Scientific Reports

https://doi.org/10.1038/s41598-025-31798-1

Article in Press

Light cone cancellation for variational quantum eigensolver in solving noisy Max-Cut

Received: 23 September 2025

Accepted: 5 December 2025

Published online: 23 February 2026

Cite this article as: Lee X., Yan X., Xie N. *et al.* Light cone cancellation for variational quantum eigensolver in solving noisy Max-Cut. *Sci Rep* (2025). https://doi.org/10.1038/s41598-025-31798-1

Xinwei Lee, Xinjian Yan, Ningyi Xie, Yoshiyuki Saito, Leo Kurosawa, Nobuyoshi Asai, Dongsheng Cai & Hoong Chuin Lau

Light Cone Cancellation for Variational Quantum Eigensolver in Solving Noisy Max-Cut

Xinwei Lee[1,*,+], Xinjian Yan[2,*,+], Ningyi Xie[2], Yoshiyuki Saito[3], Leo Kurosawa[3], Nobuyoshi Asai[4], Dongsheng Cai[5], and Hoong Chuin LAU[1]

[1]School of Computing and Information Systems, Singapore Management University
[2]Graduate School of Science and Technology, University of Tsukuba
[3]Graduate School of Computer Science and Engineering, University of Aizu
[4]School of Computer Science and Engineering, University of Aizu
[5]Faculty of Engineering, Information and Systems, University of Tsukuba
[*]xwlee@smu.edu.sg
[*]yanxinjian@cavelab.cs.tsukuba.ac.jp
[+]These authors contributed equally to this work.

ABSTRACT

Variational Quantum Eigensolver (VQE) is a quantum-classical hybrid algorithm used to estimate the ground energy of a given Hamiltonian. It consists of a parameterized quantum circuit, which the parameters are optimized using a classical optimizer. With the increasing need in solving large-scale problems in real-world applications, solving those large problems with fewer qubits and fewer gates becomes essential, so that we reduce the simulation difficulty and mitigate the effect of noise in real quantum hardware. In this study, we applied the Light Cone Cancellation (LCC) method to reduce the number of qubits and gates required in a two-local ansatz. LCC removes redundant gates that are not required in the calculation of the expectation value for a local observable. This leads to two consequences: 1) the quantum circuit used to create the trial wavefunction of VQE can be broken down into multiple quantum subcircuits with fewer qubits, enabling large-scale problems to be solved without actually simulating the entire circuit; and 2) reduced number of quantum gates in the circuit leads to the noise mitigation in quantum hardware. The main purpose of this work is to demonstrate the effectiveness of this method (called the LCC-VQE) in mitigating the device noise when solving the Max-Cut problem up to 100 qubits, using simulations on small (7-qubit and 27-qubit) fake noisy backends. Employing a single-layer two-local ansatz circuit architecture, the results show that LCC-VQE yields higher approximation ratios than those cases without LCC, implying that the effect of noise is mitigated when LCC is applied. An analysis of more than one layer of two-local ansatz is also performed, but empirical results show that the single-layer ansatz still performs the best among them. We also compare LCC-VQE under noiseless conditions with the Goemans-Williamson algorithm.

1 Introduction

Many combinatorial optimization problems (COP) are considered to be difficult to address using classical computational approaches. COPs aim to find the optimal combination of variables that minimizes (or maximizes) a given objective function, while simultaneously satisfying a set of constraints. Recent years, people focus on using quantum-classical hybrid methods, known as the variational quantum algorithms (VQA)[1] to heuristically solve COPs. The quantum approximate optimization algorithm (QAOA)[2] is one of the VQAs that is intensively explored due to its predictable patterns in the variational parameters[3–5], and also its relation with quantum annealing[6,7]. Another VQA, the variational quantum eigensolver (VQE)[8], is also capable of solving COPs, although it is better known for its application to quantum chemistry. Unlike QAOA which has a problem-dependent ansatz, the structure of the ansatz in VQE is static and does not depend on the problem solved. Moreover, the VQE ansatz offers a greater degree of freedom in the sense that it has greater expressibility[9] and more number of variational parameters compared to the QAOA ansatz. In our previous work[10], we have shown that VQE generally achieves higher approximation ratio than QAOA and the Multi-angle QAOA[11] in solving the Max-Cut problem using noiseless simulations. In terms of quantum resources, due to the expressibility of VQE, it requires less layers to achieve the same approximation ratio as QAOA. This results in less quantum gates used in the variational ansatz, which leads to less gate noise in real quantum hardware. In terms of optimization cost, it is also shown in[10] that optimizing VQE requires less function evaluations than optimizing QAOA.

