quantum chemistry. Especially, several hybrid quantum-classical algorithms have been actively examined, as a means to reduce the computational cost required on quantum computing part. The variational quantum eigensolver (VQE) [2], which trains the parameterized quantum circuit via a classical computation to decrease a given cost function, is typically considered as such a hybrid algorithm.

The critical point in this strategy is in the difficulty to design a suitable and implementable circuit ansatz that, after the training process, may produce an exact or well-approximating solution of a given problem. The hardware efficient ansatz (HEA) [3] is often used mainly because of the implementability on a hardware; this is a relatively shallow circuit ansatz, whose parameters are embedded in the angles of single-qubit rotation gates. However, it was proven in [4] that, when those parameters are randomly chosen, the gradient of a standard cost function vanishes, meaning that the update of the cost function often gets stuck in the learning process before reaching the minimum.

For solving this vanishing gradient problem, several approaches have been proposed. Reference [5] showed numerically that, with a special type of initializing method of the HEA, the vanishing gradient problem does not occur in an ansatz with $O(1)$ qubits. Also, a quantum analogue of the natural gradient have been proposed in [6, 7], the original classical version of which is often used to avoid similar vanishing gradient problems in neural networks. In this paper, we focus on the third approach given by Ref. [8] that provides a method for devising a specific structure of the HEA ansatz, called the Alternating Layered Ansatz (ALT), which in fact provably does not suffer from the vanishing gradient problem. By definition, the class of ALT is included in that of HEA; the difference is that, while a HEA consists of multiple layers of single-qubit rotation gates and entanglers that in principle combines all qubits in each layer, the entangling gates contained in an ALT is restricted to entangle only local qubits in each layer. With this setting, the authors in [8] derived a strict lower bound of the variance of the gradient for an ALT with its parameters randomly chosen (more precisely, the ensemble of unitary matrices corresponding to each circuit block is $2$-design) under the condition that the cost function is local (that is, the cost function is composed of local functions of only a small number of local qubits). By using this lower bound, it was also shown that the vanishing gradient problem can be resolved if the number of layers is of the order $O(poly(\log n))$ where $n$ is the total number of qubits, or roughly speaking if the circuit is shallow.
Then an important question arises; does ALT have a sufficient expressive power (expressibility) for generating a rich class of states, which contains the optimal or a well-approximating state? Because the set of ALTs is a subclass of that of HEA, one might argue that the expressibility of ALT could be much lower than that of the HEA; if this is the case, the ALT circuit may not generate a desired state even though the learning process is smoothly running. Thus, it is worth examining the expressibility of ALT in order to assess the practicality of this ansatz in executing the hybrid quantum-classical algorithm. In this paper, we study this problem using the expressibility measure introduced in [9] and show that, fortunately, the class of shallow ALTs has the same level of expressibility as that of HEAs. Therefore, the expressibility and the trainability of a quantum circuit can coexist, which means that the existing HEA found in the literature can be basically replaced with a simpler ALT without degradation of expressibility while acquiring a better trainability. That is, the ALT might be taken as a new standard ansatz in NISQ computing era.

The structure of this paper is as follows. Section 2 is the preliminary, giving the definition of expressibility and the ansatzes. In Sec. 3, we show both theoretically and numerically that the expressibility of ALT is as high as that of HEA. In Sec. 4, we show how much the value of expressibility is reflected to the result of VQE. Finally, we conclude with some remarks in Sec. 5.

2 Preliminaries

In this section, we define indicators of the expressibility and introduce some circuit ansatzes.

2.1 Indicators of the expressibility

Following Ref. [9], we define the expressibility of a given circuit, by the randomness of states generated from the ansatzes. In Sec. 3, we show both theoretically and numerically that the expressibility of ALT is as high as that of HEA. In Sec. 4, we show how much the value of expressibility is reflected to the result of VQE. Finally, we conclude with some remarks in Sec. 5.

Frame Potential

To define the expressibility of a given circuit ansatz \( C \), let us consider the deviation of the state distribution generated by \( C \) from the Haar distribution, as follows;

\[
\mathcal{A}^{(t)}(C) = \left\| \int_{\text{Haar}} (|\psi\rangle\langle\psi|)^{\otimes t} \, d\psi - \int_{\Theta} (|\psi_\Theta\rangle\langle\psi_\Theta|)^{\otimes t} \, d\Theta \right\|_{HS},
\]

where \( \int_{\text{Haar}} \) denotes the integration over the state distribution with respect to the Haar measure, and \( \| \cdot \|_{HS} \) is the Hilbert Schmidt distance. Also, \( |\psi_\Theta\rangle \) is the state generated by the ansatz \( C \) characterized by the parameter \( \theta \in \Theta \), e.g., \( |\psi_\Theta\rangle = U_C(\theta)|0\rangle \), \( \theta \in \Theta \), where \( U_C(\theta) \) is the unitary operator corresponding to \( C \) and \( |0\rangle \) is an initial state. Then we call that the ansatz \( C \) with smaller \( \mathcal{A}^{(t)}(C) \) has a higher expressibility. This definition is justified by the following reason. That is, because the state \( |\psi\rangle \) generated from the Haar distribution can in principle represent an arbitrary state, the condition \( \mathcal{A}^{(t)}(C) \approx 0 \) implies that the ansatz \( C \) can generate almost all states possibly including the optimal solution (e.g., the ground state in VQE); also, in this case the states generated from \( C \) are almost equally distributed, which is particularly favorable if little is known about the problem.

To compute \( \mathcal{A}^{(t)}(C) \), we instead focus on the following \( t \)-th generalized frame potential \( \mathcal{F}^{(t)}(C) \) of \( C \):

\[
\mathcal{F}^{(t)}(C) = \int_{\Phi} \int_{\Theta} |\langle \psi_\Theta | \psi_\Phi \rangle|^2 \, d\phi \, d\theta,
\]

where both \( \Theta \) and \( \Phi \) represent the same set of parameters of \( C \). In the present paper, we simply call it the \( t \)-th frame potential. In particular, the \( t \)-th frame potential of \( N \)-dimensional states distributed with respect to the Haar measure, is given by \( \mathcal{F}^{(t)}_{\text{Haar}}(N) = \int_{\text{Haar}} \int_{\text{Haar}} |\langle \psi'| \psi \rangle|^2 \, d\psi \, d\psi' \). The point of introducing the frame potential is because these quantities are linked to \( \mathcal{A}^{(t)}(C) \) in the following form; that is, for an arbitrary positive integer \( t \), it holds

\[
\mathcal{F}^{(t)}(C) - \mathcal{F}^{(t)}_{\text{Haar}}(N) = \mathcal{A}^{(t)}(C) \geq 0.
\]

The equality in the last inequality holds if and only if the ensemble of \( |\psi_\Theta\rangle \) is a state \( t \)-design [11, 12, 13]. Thus, the ansatz \( C \) with smaller \( \mathcal{F}^{(t)}(C) \) has a higher expressibility. Also \( \mathcal{F}^{(t)}(C) \) is lower bounded by \( \mathcal{F}^{(t)}_{\text{Haar}}(N) \), meaning that the frame potential can be used as an indicator for quantifying the non-uniformity in the state distribution. In Sec. 5, we calculate \( \mathcal{F}^{(1)}(C) \) and \( \mathcal{F}^{(2)}(C) \) for several ansatz \( C \).
KL-Divergence

Note that the frame potential $\mathcal{E}(C) = D_{KL}(P(C,F)||P_{\text{Haar}}(F))$, where the circuit parameters $\theta \in \Theta$ and $\phi \in \Phi$ are randomly sampled from the circuit ansatz $C$. Hence the probability distribution of $F$, denoted by $P(C,F)$, contains more information for quantifying the randomness of $C$ than $\mathcal{F}(t)(C)$, and thereby the following measure can be used to quantify the expressibility of $C$:

$$\mathcal{E}(C) = D_{KL}(P(C,F)||P_{\text{Haar}}(F)) = \int_0^1 P(C,F) \log \frac{P(C,F)}{P_{\text{Haar}}(F)} dF,$$

where $D_{KL}(q||p)$ is the KL divergence between $q$ and $p$. Also $P_{\text{Haar}}(F)$ is the probability distribution of the fidelity $F = |\langle \psi|\psi' \rangle|^2$, where $|\psi\rangle$ and $|\psi'\rangle$ are sampled according to the Haar measure. In Ref. [14], $P_{\text{Haar}}(F) = (N - 1)(1 - F)^{N-2}$ is derived, where $N$ is the dimension of Hilbert space. Because in general $D_{KL}(q||p) = 0$ iff $q = p$, the ansatz $C$ with smaller value of $\mathcal{E}(C)$ has a higher expressibility. Thus, $\mathcal{E}(C)$ can also be used as an indicator for quantifying the non-uniformity of an ansatz.