The VQAs have shown the potential quantum advantage on Noisy Intermediate Scale Quantum (NISQ)[12] devices. However, an increase in the number of qubits often leads to higher error rates when building actual quantum hardware. Although

recent advancements have prominently featured quantum error correction algorithms, those involve intricate designs that must effectively address the inherent noise and decoherence in quantum environments. It appears that reducing the number of qubits and gates is a more feasible and efficient approach while maintaining the accuracy of algorithms.

In this paper, we apply a method known as Light Cone Cancellation (LCC)[13,14] to solve the Max-Cut problem using VQE. When computing the expectation function of variational circuits, there are many redundant operators that need not be included in the computations. LCC exploits the preliminary knowledge of which operators are redundant, so that we do not include them in the calculation at the first place. LCC is widely applied for QAOA[15], and also inspired applications in tensor network[16–21] and quantum machine learning[22].

The primary contribution of our work is that we demonstrate the effectiveness of LCC in reducing the number of qubits required for the simulation of quantum circuits, and also noise mitigation caused by the reduction in the number of gates in the circuit simulation. We simulate using Qiskit fake noisy backends with 7 qubits and 27 qubits. The demonstration results show that, compared to the original VQE, the implementation of LCC achieves better performance. Additionally, to further quantify the performance of LCC-VQE under a noiseless condition, we benchmark it against the Goemans–Williamson (GW) algorithm[23] on 100-vertex graph instances. The simulation results indicate that LCC-VQE achieves better performance compared to the GW algorithm on denser graphs.

The rest of the paper is structured as follows. Section II provides background information and details the construction of the LCC architecture. The detailed results of comparative simulations under both noisy and noiseless conditions are articulated in Section III. Section IV contains the concluding remarks of this study.

2 Background

2.1 Max-Cut

Max-Cut is a fundamental and widely studied NP-hard combinatorial optimization problem in the field of graph theory[24]. The primary objective of Max-Cut is to partition the nodes of an undirected graph into two disjoint subsets such that the number of edges connecting the two subsets is maximized. This problem is relevant in various fields, including network design, VLSI layout, community detection, and social network analysis[25].

Consider an n-node unweighted, undirected graph $G = (V, E)$, where V represents the set of the nodes and E represents the set of the edges of graph G. A cut is defined as a partition of the original set V into two subsets. The cost function $C(\mathbf{x})$ to be maximized is the sum of the edges connecting points in the two different subsets, which can be expressed as

$$C(\mathbf{x}) = \sum_{(i,j)\in E} (x_i \oplus x_j), \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $x_i \in \{0, 1\}$ represent the binary variable of node i, and $n = |V|$ is the number of nodes in the graph. The symbol $\oplus$ denotes the XOR (exclusive-OR) operation. We want to find the combinations of $\mathbf{x}$ such that the cost function is maximized, i.e. the number of edges cut is maximum. A brute-force approach on a classical computer would require $\mathcal{O}(2^n)$ time to solve this problem.

In the quantum realm, the cost function in Eq. (1) can be formulated as the cost Hamiltonian H_C, whose expectation value is to be maximized:

$$H_C = \frac{1}{2} \sum_{(i,j)\in E} (I - Z_i Z_j), \tag{2}$$

where I is the identity matrix of size $2^n \times 2^n$, and Z_i represents Pauli-Z observable on qubit i. One node in the graph is mapped to one qubit in the quantum circuit, so it requires n qubits to encode the solution of Max-Cut.

2.2 Variational Quantum Eigensolver

The variational quantum eigensolver (VQE) is initially developed for calculating the minimum energy states of molecules. When re-formulated to address the Max-Cut problem, the expectation value of the cost Hamiltonian H_C over a trial state $|\psi(\boldsymbol{\theta})\rangle$ is defined as

$$\mathscr{E}(\boldsymbol{\theta}) = \langle \psi(\boldsymbol{\theta}) | H_C | \psi(\boldsymbol{\theta}) \rangle . \tag{3}$$

The objective is to maximize $\mathscr{E}(\boldsymbol{\theta})$, which is equivalent to minimizing $-\mathscr{E}(\boldsymbol{\theta})$, using a classical optimizer. $\boldsymbol{\theta}$ is the collection of variational parameters for the VQE ansatz circuit. In this paper, we employ a two-local circuit with $R_y(\theta)$ single-qubit rotation gates and CZ (controlled-Z) circular entanglement, with only a single layer. Fig. 1 shows an example of a 4-qubit ansatz circuit. The circular entanglement has CZ gates between adjacent qubits, and also between the first and the last

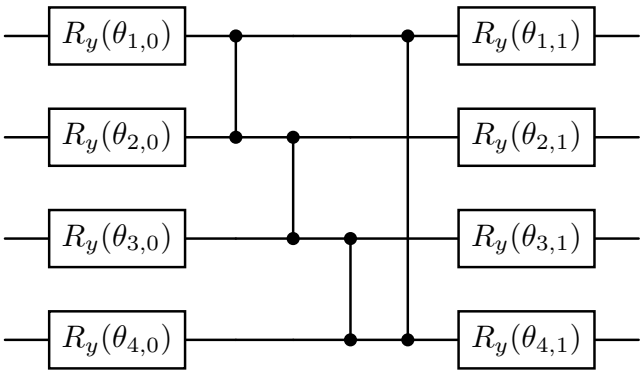

Figure 1. Two-local ansatz used in the simulations in our work. The architecture of a single layer of R_y gates with circular CZ entanglement is used. The figure shows a 4-qubit example. The parameters $\theta_{k,m}$ are based on the notation specified in Eq. (4).

qubit. Instead of using a flattened array $\boldsymbol{\theta}$, we use a matrix Θ to represent the variational parameters, aligning them with their geometrical positions in the circuit:

$$\Theta = \begin{bmatrix} \theta_{1,0} & \theta_{1,1} \\ \theta_{2,0} & \theta_{2,1} \\ \vdots & \vdots \\ \theta_{n,0} & \theta_{n,1} \end{bmatrix}. \tag{4}$$

It is important to note that Θ is simply the re-shaped version of $\boldsymbol{\theta}$, and the notation $\theta_{k,m}$ denotes the parameters on the k-th qubit in the original circuit. $\theta_{k,0}$ corresponds to the rotation angle of the initial R_y gate, whereas $\theta_{k,1}$ corresponds to the rotation angle of the R_y gate applied after the circular CZ entanglement. Also, due to the periodicity of the R_y rotation gate, the range of the values for $\theta_{k,m}$ are restricted to $[0, 2\pi)$.

Numerous metrics are available for assessing the performance of VQE. Within the scope of unconstrained COP, our focus is on studying the approximation ratio (AR). This metric compares the expected solution obtained through VQE with the optimal solution, essentially measuring how close VQE comes to the best possible outcome. It is defined as

$$\mathrm{AR} = \frac{\mathscr{E}(\boldsymbol{\theta}^*)}{\mathrm{MaxCut}(G)}, \tag{5}$$

where $\boldsymbol{\theta}^*$ is the quasi-optimal parameters returned by the optimizer, $\mathscr{E}(\boldsymbol{\theta}^*)$ is its corresponding expectation, and $\mathrm{MaxCut}(G)$ is the exact solution of graph G. The closer the value of AR is to 1, the closer it is to the true solution of Max-Cut.