Lastly recall that the $t$-th moment of $P(C,F)$ and $P_{\text{Haar}}(F)$ are $\mathcal{F}(t)(C)$ and $\mathcal{F}(t)_{\text{Haar}}(N)$, respectively. Thus, if the values of $\mathcal{F}(t)(C)$ is close to $\mathcal{F}(t)_{\text{Haar}}(N)$, the value of $\mathcal{E}(C)$ is close to zero.

2.2 Ansätze

Here we describe the three types of ansätze investigated in this paper.

Hardware Efficient Ansatz

The HEA circuit consists of multiple layers of parametrized single qubit gates and entanglers which entangle all qubits. In the following, let $C^{\ell,m,n}_{\text{HEA}}$ be the class of HEA with $\ell$ layers where each layer contains $n$-qubits.

Alternating Layered Ansatz

The ALT introduced in [8] also consists of multiple layers but the components of each layer are different from those of the HEA as follows. That is, each layer has some separated blocks, where each block has parametrized single-qubit rotation gates and fixed entanglers that entangle all qubits inside the block. The probability distributions of those angle parameters are independent in all blocks and all layers. In this paper, we further restrict the class of ALT as follows. First, as in HEA, the entire circuit is composed of probability distributions of those angle parameters are independent in all blocks and all layers. In this paper, parametrized single-qubit rotation gates and fixed entanglers that entangle all qubits inside the block. The components of each layer are different from those of the HEA as follows. That is, each layer has some separated blocks, where each block operates on $\ell$ $n/m$ blocks which operate on $\{1, \ldots, m\}$, $\{m+1, \ldots, 2m\}$, ..., and $\{n-m+1, \ldots, n\}$ qubits. As for the even-number-labeled layers, they contain $n/m+1$ blocks which operate on $\{1, \ldots, m/2\}$, $\{m/2+1, \ldots, 3m/2\}$, ..., and $\{n-m/2+1, \ldots, n\}$ qubits; that is, the first and the last block operate on $m/2$ qubits, but the others operate on $m$ qubits. In the following, we use $C^{\ell,m,n}_{\text{ALT}}$ to denote the class of ALT with the above-defined indices.

In [8], it is proved that the vanishing gradient problem can be avoided if the following conditions are satisfied; (i) each term of the cost Hamiltonian, $H = \sum_k H_k$ is local, meaning that each term $H_k$ is composed of less than $m$ neighboring qubits, (ii) the ensemble of unitary matrices in each block is 2-design, and (iii) the number of layers, $\ell$, is of the order $O(\text{poly}(\log n))$. Note that each block needs to be deep enough for the distribution of corresponding unitary matrices to be close to 2-design.

Tensor Product Ansatz

In addition to HEA and ALT, we here introduce the class of tensor product ansatz (TEN) as a relatively weak ansatz. This ansatz also consists of $\ell$ layers, where each layer contains $n$ qubits, and each layer contains $n/m$ blocks ($n/m$ is assumed to be an integer), which contain single-qubit rotation gates and entanglers combining all qubits in the block. Throughout all the layers, the blocks operate on $\{1, \ldots, m\}$, $\{m+1, \ldots, 2m\}$, ..., and $\{n-m+1, \ldots, n\}$ qubits. Thus, TEN always generates a product state of the form $|\psi_1\rangle \otimes \cdots \otimes |\psi_{n/m}\rangle$ where each state is composed of $m$ qubits. Let $C^{\ell,m,n}_{\text{TEN}}$ be the class of TEN described above.

In Fig. [4] we show examples of the structures of the above defined ansätze.

3 Expressibility of the circuit ansätze

In this section, we give some analytical expressions as well as upper bounds of the first and the second frame potentials of the three ansätze introduced in Section 2.2 showing that the shallow ALT has almost the same expressibility as that of HEA. This result will be further confirmed by a numerical simulation in terms of the KL-divergence.
Figure 1: Examples of TEN, ALT and HEA introduced in Section 2.2. The upper figures are the overall structures of the ansatzes in the case $n = 8$ and $\ell = 3$ (and $m = 4$ for TEN and ALT). The circuits inside the blocks are exemplified in the lower figure, where $\sigma_{a_i}$ ($a_i \in \{x, y, z\}$) is the Pauli matrix operating on the qubit and $\theta_i \in [0, 2\pi]$ is the parameter.

3.1 Analytical expression of the frame potential of the ansatzes

First of all, to compute the frame potentials of each ansatz, we assume that the ensemble of the unitary matrices corresponding to $C_{\text{HEA}}^{m,n}$ is 2-design. Similarly, the ensemble of the unitary matrices corresponding to each block of $C_{\text{ALT}}^{\ell,m,n}$ and $C_{\text{TEN}}^{\ell,m,n}$ are assumed to be 2-design. The assumptions are also adopted in the discussion of [4] and [8]. Such ensemble of unitary matrices can be generated by randomly choosing the parameters of the circuit having a specific structure.

Before going into the detail, we show some integration formula for random unitary matrices [15]. First, if the ensemble of $n \times n$ unitary matrices $\{U\}$ is 1-design, the following formula holds:

$$\int_{1\text{-design}} dUU_1U_2^* = \frac{\delta_{11}\delta_{22}}{n},$$  \hspace{1cm} (4)

where $\int_{1\text{-design}} dU$ is the integral over the 1-design ensemble of the unitary matrices. Second, if the ensemble of $n \times n$ unitary matrices $\{U\}$ is 2-design, the following formula holds:

$$\int_{2\text{-design}} dUU_{1_1j_1}U_{1_2j_2}U_{1_3j_3}U_{1_4j_4}^*U_{1_5j_5}^*U_{1_6j_6}^*U_{1_7j_7}^*U_{1_8j_8}^* = \sum_{k=1}^{4} \lambda_k^{(n)} \Delta_{1_{11j_1j_2j_3j_4j_5j_6j_7j_8}}^k,$$  \hspace{1cm} (5)

where

$$\lambda_1^{(n)} = \lambda_2^{(n)} = \frac{1}{2^{2n} - 1}, \quad \lambda_3^{(n)} = \lambda_4^{(n)} = \frac{1}{(2^{2n} - 1)2^{n}},$$

$$\Delta_{1_{11j_1j_2j_3j_4j_5j_6j_7j_8}}^1 = \delta_{11} \delta_{j_1j_2j_3j_4j_5j_6j_7j_8},$$

$$\Delta_{1_{11j_1j_2j_3j_4j_5j_6j_7j_8}}^2 = \delta_{11} \delta_{j_3j_4j_5j_6j_7j_8} \delta_{j_2j_1j_6j_5j_7j_8},$$

$$\Delta_{1_{11j_1j_2j_3j_4j_5j_6j_7j_8}}^3 = \delta_{11} \delta_{j_3j_4j_5j_6j_7j_8} \delta_{j_2j_1j_6j_5j_7j_8},$$

$$\Delta_{1_{11j_1j_2j_3j_4j_5j_6j_7j_8}}^4 = \delta_{11} \delta_{j_3j_4j_5j_6j_7j_8} \delta_{j_2j_1j_6j_5j_7j_8},$$  \hspace{1cm} (6)

and $\int_{2\text{-design}} dU$ is the integral over the 2-design ensemble of the unitary matrices. These formula are effectively used to derive the theorems shown below.
The First Frame Potential

For the first frame potentials, the following theorem holds.

**Theorem 1.** If the ensemble of the unitary matrices corresponding to $C^{d,n}_{\text{HEA}}$ and the ensemble of the unitary matrices corresponding to each block of $C_{\text{ALT}}^{d,m,n}$ and $C_{\text{TEN}}^{d,m,n}$ are 2-design, then the following equalities hold:

$$F^{(1)}(C^{d,n}_{\text{HEA}}) = F^{(1)}(C^{d,m,n}_{\text{ALT}}) = F^{(1)}(C^{d,m,n}_{\text{TEN}}) = F^{(1)}_{\text{Haar}}(2^n).$$  (7)

The equality, $F^{(1)}(C^{d,n}_{\text{HEA}}) = F^{(1)}_{\text{Haar}}(2^n)$, can be readily proved from the assumption of the theorem, because, if the ensemble of the unitary matrices corresponding to the circuit is 2-design, the ensemble of the states generated by the circuit is a state 2-design (and therefore a state 1-design). For the other equalities, the proof is given in Appendix. Note that, accordingly, the ensembles of the states generated by $C^{d,n}_{\text{HEA}}, C_{\text{ALT}}^{d,m,n}$, and $C_{\text{TEN}}^{d,m,n}$ are all 1-design.

The Second Frame Potential

The second frame potential of the Haar random circuits, $F^{(2)}_{\text{Haar}}(2^n)$, can be computed as

$$F^{(2)}_{\text{Haar}}(2^n) = \int_0^1 dP dQ (2^n - 1)(1 - P)^{2^n - 2} = \frac{1}{(2^n + 1)^{2^n - 1}}.$$  (8)

Then, for $C^{d,n}_{\text{HEA}}$ and $C_{\text{TEN}}^{d,m,n}$, the following theorem holds.

**Theorem 2.** If the ensemble of the unitary matrices corresponding to $C^{d,n}_{\text{HEA}}$ and the ensemble of the unitary matrices corresponding to each block of $C_{\text{ALT}}^{d,m,n}$ and $C_{\text{TEN}}^{d,m,n}$ are 2-design, then the following equalities hold:

$$F^{(2)}(C^{d,n}_{\text{HEA}}) = F^{(2)}_{\text{Haar}}(2^n),$$  (9)

$$F^{(2)}(C^{d,m,n}_{\text{TEN}}) = 2^{n-1} (\frac{2^n + 1}{2^{m+1}}) F^{(2)}_{\text{Haar}}(2^n).$$  (10)

The equality [9] is readily proved from the assumption of the theorem, because, as we mentioned above, if the ensemble of the unitary matrices corresponding to the circuit is 2-design, the ensemble of the states generated by the circuit is a state 2-design. For Eq. [10], we give the proof in Appendix. From this theorem we find that $F^{(2)}(C_{\text{TEN}}^{d,m,n})$ is always larger than $F^{(2)}_{\text{Haar}}(2^n)$; in particular, $F^{(2)}(C_{\text{ALT}}^{d,m,n}) \approx 2^{n/m - 1} F^{(2)}_{\text{Haar}}(2^n)$ for large $n$, meaning that the expressibility of TEN is much smaller than that of HEA in the sense of the frame potential.

As for ALT, it is difficult to obtain an explicit formula like the case of HEA and TEN. Hence in Theorem 3 below, we provide a formula for computing the values of $F^{(2)}(C_{\text{ALT}}^{d,m,n})$ and $F^{(2)}(C_{\text{TEN}}^{d,m,n})$; the computation methods for the other $f$s are left for future work. Before stating the theorem, we define a 16-dimensional vector $a(2,m)$, a $16 \times 16$ matrix $B(2,m)$, a 64-dimensional vector $a(3,m)$ and a $64 \times 64$ matrix $B(3,m)$. Given integers $k_a, k_b \in \{1, 2, 3, 4\}$, the $(4(k_a - 1) + k_b)$-th component of the vector $a(2,m)$ is defined as

$$a(2,m)_{4(k_a-1)+k_b} = \int_{2\text{design}} dP dQ \sqrt{\lambda^{(m/2)}_{k_a} \lambda^{(m/2)}_{k_b} \Delta^{(k_a,k_b)}(P,Q)},$$  (11)

where $\Delta^{(k_a,k_b)}(P,Q)$ is the function of $m/2 \times m/2$ unitary matrices $P$ and $Q$:

$$\Delta^{(k_a,k_b)}(P,Q) = \sum_{i,j} \Delta^{(k_a)}_{01230123} \Delta^{(k_b)}_{01230123} P_{j_1,j_2} P_{j_3,j_4} P_{j_5,j_6} Q_{j_1,j_2} Q_{j_3,j_4} Q_{j_5,j_6}.$$  (12)

Next, given integers $k_a, k_b, k_c \in \{1, 2, 3, 4\}$, the $(16(k_a - 1) + 4(k_b - 1) + k_c)$-th component of the vector $a(3,m)$ is defined as

$$a(3,m)_{16(k_a-1)+4(k_b-1)+k_c} = \int_{2\text{design}} dP dQ \sqrt{\lambda^{(m/2)}_{k_a} \lambda^{(m/2)}_{k_b} \lambda^{(m/2)}_{k_c} \Delta^{(k_a,k_b,k_c)}(P,Q)},$$  (13)

where $\Delta^{(k_a,k_b,k_c)}(P,Q)$ is a function of $m/2 \times m/2$ unitary matrices $P$ and $Q$:

$$\Delta^{(k_a,k_b,k_c)}(P,Q) = \sum_{i,j} \Delta^{(k_a)}_{01230123} \Delta^{(k_b)}_{01230123} \Delta^{(k_c)}_{01230123} P_{j_1,j_2} P_{j_3,j_4} P_{j_5,j_6} Q_{j_1,j_2} Q_{j_3,j_4} Q_{j_5,j_6}.$$  (14)

Also, given integers $k_a, k_b, k_c, k_d \in \{1, 2, 3, 4\}$, the $(4(k_a - 1) + 4k_b, 4(k_c - 1) + k_d)$-th component of the matrix $B(m, 2)$ is defined as

$$B(2,m)_{4(k-a-1)+k_b,4(k-c-1)+k_d} = \int_{2\text{design}} dP dQ \sqrt{\lambda^{(m/2)}_{k_a} \lambda^{(m/2)}_{k_b} \lambda^{(m/2)}_{k_c} \lambda^{(m/2)}_{k_d} \Delta^{(k_a,k_b,k_c,k_d)}(P,Q)},$$  (15)
Figure 2: The values of $\mathcal{F}^{(2)}(C)/\mathcal{F}^{(2)}_{\text{Haar}}(2^2)$ for ALT and TEN. The left and right figures show the case $\ell = 2$ and $\ell = 3$, respectively.

where $\Delta^{(k_a, k_b, k_c, k_d)}(P, Q)$ is a function of $m \times m$ matrices $P$ and $Q$:

$$
\Delta^{(k_a, k_b, k_c, k_d)}(P, Q) = \sum_{i,j,a,t} \Delta_{a}^{k_a} \Delta_{b}^{k_b} \Delta_{c}^{k_c} \Delta_{d}^{k_d} 
\times P_{(i_1, j_1)}^{a} P_{(i_2, j_2)}^{a} P_{(i_3, j_3)}^{a} P_{(i_4, j_4)}^{a} Q_{(m_1, n_1)}^{a} Q_{(m_2, n_2)}^{a} Q_{(m_3, n_3)}^{a} Q_{(m_4, n_4)}^{a}.
$$

(16)

For the matrix component $M^{(l)}_{i,j}$, the upper indices correspond to the first $m/2$ qubits and the lower indices correspond to the last $m/2$. Given integers $k_a, k_b, k_c, k_d, k_f \in \{1, 2, 3, 4\}$ the $(16(k_a - 1) + 4(k_b - 1) + k_c, 16(k_d - 1) + 4(k_e - 1) + k_f)$-th component of the matrix $B(m, 3)$ is defined as

$$
B(3, m)_{16(k_a - 1) + 4(k_b - 1) + k_c, 16(k_d - 1) + 4(k_e - 1) + k_f} = \int_{2\text{dHaar}} \! dP \! dQ \! \sqrt{\lambda_{k_a}^{(m)} \lambda_{k_b}^{(m)} \lambda_{k_c}^{(m)} \lambda_{k_d}^{(m)} \lambda_{k_f}^{(m)}} \Delta^{(k_a, k_b, k_c, k_d, k_e, k_f)}(P, Q),
$$

(17)

where $\Delta^{(k_a, k_b, k_c, k_d, k_e, k_f)}(P, Q)$ is a function of $m \times m$ matrices $P$ and $Q$:

$$
\Delta^{(k_a, k_b, k_c, k_d, k_e, k_f)}(P, Q) = \sum_{i,j,a,t} \Delta_{a}^{k_a} \Delta_{b}^{k_b} \Delta_{c}^{k_c} \Delta_{d}^{k_d} \Delta_{e}^{k_e} \Delta_{f}^{k_f} 
\times P_{(i_1, j_1)}^{a} P_{(i_2, j_2)}^{a} P_{(i_3, j_3)}^{a} P_{(i_4, j_4)}^{a} Q_{(m_1, n_1)}^{a} Q_{(m_2, n_2)}^{a} Q_{(m_3, n_3)}^{a} Q_{(m_4, n_4)}^{a}.
$$

(18)

Now we can give the theorem as follows.