The performance of VQE and QAOA in addressing Max-Cut problems under noiseless conditions is evaluated in previous research[10]. When both algorithms are initialized with the same number of parameters, our findings indicate that VQE outperforms QAOA under random initialization. Furthermore, the comparison with Multi-angle quantum approximate optimization algorithm (ma-QAOA) reveals that VQE also achieves superior performance on different undirected graphs.

2.3 Light Cone Cancellation

Light cone cancellation (LCC) is a method that utilizes the intrinsic property of the expectation function, so that the redundant unitaries in the expectation function are not included in its computation in the first place[26]. The LCC property was originally used in QAOA to reduce the problem graphs to their constituent subgraphs, hence simplifying the problem to be solved[27–29].

The following formally describes LCC for the circular-entangled two-local circuit using mathematical derivation. We start by substituting Eq. (2) into the expectation function Eq. (3), the Max-Cut expectation can be rewritten as

$$\mathscr{E}(\boldsymbol{\theta}) = \frac{|E|}{2} - \frac{1}{2} \sum_{(i,j)\in E} \langle \psi(\boldsymbol{\theta}) | Z_i Z_j | \psi(\boldsymbol{\theta}) \rangle . \tag{6}$$

The trial wavefunction (or ansatz) on both sides, $\langle \psi(\boldsymbol{\theta}) |$ and $| \psi(\boldsymbol{\theta}) \rangle$, can be partially cancelled out. This is because some of the unitary gates used to prepare $| \psi(\boldsymbol{\theta}) \rangle$ commute through the central local observables $Z_i Z_j$. Depending on i and j, the state $| \psi(\boldsymbol{\theta}) \rangle$ prepared by the original full circuit of L alternating layers becomes $\left| \psi_{i,j}(\boldsymbol{\theta}) \right\rangle$, which can be prepared by subcircuits

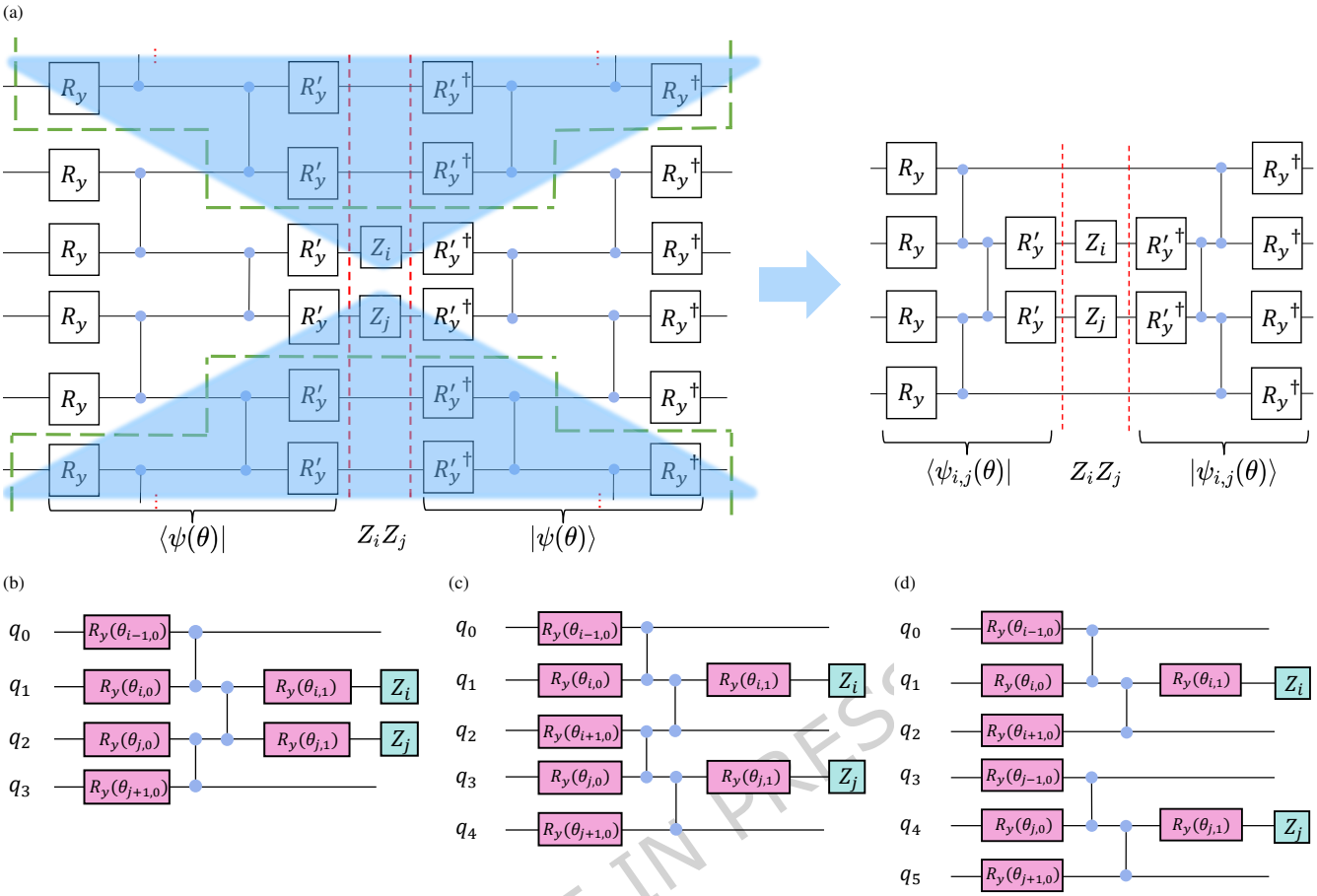

Figure 2. (a) The light cone cancellation (LCC) in a single layer two-local ansatz. The expectation function $\langle\psi(\boldsymbol{\theta})|Z_iZ_j|\psi(\boldsymbol{\theta})\rangle$ is visualized as a quantum circuit on the left figure. R_y's are the single-qubit gates and C_z's are the two-qubit gates. The gates on the left of the red dashed line show $\langle\psi(\boldsymbol{\theta})|$ and the gates on the right show $|\psi(\boldsymbol{\theta})\rangle$. The blue shaded regions show the redundant gates that can be cancelled during the calculation of the expectation. The figure on the right shows the resulting circuit after cancellation. (b), (c) and (d) show the possible resulting circuit for the Max-cut Hamiltonian, depending on the indices i and j (positions of the Pauli-Z operators). (b) When Z_i and Z_j are adjacent; (c) when Z_i and Z_j are one qubit apart; and (d) when Z_i and Z_j are two or more qubits apart.

using less qubits:

$$\left|\psi_{i,j}(\boldsymbol{\theta})\right\rangle = U_1U_2\cdots U_L \prod_{s\in\mathscr{S}^{(L)}_{i,j}} R_y(\theta_{s,0})\,|0\rangle^{\otimes n_q}, \tag{7}$$

The operator U_m is the unitary in the $(L-m)$-th layer (larger m means further from the center of the light cone) of the subcircuit and is given by:

$$U_m = \prod_{s\in\mathscr{S}^{(m)}_{i,j}} R_y(\theta_{s,m}) \prod_{(s,s')\in\mathscr{T}^{(m)}_{i,j}} CZ_{s,s'}, \tag{8}$$

where $R_y(\theta_{s,m})$ denotes a single-qubit R_y gate acting on qubit s at layer m, and $CZ_{s,s'}$ denotes a CZ gate acting on qubit s and s'. For all $m\in\{1,2,...,L\}$, we respectively define the sets $\mathscr{S}^{(m)}_i$ and $\mathscr{T}^{(m)}_i$ for the R_y and CZ gates:

$$\mathscr{S}^{(m)}_i := \{s\in\{1,2,\ldots,n\} \mid (i-m\le s\le i+m) \bmod n\} \tag{9}$$

$$\mathscr{T}^{(m)}_i := \{(s,s') | s\in\mathscr{S}^{(m)}_i, s' = (s+1) \bmod n\}. \tag{10}$$