**Theorem 3.** If the ensemble of the unitary matrices corresponding to each block of $C^{3, m, n}_{\text{ALT}}$ and $C^{3, m, n}_{\text{ALT}}$ are 2-design, then we have

$$
\mathcal{F}^{(2)}(C^{3, m, n}_{\text{ALT}}) = a(3, m)^T B(3, m) a(3, m),
$$

(19)

$$
\mathcal{F}^{(2)}(C^{3, m, n}_{\text{ALT}}) = a(3, m)^T B(3, m) a(3, m).
$$

(20)

The vectors $a(\ell, m)$ and the matrices $B(\ell, m)$ for $\ell = 2, 3$ are obtained by directly computing Eqs. (11, 13, 15, and 17), then we lead to $\mathcal{F}^{(2)}(C^{3, m, n}_{\text{ALT}})$ and $\mathcal{F}^{(2)}(C^{3, m, n}_{\text{ALT}})$. Now our interest is in the gap of these quantities from $\mathcal{F}^{(2)}(\text{Haar}(2^2))$, to see the expressibility of ALT. For this purpose, in Fig. 2 we show $\mathcal{F}^{(2)}(C^{3, m, n}_{\text{ALT}})/\mathcal{F}^{(2)}(\text{Haar}(2^2))$ and $\mathcal{F}^{(2)}(C^{3, m, n}_{\text{ALT}})/\mathcal{F}^{(2)}(\text{Haar}(2^2))$ as a function of $n/m$, for several values of $(m, n)$. For comparison, $\mathcal{F}^{(2)}(C^{3, m, n}_{\text{ALT}})/\mathcal{F}^{(2)}(\text{Haar}(2^2))$ and $\mathcal{F}^{(2)}(C^{3, m, n}_{\text{ALT}})/\mathcal{F}^{(2)}(\text{Haar}(2^2))$ are shown in the figure. Recall that, if this measure takes a smaller value, this means that the corresponding ansatz has a higher expressibility. Here is the list of notable points:

- For any pair of $(n, m)$, it is clear that $\mathcal{F}^{(2)}(C^{3, m, n}_{\text{ALT}})$ is much smaller than $\mathcal{F}^{(2)}(C^{3, m, n}_{\text{ALT}})$ for both $\ell = 2, 3$. This means that, as expected, ALT has a much higher expressibility than TEN.
Corollary 1, which can be readily derived from the theorem, provides a means for evaluating such the number \( m \). The last point is of particular importance in our scenario. That is, we are concerned with the condition on the ensemble of the unitary matrices corresponding to each block of \( F \) to those of \( P \) random. (Recall that, if every block is completely random, then the set of HEA constitutes the Haar ensemble.) For any pair of \( (n, m) \), \( C^{2,m,n}_{\text{ALT}} > F^{(2)}(C^{3,m,n}_{\text{ALT}}) \) hold, i.e., as \( \ell \) increases, the expressibility increases.

For any fixed \( n/m \), the ALT with bigger \( m \) always has a higher expressibility. For instance, the ALT with \( (n, m) = (50, 10) \) has a higher expressibility than the ALT with \( (n, m) = (20, 4) \). This is simply because, if the structure of the circuit (the number of division in each layer for making the block) is the same, then an ALT with bigger block components has a higher expressibility.

For a fixed \( n \), we have ALT with the smaller second order frame potential by taking \( m \) bigger. For instance \( n = 100 \), we have \( F^{(2)}(C^{2,100,100}_{\text{ALT}}) > F^{(2)}(C^{2,4,100}_{\text{ALT}}) > F^{(2)}(C^{2,10,100}_{\text{ALT}}) \). That is, for a limited number of available qubits, the ALT with less blocks has a higher expressibility.

If \( F^{(2)}(C^{\ell,m,n}_{\text{ALT}}) \geq F^{(2)}(C_{\text{Haar}}(2^n)) \) when \( m = 10 \) for all \( n/m \) within the figure and for both \( \ell = 2, 3 \). Hence the ALT composed from the blocks with \( m = 10 \) qubits in each layer has almost the same expressibility as the HEA without respect to the total qubits number, \( n \). In other words, for a given HEA with fixed \( n \), we can divide each layer into separated 10-qubits blocks to make an ALT, without decreasing the expressibility.

The last point is of particular importance in our scenario. That is, we are concerned with the condition on the number \( m \) such that \( F^{(2)}(C^{2,m,n}_{\text{ALT}}) \simeq F^{(2)}(C_{\text{Haar}}(2^n)) \) holds. The following Theorem 4 and the subsequent Corollary 1, which can be readily derived from the theorem, provide a means for evaluating such \( m \).

**Theorem 4.** If the ensemble of the unitary matrices corresponding to each block of \( C^{2,m,n}_{\text{ALT}} \) and \( C^{3,m,n}_{\text{ALT}} \) are 2-design, then the following inequalities hold:

\[
F^{(2)}(C^{2,m,n}_{\text{ALT}}) < \left(1 + \frac{1}{2^n}\right) \left(1 + \frac{1.2}{2^m}\right)^2 \left(1 + \frac{20.8}{2^m/2} \right)^{\frac{n}{m} - 1} - 1 \right) F^{(2)}(C_{\text{Haar}}(2^n)).
\]

If \( m = 2a \log_2 n + 143/(an^{a-1} \log_2 n) < 1 \),

\[
F^{(2)}(C^{3,m,n}_{\text{ALT}}) < \left(1 + \frac{1}{2^n}\right) \left(1 + \frac{1.2}{2^m}\right)^2 \left(1 + \frac{143}{an^{a-1} \log_2 n}\right) F^{(2)}(C_{\text{Haar}}(2^n)).
\]

Recall from Eq. 4 that \( F^{(1)}(C) \geq F^{(1)}(C_{\text{Haar}}(N)) \) holds for any ansatz \( C \). Therefore, if \( m \geq 4 \log n \) and \( n \) is enough large, Corollary 1 implies that \( F^{(2)}(C^{2,m,n}_{\text{ALT}}) \sim F^{(2)}(C_{\text{Haar}}(2^n)) \) and \( F^{(2)}(C^{3,m,n}_{\text{ALT}}) \sim F^{(2)}(C_{\text{Haar}}(2^n)) \). This means that the ensembles of the states generated by \( C^{2,m,n}_{\text{ALT}} \) and \( C^{3,m,n}_{\text{ALT}} \) are almost 2-design. Hence in this case, from Theorem 2, the expressibility of ALT is as high as that of HAE. It is worth mentioning that, when \( m = O(\log_2 n) \), the vanishing gradient problem does not happen in ALT as long as the cost function is local and \( \ell \) is small enough. More precisely, it was shown there that the variance of the gradient of such a cost function is larger than the value proportional to \( O(1/2^m) \); thus, by taking \( m = O(\log_2 n) \), the variance decreases with only \( O(1/poly(n)) \) as a function of \( n \), whereas in the HEA case the same variance decreases exponentially fast as \( n \) becomes large. Therefore, the expressibility and the trainability coexist in the shallow ALT with \( m = O(\log_2 n) \).

**3.2 Expressibility measured by KL divergence**

In Subsection 3.1, we have shown that the first two moments of \( P(C^{\ell,m,n}_{\text{ALT}}, F) \) and \( P(C^{\ell,m,n}_{\text{HEA}}, F) \) are close to those of \( P_{\text{Haar}}(F) \), as long as \( m = O(\log_2 n) \) and the block components of the ansatz are sufficiently random. (Recall that, if every block is completely random, then the set of HEA constitutes the Haar ensemble.) The result implies that both \( P(C^{\ell,m,n}_{\text{ALT}}, F) \) and \( P(C^{\ell,m,n}_{\text{HEA}}, F) \) are close to \( P_{\text{Haar}}(F) \) itself, meaning that \( P(C^{\ell,m,n}_{\text{ALT}}, F) \simeq P(C^{\ell,m,n}_{\text{HEA}}, F) \simeq P_{\text{Haar}}(F) \). In this subsection, to support this conjecture, we evaluate the values of KL-divergence \( E(C) = D_{KL}(P(C, F) || P_{\text{Haar}}(F)) \) for the case \( C = C^{\ell,m,n}_{\text{ALT}} \) and \( C = C^{\ell,m,n}_{\text{HEA}} \), in addition to \( C = C^{\ell,m,n}_{\text{TEN}} \) for comparison with various sets of \( (\ell, m, n) \). Especially, we focus on the relationship between the values of \( F^{(2)}(C) \) and \( E(C) \), and check if \( F^{(2)}(C) \simeq F^{(2)}(C') \) would lead to \( E(C) \simeq E(C') \) for a fixed \( n \).

The parameters taken for calculating the KL-divergence are summarized in Table 1. Note that the circuits are chosen to be similar to those used in Section 3.1 for TEN and ALT; the depth of the circuits inside the blocks are all set to \( n \) so that the ensemble of the unitary matrices corresponding to those circuits become close to 2-design [16] [17]; for HEA, \( \ell \) is set to \( n \) so that the ensemble of the unitary matrices corresponding to the whole circuits becomes close to 2-design. It is expected that \( F^{(2)}(C^{2,2,n}_{\text{ALT}}) \simeq F^{(2)}(C^{3,2,n}_{\text{HEA}}) \)
| n  | Ansatz | $\ell$ | $m$ | Depth of each block | # of gate parameters | Example of the circuit |
|----|--------|------|-----|-------------------|---------------------|------------------------|
| 4  | TEN    | 2    | 2   | 2                 | 16                  | -                      |
|    |        | 3    | 2   | 2                 | 24                  | Fig. 3a                |
|    | ALT    | 2    | 2   | 2                 | 16                  | -                      |
|    |        | 3    | 2   | 2                 | 24                  | Fig. 3b                |
|    | HEA    | 4    | -   | -                 | 16                  | Fig. 3c                |
| 6  | TEN    | 2    | 2   | 2                 | 24                  | -                      |
|    |        | 3    | 2   | 2                 | 36                  | -                      |
|    | ALT    | 2    | 2   | 2                 | 24                  | -                      |
|    |        | 3    | 2   | 2                 | 36                  | -                      |
|    | HEA    | 6    | -   | -                 | 36                  | -                      |
| 8  | TEN    | 2    | 2   | 2                 | 32                  | -                      |
|    |        | 4    | 4   | 64                | -                   |
|    | ALT    | 2    | 2   | 2                 | 32                  | -                      |
|    |        | 4    | 4   | 96                | -                   |
|    |        | 3    | 2   | 4                 | 46                  | -                      |
|    | HEA    | 8    | -   | -                 | 64                  | -                      |

Table 1: Parameters chosen for computing the KL-divergence. The number of gate parameters are computed by $n \times \ell \times \text{(Depth of each block)}$ for TEN and ALT, and $n \times \ell$ for HEA.

and $F^{(2)}(C_{\text{TEN}}^{3,4,8}) \approx F^{(2)}(C_{\text{HEA}}^{8})$ are realized, because, in Fig. 2, we see that $F^{(2)}(C_{\text{TEN}}^{3,2,4})$ and $F^{(2)}(C_{\text{ALT}}^{3,4,8})$ are almost equals to the Haar values when the ensembles of unitary matrices corresponding to each block are 2-design. Thus, we here check if $F^{(2)}(C) \approx F^{(2)}(C')$ would mean $\mathcal{E}(C) \approx \mathcal{E}(C')$ in these parameter sets. As an example of the circuit, the whole structure of $C_{\text{TEN}}^{3,2,4}$, $C_{\text{ALT}}^{3,2,4}$, and $C_{\text{HEA}}^{3,2,4}$ in our settings are shown in Figs. 3a, 3b, and 3c, respectively. As illustrated in the figures, each layer is composed of parametrized single qubit gates and fixed 2-qubit CNOT gates.

In each trial of computing KL-divergence, we generate 200 states. When generating a state in each trial, we randomly choose the parameters and the type of single-qubit gate of the circuit. That is, for the qubit gates and fixed 2-qubit CNOT gates.

Figs. 3a, 3b, and 3c, respectively. As illustrated in the figures, each layer is composed of parametrized single qubit gates and fixed 2-qubit CNOT gates.

In this setting, Fig. 4 shows the KL divergences $\mathcal{E}(C_{\text{TEN}}^{3,4,n})$, $\mathcal{E}(C_{\text{HEA}}^{3,4,n})$, and $\mathcal{E}(C_{\text{TEN}}^{3,4,n})$. As a reference, we also show the values of $F^{(2)}(C)/F^{(2)}(\text{Haar}(2^n))$ computed from the second moment of the fidelity distributions.

Each data point and associated error bar is the average and the standard deviation of 10 trials of computation, respectively. Here is the list of points:

- For a fixed $n$, $\mathcal{E}(C_{\text{TEN}}^{3,4,n})$ is always bigger than $\mathcal{E}(C_{\text{TEN}}^{3,4,n})$ and $\mathcal{E}(C_{\text{HEA}}^{3,4,n})$.
- As the number of layers increases, the KL-divergence decreases for fixed $(m, n)$.
- For a fixed $n$, the tendency of the values of $F^{(2)}(C)/F^{(2)}(\text{Haar}(2^n))$ is strongly correlated with that of KL-divergence.
- As expected, $F^{(2)}(C_{\text{ALT}}^{3,4,m,n}) \approx F^{(2)}(C_{\text{HEA}}^{3,4,m,n})$ is realized when $(m, n) = (2, 4)$ and $(4, 8)$.
- $\mathcal{E}(C_{\text{ALT}}^{3,4,m,n})$ is as small as $\mathcal{E}(C_{\text{HEA}}^{3,4,m,n})$ in the parameter sets where $F^{(2)}(C_{\text{ALT}}^{3,4,m,n}) \approx F^{(2)}(C_{\text{HEA}}^{3,4,m,n})$ is realized, i.e., $(m, n) = (2, 4)$ and $(4, 8)$. This result implies that the state distribution in ALT almost close to that in HEA, in the setting where the second frame potential is close to $F^{(2)}(\text{Haar}(2^n))$.

From some of the above observations, we find the strong correlation between $F^{(2)}(C)/F^{(2)}(\text{Haar}(2^n))$ and $\mathcal{E}(C)$; that is, as $F^{(2)}(C)$ becomes close to 1, then $\mathcal{E}(C)$ becomes close to 0. Therefore, combining the result obtained in Section 3.1, we get a clear evidence that, as far as $m = O(\log_2 n)$, $F^{(2)}(C_{\text{TEN}}^{3,4,m,n}) \approx 0$ and $\mathcal{E}(C_{\text{ALT}}^{3,4,m,n}) \approx 0$ hold. That is, the high expressibility and trainability in ALT proven in Section 3.1 are assured also in terms of KL-divergence.
(a) The structure of $C_{\text{TEN}}^{3,2,4}$. The red boxes correspond to blocks.

(b) The structure of $C_{\text{ALT}}^{3,2,4}$. The red boxes correspond to blocks.

(c) The structure of $C_{\text{HEA}}^{4,4}$.

Figure 3: The structures of ansatzes in our setting.

Figure 4: $F(2)(C)/F_{\text{TEN}}^{2}(2^n)$ (top) and KL-divergence (bottom) for each ansatz. The sets of points with which $F^{(2)}(C_{\text{ALT}}^{3,m,n}) \approx F^{(2)}(C_{\text{HEA}}^{3,m,n})$ hold are enclosed by the red rectangles.
4 Application to VQE

Recall that ALT was originally introduced with the motivation to resolve the vanishing gradient problem in VQE, which has been often observed when using HEA; then we were concerned with the expressibility of ALT in VQE, meaning that the ALT does not offer a chance to reach the optimal solution due to the possible loss of expressibility. But we now know that this concern has been resolved under some conditions, as concluded in the previous section; that is, the expressibility and the trainability coexists in the shallow ALT with \( m = O(\log_2 n) \). In this section, let us see that this desirable fact indeed holds in a VQE problem.