These sets contain the qubit indices that span l neighboring qubits of qubit i (because of the observable Z_i on qubit i). Since

there are two observables, Z_i and Z_j, the qubits spanned by these observables are the union of the two sets, hence we define:

$$\mathscr{S}_{i,j}^{(m)} := \mathscr{S}_i^{(m)} \cup \mathscr{S}_j^{(m)} \tag{11}$$

$$\mathscr{T}_{i,j}^{(m)} := \mathscr{T}_i^{(m)} \cup \mathscr{T}_j^{(m)} \tag{12}$$

to represent the union sets in the subscript of the operators in Eq. (8). The total number of qubits required for each subcircuit is

$$n_q' = \left|\mathscr{S}_{i,j}^{(L)}\right|, \tag{13}$$

where $|\cdot|$ is the set cardinality. Thus, the number of qubits required for each circuit is $\mathscr{O}(L)$. It is important to know that $n_q' \leq n$, in which LCC leads to a reduction in the number of qubits.

Consequently, the states $|\psi_{i,j}(\boldsymbol{\theta})\rangle$ prepared by the subcircuit is used to calculate the expectation $\mathscr{E}(\boldsymbol{\theta})$ instead of using the full circuit state $|\psi(\boldsymbol{\theta})\rangle$:

$$\mathscr{E}_{\mathrm{LCC}}(\boldsymbol{\theta}) = \frac{|E|}{2} - \frac{1}{2}\sum_{(i,j)\in E} \langle \psi_{i,j}(\boldsymbol{\theta})| Z_i Z_j |\psi_{i,j}(\boldsymbol{\theta})\rangle . \tag{14}$$

Since LCC is just cancelling the redundant operators, the expectation computed using LCC is technically the same as the original expectation:

$$\mathscr{E}(\boldsymbol{\theta}) \equiv \mathscr{E}_{\mathrm{LCC}}(\boldsymbol{\theta}). \tag{15}$$

Fig. 2(a) shows an example of the LCC of a two-local circuit used to prepare the trial wavefunction for VQE. The figure visualizes the expectation function Eq. (3) as a quantum circuit. The circuit on the left of the red dashed line shows the term $\langle\psi(\boldsymbol{\theta})|$, and the circuit on the right of the dashed line shows $|\psi(\boldsymbol{\theta})\rangle$, with the central observable Z_iZ_j (between the red dashed lines) acting on qubit i and j. Since $|\psi(\boldsymbol{\theta})\rangle$ is just the conjugate transpose of $\langle\psi(\boldsymbol{\theta})|$, they are the counterpart of each other in the circuit. The blue shaded regions indicate the gates that are not related to qubits i and j can commute through the center are cancelled. The result of this cancellation is shown on the right side of Fig. 2(a), which has reduced the number of qubits and gates.

The Max-Cut Hamiltonian in Eq. (2) has a two-local Z observable on every term. After LCC on the two-local ansatz that we considered (single layer R_y and circular CZ entanglement), we can get 3 different types of subcircuits as shown in Fig. 2(b), (c) and (d), which requires 4-, 5-, and 6-qubit quantum circuits, respectively. Circular entanglement means the adjacent qubits, as well as the first and the last qubits, are entangled. Note that one subcircuit corresponds to the expectation of a local term in the Hamiltonian, so the simulation of the subcircuits can be done separately. Also, the subcircuit in Fig. 2(d) can be further divided into two separate circuits as the first 3 qubits and the last 3 qubits are not entangled. Thus, we only require a maximum of 5 qubits to simulate the expectation of the entire Max-Cut Hamiltonian, regardless of the problem size. In fact, the maximum number of qubits,

$$n_q = 2k+1, \tag{16}$$

are required to simulate the expectation of a Hamiltonian with k-local observables, for this kind of ansatz (one layer of single-qubit gates and circular entanglement). The architecture of the entanglement is crucial in deciding how many qubits we can reduce. For linear entanglement (only adjacent qubits entangled, first and last not entangled), there would be another case of the subcircuits where the observable is located on the first qubit or last qubit. For full entanglement (all qubits entangled), LCC will not be possible.