We choose the Hamiltonian of 4-qubits Heisenberg model on a 1-dimensional lattice with periodic boundary conditions:

\[
\mathcal{H} = \sum_{i=1}^{4} (\sigma_i^x \sigma_i^{x+1} + \sigma_i^y \sigma_i^{y+1} + \sigma_i^z \sigma_i^{z+1}),
\]

where \( \sigma_i^a (a \in \{x,y,z\}) \) is the Pauli matrix that operates on the \( i \)-th qubit and \( \sigma_5^a = \sigma_i^a \). The goal of VQE problem is to find the minimum eigenvalue of \( \mathcal{H} \), by calculating the mean energy \( \langle \mathcal{H} \rangle = \langle \psi_\theta | \mathcal{H} | \psi_\theta \rangle \) via a quantum computer and updating the parameter \( \theta \in \Theta \) to decrease \( \langle \mathcal{H} \rangle \) via a classical computer, in each iteration. As ansatzes, \( \mathcal{C}^{3,2,4}_{\text{TEN}}, \mathcal{C}^{3,2,4}_{\text{ALT}} \) and \( \mathcal{C}^{4,4}_{\text{HEA}} \) are chosen. As indicated in Fig. 4, the values of KL divergence corresponding to these ansatzes show that \( E(\mathcal{C}^{3,2,4}_{\text{TEN}}) > E(\mathcal{C}^{3,2,4}_{\text{ALT}}) \approx E(\mathcal{C}^{4,4}_{\text{HEA}}) \). That is, this ALT has the expressibility as high as that of HEA, and further, it is expected to enjoy the trainability unlike the HEA.

The simulation results are shown in Fig. 5. The blue lines and the associated error bars represent the average and the standard deviation of the mean energies in total 100 trials, respectively; in each trial, the initial parameters of the ansatz are randomly chosen; also the optimization to decrease \( \langle \mathcal{H} \rangle \) in each iteration is performed by using the Adam Optimizer with learning rate 0.001 [18]. The green line shows the theoretical minimum energy (i.e., the ground energy) of \( \mathcal{H} \). In each subfigure, the orange line is constructed by combining trajectories having the least mean energy at each iteration; in other words, it is the minimum-energy envelope of the whole 100 trajectories.

The ansatz \( \mathcal{C}^{3,2,4}_{\text{TEN}} \), which has the least expressibility in the sense of frame potentials and the KL divergence analysis, clearly gives the worst result; its least mean-energy is far above from the ground energy. This is simply because the state generated via \( \mathcal{C}^{3,2,4}_{\text{TEN}} \) cannot represent the ground state for any parameter choice. The result on \( \mathcal{C}^{4,4}_{\text{HEA}} \) is the second worst, which also does not reach the ground energy as in the case of TEN. Note that increasing the number of parameters does not change this result; we also executed the simulation with \( \mathcal{C}^{4,6}_{\text{HEA}} \) that has the same number of parameters as \( \mathcal{C}^{3,2,4}_{\text{ALT}} \) but did not find a better result than that of \( \mathcal{C}^{4,4}_{\text{HEA}} \). On the other hand, \( \mathcal{C}^{3,2,4}_{\text{ALT}} \) succeeds in finding the ground state; in fact, 5 of the total 100 trajectories generated via this ALT reach the ground energy. Even though HEA can represent the ground state by suitably choosing the parameters, it failed to find the optimal parameters due to being trapped in a plateau of the mean-energy landscape, i.e., due to the vanishing gradient problem. In contrast, \( \mathcal{C}^{3,2,4}_{\text{ALT}} \) can circumvent the plateau, as predicted in [3], and at the same time, it can represent the ground state. Therefore, we end up with a clear evidence that the expressibility and the trainability certainly coexist in ALT.
5 Conclusion

This paper has examined the expressibility power of the shallow ALT, which was proposed before as a solution to the vanishing gradient problem found in various types of hybrid quantum-classical algorithms. Our conclusion is that, in addition to such a well trainability, shallow ALTs have the expressibility; that is, in the measure of the frame potential and the KL-divergence, those shallow ALTs have almost the same expressibility as HEAs, which are often used in hybrid algorithms but suffer from the vanishing gradient problem. In particular, we have proven that such expressibility holds if the number of entangled qubits in each block is of the order of the logarithm of the number of all resource qubits, which is consistent to the previous result discussing the trainability of ALT. We also confirmed that the ALT certainly enjoys both the expressibility and the trainability in VQE.

Even though our results are limited to the case \( \ell = 2, 3 \), we have numerically observed that the ALT acquires even higher expressibility when making \( \ell \) bigger. Therefore, we conjecture that the above conclusion still holds for ALT with \( \ell \geq 4 \). The rigorous proof is left for future work.

This work is supported by the MEXT Quantum Leap Flagship Program Grant Number JPMXS0118067285.

A Proof of Theorems

Notation: For the unitary matrix \( U_a \), corresponding to the entire circuit, we denote the unitary matrix corresponding to the \( i \)-th layer of the circuit to be \( U_a(i) \) and the unitary matrix corresponding to the \( j \)-th block in the \( i \)-th layer as \( U_a(i, j) \).

A.1 Proof of Theorem 1

First, the value of \( \mathcal{F}^{(1)}_{\text{Haar}}(2^n) \) is straightforwardly computed as

\[
\mathcal{F}^{(1)}_{\text{Haar}}(2^n) = \int_0^1 dFF(2^n - 1)(1 - F)^{2^n - 2} = \frac{1}{2^n}. \tag{26}
\]

Next, we provide the proof of \( \mathcal{F}^{(1)}(C^{m, n}_{\text{ALT}}) = \mathcal{F}^{(1)}_{\text{Haar}}(2^n) \). Given two final states |\( \phi \rangle = U_a|0\rangle \) and |\( \psi \rangle = U_b|0\rangle \) generated by \( C^{m, n}_{\text{ALT}} \), we have

\[
\mathcal{F}^{(1)}(C^{m, n}_{\text{ALT}}) = \int dU_a dU_b |0\rangle \langle 0| U_a U_b |0\rangle \langle 0|
= \int \left( \prod_{i=1}^\ell dU_a(i) \right) \left( \prod_{i=1}^\ell dU_b(i) \right)
\times |0\rangle U_a(1) U_a(2)^\dagger \ldots U_a(\ell)^\dagger U_b(\ell) \ldots U_b(2) U_b(1) \langle 0|
\times |0\rangle U_b(1)^\dagger U_b(2) \ldots U_b(\ell)^\dagger U_a(\ell) \ldots U_a(2) U_a(1) \langle 0|
\times |0\rangle U_a(1)^\dagger U_a(2)^\dagger \ldots U_a(\ell)^\dagger U_b(\ell) \ldots U_b(2) U_b(1) \langle 0|
\times |0\rangle U_b(1)^\dagger U_b(2)^\dagger \ldots U_b(\ell)^\dagger U_a(\ell) \ldots U_a(2) U_a(1) \langle 0|, \tag{27}
\]

where \( k(i) \) is the number of blocks in the \( i \)-th layer and each \( \int dU \) is the average over the ensemble of the unitary matrix \( U \). Because the distribution of each \( U_a(i, j) \) is 2-design (and is therefore 1-design), we can apply the formula (4) to the integrals with respect to \( U_a(i, j) \). Actually, by integrating \( \prod_{i=1}^\ell dU_a(\ell, j) \) for all \( \alpha \) in the last line of (27), we have

\[
\mathcal{F}^{(1)}(C^{m, n}_{\text{ALT}}) = \int \left( \prod_{i=1}^\ell \left( \prod_{i=1}^{k(i)} dU_a(i, j) \right) \right) \left( \prod_{i=1}^\ell \left( \prod_{j=1}^{k(i)} dU_b(i', j') \right) \right)
\times |0\rangle U_a(1)^\dagger U_a(2)^\dagger \ldots U_a(\ell)^\dagger U_b(\ell) \ldots U_b(2) U_b(1) \langle 0|
\times |0\rangle U_b(1)^\dagger U_b(2)^\dagger \ldots U_b(\ell)^\dagger U_a(\ell) \ldots U_a(2) U_a(1) \langle 0|
= \left( \frac{1}{2^n} \right)^\ell \int \left( \prod_{i=1}^{L-1} \left( \prod_{i=1}^{k(i)} dU_a^{i}\alpha \right) \right) \left( \prod_{i=1}^{L-1} \left( \prod_{j=1}^{k(i)} dU_b^{j}\beta \right) \right) \times 1
= \frac{1}{2^n} \mathcal{F}^{(1)}_{\text{Haar}}(2^n). \tag{28}
\]

The other equality in Eq. (7) can be proved in the same manner.
A.2 Proof of Theorem 2

Similar to the first frame potential, the value of $\mathcal{F}_{\text{Haa}}^{(1)}(2^n)$ is straightforwardly computed as

$$\mathcal{F}_{\text{Haa}}^{(2)}(2^n) = \int_0^1 dF F^2 (2^n - 1) (1 - F)^{2^n - 2} = \frac{1}{2^{m-1}/(2^m + 1)}.$$  \hspace{1cm} (29)

Next, we compute $\mathcal{F}_{\text{TEN}}^{(2)}(C^{d,m,n})$. Given two final states $|\phi\rangle = U_a|0\rangle$ and $|\psi\rangle = U_b|0\rangle$, it is computed as

$$\mathcal{F}_{\text{TEN}}^{(2)}(C^{d,m,n}) = \int_{2\text{design}} \left( \prod_{i=1}^{\ell} du_a(i,1)du_b(i,1) \right) \left| \langle 0|U_a(1,1)^T U_a(2,1)^T \cdots U_a(\ell,1)^T U_b(\ell,1) \cdots U_b(2,1) U_b(1,1) |0\rangle \right|^4$$

$$\times \int_{2\text{design}} \left( \prod_{i=1}^{\ell} du_a(i,2)du_b(i,2) \right) \left| \langle 0|U_a(1,2)^T U_a(2,2)^T \cdots U_a(\ell,2)^T U_b(\ell,2) \cdots U_b(2,2) U_b(1,2) |0\rangle \right|^4$$

$$\times \cdots \times \int_{2\text{design}} \left( \prod_{i=1}^{\ell} du_a(i,\frac{n}{m})du_b(i,\frac{n}{m}) \right) \left| \langle 0|U_a(1,\frac{n}{m}) U_a(2,\frac{n}{m}) \cdots U_a(\ell,\frac{n}{m}) U_b(\ell,\frac{n}{m}) \cdots U_b(2,\frac{n}{m}) U_b(1,\frac{n}{m}) |0\rangle \right|^4$$

$$= \left( \frac{1}{(2^m + 1)2^{m-1}} \right)^{\frac{n}{m}} = \frac{2^n + 1}{(2^m + 1)^n} \mathcal{F}_{\text{Haa}}^{(2)}(2^n).$$ \hspace{1cm} (30)

A.3 Proof of Theorem 3

Here we only show the computation of $\mathcal{F}_{\text{ALT}}^{(2)}(C^{d,m,n})$. The computation of $\mathcal{F}_{\text{ALT}}^{(2)}(C^{d,m,n})$ can be done in a similar manner. Given two final states $|\phi\rangle = U_a|0\rangle$ and $|\psi\rangle = U_b|0\rangle$,

$$\mathcal{F}_{\text{ALT}}^{(2)}(C^{d,m,n}) = \int dU_a(1) dU_a(2) dU_a(3) dU_b(1) dU_b(2) dU_b(3) \left| \langle 0|U_a(1)^T U_a(2)^T U_a(3)^T U_b(3) U_b(2) U_b(1) |0\rangle \right|^4$$

$$= \int_{2\text{design}} \left( \prod_{i=1}^{\ell} du_a(i,j) \right) \left( \prod_{j'=1}^{\ell} du_b(i',j') \right) \left| \langle 0|U_a(1)^T U_a(2)^T U_a(3)^T U_b(3) U_b(2) U_b(1) |0\rangle \right|^4.$$ \hspace{1cm} (31)

Executing integrals $\int_{2\text{design}} \prod_{i=1}^{n/m} du_a(i,j)$ and $\int_{2\text{design}} \prod_{j'=1}^{n/m} du_b(i',j')$ for $\ell = 1, 3$,

$$\mathcal{F}_{\text{ALT}}^{(2)}(C^{d,m,n}) = \int_{2\text{design}} \left( \prod_{j=1}^{n/m} du_a(a,j) \right) \left( \prod_{j'=1}^{n/m} du_b(a',j') \right) \sum_{k_{1,1}^{d}} \sum_{k_{2,1}^{d}} \cdots \sum_{k_{\frac{m}{2},1}^{d}} \sum_{k_{1,1}^{d}} \sum_{k_{2,1}^{d}} \cdots \sum_{k_{\frac{m}{2},1}^{d}} \sum_{k_{1,1}^{d}} \sum_{k_{2,1}^{d}} \cdots \sum_{k_{\frac{m}{2},1}^{d}} \frac{1}{(2^m + 1)^n} \frac{1}{\left( (2^m + 1)^n \right)^{\frac{n}{m}}}$$

$$\times \Delta(k_{1,1}^{d},k_{2,1}^{d},k_{1,1}^{d}) (U_a(2,1),U_a(2,1)) \times \Delta(k_{1,1}^{d},k_{2,1}^{d},k_{1,1}^{d}) (U_a(2,2),U_b(2,2))$$

$$\times \Delta(k_{1,1}^{d},k_{2,1}^{d},k_{1,1}^{d},U_a(2,3),U_b(2,3)) \times \cdots \times \Delta(k_{1,1}^{d},k_{2,1}^{d},k_{1,1}^{d},U_a(2,n/m),U_b(2,n/m))$$

$$= \frac{4}{(2^m + 1)^n} \frac{4}{\left( (2^m + 1)^n \right)^{\frac{n}{m}}} \left( \frac{1}{(2^m + 1)^n} \frac{1}{\left( (2^m + 1)^n \right)^{\frac{n}{m}}} \right)$$

$$\int_{2\text{design}} dU_a(2,1) \sqrt{\frac{1}{k_{1,1}^{d}} \Delta(k_{1,1}^{d},k_{2,1}^{d},k_{1,1}^{d}) (U_a(2,1),U_a(2,1))}$$

$$\times \int_{2\text{design}} dU_a(2,2) \sqrt{\frac{1}{k_{1,1}^{d}} \Delta(k_{1,1}^{d},k_{2,1}^{d},k_{1,1}^{d}) (U_a(2,2),U_a(2,2))}$$

$$\times \int_{2\text{design}} dU_a(2,3) \sqrt{\frac{1}{k_{1,1}^{d}} \Delta(k_{1,1}^{d},k_{2,1}^{d},k_{1,1}^{d}) (U_a(2,3),U_a(2,3))}$$

$$\times \cdots \times \int_{2\text{design}} dU_a(2,\frac{n}{m}) \sqrt{\frac{1}{k_{1,1}^{d}} \Delta(k_{1,1}^{d},k_{2,1}^{d},k_{1,1}^{d}) (U_a(2,\frac{n}{m}),U_a(2,\frac{n}{m}))}$$

$$= a(3,m)^T B(3,m) a(3,m).$$ \hspace{1cm} (33)
A.4 Proof of Theorem 4

Similar to Theorem 3, we only show the inequality for \( F^{(2)}(C_{\text{ALT}}^{3,m,n}) \) here. The inequality for \( F^{(2)}(C_{\text{ALT}}^{2,m,n}) \) can be shown in the same manner. Evaluating Eq. (15), we can expand \( a(3, m) \) as

\[
a(3, m) = \frac{1}{2^m} \left( v_0 + \frac{1.2}{2^{m/2}} v_1 \right),
\]

where

\[
v_0 = \begin{cases} 1 & i = 1, \ (k_a, k_b, k_c = 1), \\ 1 & i = 22, \ (k_a, k_b, k_c = 2), \\ 0 & \text{otherwise}, \end{cases} \quad |v_{11}| < 1.
\]