Another advantage of LCC is that the number of gates is reduced, which in turn reduces the effect of gate noise in quantum devices. However, it is also worth noting that even though the numbers of qubits and gates are reduced, the number of parameters remains unchanged after LCC. This is because the subcircuits after LCC will have different parameters corresponding to the indices i and j, depending on where the observables Z_iZ_j are. Therefore, the difficulty in parameter optimization remains the same before and after LCC.

LCC is also applicable for other circuits like QAOA and ma-QAOA. However, QAOA or ma-QAOA usually needs more layers (larger circuit depths) to achieve higher ARs. Meanwhile, VQE with even one layer is enough to reach most of the states and hence easier to reach higher ARs than QAOA. This is due to the difference in expressibility between the VQE ansatz and the QAOA ansatz[9]. Since VQE yields higher ARs with fewer layers than QAOA, LCC-VQE can be done with fewer qubits, compared to LCC-QAOA or LCC-ma-QAOA.

Fig. 3 shows the overall workflow of the LCC in two-local ansatz. The ansatz $U(\boldsymbol{\theta})$ prepares a parameterized quantum state $|\psi(\boldsymbol{\theta})\rangle$, which is then used to evaluate the expectation function $\langle\psi(\boldsymbol{\theta})|H_C|\psi(\boldsymbol{\theta})\rangle$ for the Max-Cut Hamiltonian H_C. By