Also, evaluating Eq. (17), we can expand \( B(3, m) \) as

\[
B(3, m) = \frac{1}{2^m} \left( D + \frac{1.3}{2^{m/2}} X \right),
\]

where

\[
D_{ij} = \begin{cases} 1 & i = 1, j = 1 \ (k_a, k_b, k_c, k_d, k_e, k_f = 1), \\ 1 & i = 22, j = 22 \ (k_a, k_b, k_c, k_d, k_e, k_f = 2), \\ 0 & \text{otherwise}, \end{cases} \quad |X_{ij}| < \frac{1}{64}.
\]

With \( \alpha = n/m \), let \( g_{\alpha}(X, D) \) be as the set of matrices expressed by \( \prod_{i=1}^{n} R_i \), where \( R_i = D \) or \( X \) and the number of \( X \)s in \( \{ R_i \} \) is \( k \). For example, \( X D X X \in g_{\alpha}^3(X, D) \) and \( X X D D D \in g_{\alpha}^4(X, D) \). Then, \( F^{(2)}(C_{\text{ALT}}^{3,m,n}) \) is expanded as

\[
F^{(2)}(C_{\text{ALT}}^{3,m,n}) = \frac{1}{2^{2m}} \left( v_0^T + \frac{1.2}{2^{m/2}} v_1^T \right) \frac{1}{2^{2m-2m}} \left( D + \frac{1.3}{2^{m/2}} X \right)^{\alpha - 1} \left( v_0 + \frac{1.2}{2^{m/2}} v_1 \right)
\]

\[
= \frac{1}{2^{2m}} \left( v_0^T D^{\alpha - 1} v_0 + \left( \frac{2.4}{2^{m/2}} \right) v_1^T D^{\alpha - 1} v_0 + \left( \frac{1.2^2}{2^m} \right) v_0^T D^{\alpha - 1} v_1 \right)
\]

\[
+ \frac{1}{2^{2m}} \sum_{k=1}^{n-1} \left( \frac{1.3}{2^{m/2} - 2} \right)^{k-1} \sum_{i=1}^{2^k} \left( v_0^T g_{\alpha}^i v_0 + \left( \frac{2.4}{2^{m/2}} \right) v_1^T g_{\alpha}^i v_0 + \left( \frac{1.2^2}{2^m} \right) v_0^T g_{\alpha}^i v_1 \right).
\]

where \( g_{\alpha}^i \) \( i = 1, 2, \ldots, 2^k \) is an element of \( g_{\alpha}^k(X, D) \). For an arbitrary \( g \in g_{\alpha}^k(X, D) \) with \( k \geq 1 \),

\[
v_1^T g v_s < (1, 1, \cdots, 1) \begin{pmatrix} 1 & 0 \cdots & 0 \\ 0 & 1 \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 \cdots & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{2^m} \frac{1}{2^m} \cdots \frac{1}{2^m} \\ \frac{1}{2^m} \frac{1}{2^m} \cdots \frac{1}{2^m} \\ \vdots & \vdots & \vdots \\ \frac{1}{2^m} \frac{1}{2^m} \cdots \frac{1}{2^m} \end{pmatrix}^{k-1} \begin{pmatrix} 1 \\ 1 \cdots \ 1 \end{pmatrix} = 64
\]

holds. For \( D \) and \( v_0, v_1 \)

\[
v_1^T D \frac{1}{2^m} v_s < \sum_{i=1}^{64} (1 \cdot (\delta_{i1} + \delta_{i22}) \cdot 1) = 2
\]

\[
v_0^T D \frac{1}{2^m} v_0 = 2
\]

holds where \( r, s = 0, 1 \). By using the inequalities (39), (40), and (41), the upper bound for \( F^{(2)}(C_{\text{ALT}}^{3,m,n}) \) is derived as follows:

\[
F^{(2)}(C_{\text{ALT}}^{3,m,n}) < \frac{1}{2^{2m}} \left( 2 + 2 \left( \frac{2.4}{2^{m/2}} \right) + 2 \left( \frac{1.2^2}{2^m} \right) + 64 \sum_{k=1}^{n-1} \alpha - 1 C_k \left( \frac{1.3}{2^{m/2} - 2} \right) \right) \left( \frac{1}{2^{m/2}} \right)^2 \left( 1 + \frac{1.2}{2^{m/2}} \right)^2
\]

\[
= \frac{1}{2^{2m}} \left( 1 + \frac{1.2}{2^m} \right)^2 \left( 1 + 32 \left( \frac{1 + \frac{1.3}{2^{m/2} - 2}}{1} \right) \right)
\]

\[
= \left( 1 + \frac{1}{2^m} \right) \left( 1 + \frac{1.2}{2^m} \right)^2 \left( 1 + 32 \left( \frac{1 + \frac{83.2}{2^{m/2} - 2}}{1} \right) \right) F_{\text{Haar}}^{(2)}(2^n).
\]
A.5 Proof of Corollary 1

As in the above theorems, we only show the inequality for $F^{(2)}(C^3_{\text{ALT}})$ here. When $m = 2a \log_2 n$,

$$
\left(1 + \frac{83.2}{2^{m/2}}\right)^{a-1} < \left(1 + \frac{83.2}{2^{m/2}}\right)^{\frac{n}{2^{m/2}} n} < e^{\frac{83.2 n}{2^{m/2}}} = e^{\frac{41.6}{an^{a-1} \log_2 n}}.
$$

(43)

If $41.6/(an^{a-1} \log_2 n) < 1$,

$$
e^{\frac{41.6}{an^{a-1} \log_2 n}} < 1 + (e - 1) \frac{41.6}{an^{a-1} \log_2 n}.
$$

(44)

Substituting Eqs. (43), (44), and $m = 2a \log_2 n$ into (42), we get

$$
F^{(2)}(C^3_{\text{ALT}}) < \left(1 + \frac{1}{2^n}\right) \left(1 + \frac{1.2}{n^2} n\right)^{2} \left(1 + \frac{2288}{an^{a-1} \log_2 n}\right) F^{(2)}(C^3_{\text{Haar}}(2^n)).
$$

(45)

References

[1] J. Preskill. Quantum Computing in the NISQ era and beyond Quantum 2, 79 (2018); arXiv:1801.00862. DOI: 10.22331/q-2018-08-06-79.

[2] A. Peruzzo, Jarrod R. McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Aln Aspuru-Guzik and Jeremy L. O’Brien. A variational eigenvalue solver on a quantum processor Nature Communications, 5:4213 (2014).

[3] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow and J. M. Gambetta. Hardware-efficient Variational Quantum Eigensolver for Small Molecules and Quantum Magnets Nature 549, 242 (2017);

[4] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush and H. Neven. Barren plateaus in quantum neural network training landscapes Nature Communications, Volume 9, Article Number: 4812 (2018).

[5] E. Grant, L. Wossnig, M. Ostaszewski and M. Benedetti. An initialization strategy for addressing barren plateaus in parametrized quantum circuits Quantum 3, 214 (2019).

[6] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum natural gradient, arXiv:1909.02108.

[7] M. Cerezo, A. Sone, T. Volkoff, L. Cincio and P. J. Coles. Cost-Function-Dependent Barren Plateaus in Shallow Quantum Neural Networks arXiv:2001.00550.

[8] S. Sim, P. D. Johnson and A. Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms Adv. Quantum Technol. 2 (2019) 1900070.

[9] K. Zyczkowski and HJ Sommers. Average fidelity between random quantum states Phys. Rev. A 71, 032313 (2005).

[10] D. A. Roberts and B. Yoshida. Chaos and complexity by design JHEP 04 (2017) 121.

[11] J. M. Renes, R. Blume-Kohout, A. J. Scott and C. M. Caves. Symmetric Informationally Complete Quantum Measurements J. Math. Phys. 45, 2171 (2004).

[12] A. Klappenecker and M. Roetteler. Mutually Unbiased Bases are Complex Projective 2-Designs Proceedings 2005 IEEE International Symposium on Information Theory (ISIT 2005), Adelaide, Australia, pp. 1740-1744, 2005.

[13] I. Bengtsson and K. Zyczkowski Geometry of Quantum States: An Introduction to Quantum Entanglement (Cambridge University Press, 2017) pp. 1619.

[14] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization In Proceedings of 3rd International Conference on Learning Representations, 2015.