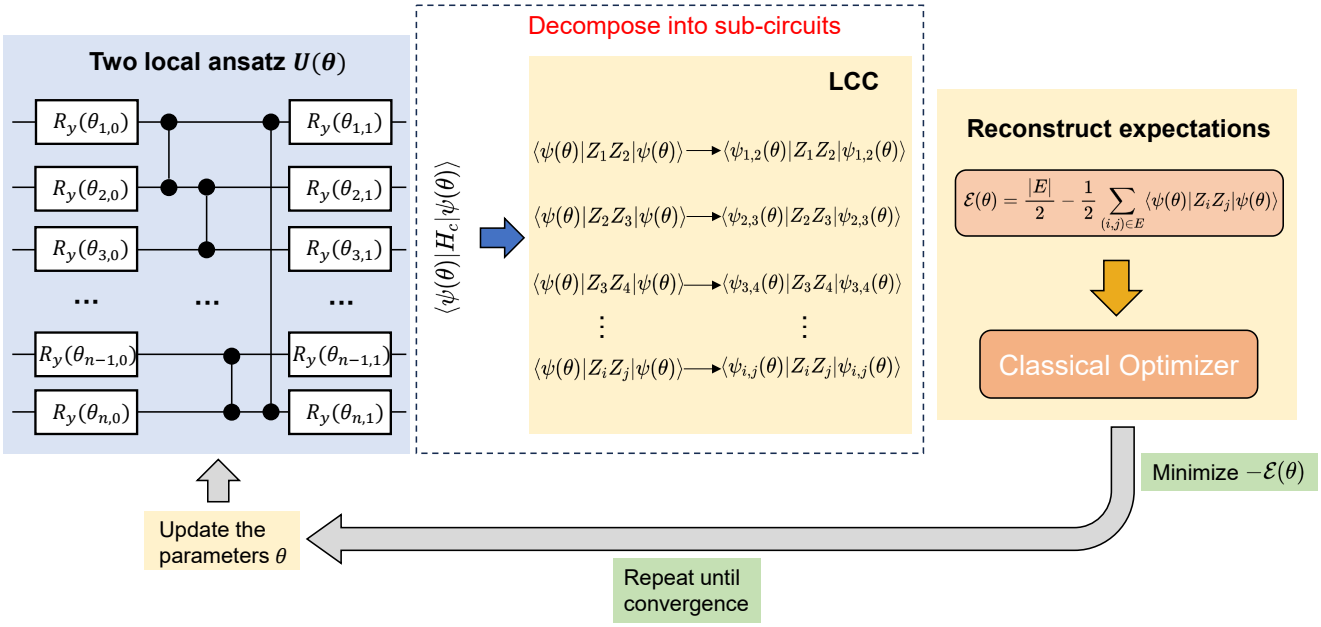

Figure 3. Overall workflow of the LCC in two-local ansatz. The expectation function $\langle\psi(\boldsymbol{\theta})|H_c|\psi(\boldsymbol{\theta})\rangle$ of the original circuit is decomposed into separate expectation values of subcircuits using Light Cone Cancellation. These sub-expectations are used to reconstruct the total expectation function $\mathscr{E}(\boldsymbol{\theta})$ of the Max-Cut instances, which is then iteratively optimized within a hybrid quantum-classical framework.

applying the Light Cone Cancellation (LCC), the original expectation value is decomposed into a set of smaller subcircuit expectations, $\langle\psi_{i,j}(\boldsymbol{\theta})|Z_iZ_j|\psi_{i,j}(\boldsymbol{\theta})\rangle$, each involving only a subset of qubits associated with local interactions in the Max-Cut instances. These local expectations are subsequently combined to reconstruct the total expectation function $\mathscr{E}(\boldsymbol{\theta})$, which is optimized using a classical optimizer. The parameters $\boldsymbol{\theta}$ are then iteratively updated in a hybrid quantum-classical loop until convergence.

2.4 Number of layers in the ansatz

In this work, we only consider a single layer of the VQE ansatz as it is sufficient to address the Max-Cut problem. To justify this, we conducted a simulation to investigate how the quality of the solution varies with the number of layers in the two-local ansatz. Fig. 4 shows that as the number of layers increases, the chance of obtaining a high-quality solution (AR ≥ 0.99) decreases. The chance of obtaining a high-quality solution is quantified by the percentage, i.e., the total number of times AR ≥ 0.99 obtained, divided by the total number of trials. For $n = 10, 11, 12$, each point shows the percentage calculated from 8 graph instances, each with 24 different random initial parameters, with a grand total of $8 \times 24 = 192$ trials. For $n = 13, 14, 15$, each point shows the percentage for 4 graph instances, each with 24 trials, with a total of 96 trials. The decline in the percentage can be explained by the increasing difficulty of optimization as the number of layers (number of parameters) increases, possibly the increased number of local minima, causing overparameterization where the AR could not increase further despite increasing the number of parameters. Hence, we conclude that a single-layer ansatz is sufficient for the problem instances considered.

Let us consider what happens to LCC when we have more than one layer of the ansatz. As the number of layers increases, the number of gates that can be cancelled decreases, resulting in a larger subcircuit after LCC. This is because entangling gates (CZ gate in our case) farther from the center cannot commute through the layers nearer to the center (where the observables are) to get cancelled out on the other side, causing them to remain in the circuit after LCC (refer to Fig. 2, where the area outside the light cone gets larger if the circuit has more layers). The number of qubits remaining after LCC depends on the entanglement structure of the ansatz. The entanglement map of a quantum circuit can be viewed as a graph with the qubits as the nodes and the entanglement as the edges, e.g., if there is an entangling gate between qubit i and qubit j, then there is an edge between node i and node j. The resulting entanglement map, after LCC, can then be viewed as a subgraph spanned by the observables Z_i and Z_j, from the distance L nodes away from node i and node j, where L is the number of layers in the original ansatz. This is analogous to the idea where QAOA is said to search deeper subgraphs as its circuit depth p increases[2,27,30]. Fig. 5 shows the visualization of the entanglement map after LCC, considering the largest subgraph that can be spanned by the two observables. For linear and circular entanglements, the subgraphs are represented by line graphs. To calculate the maximum number of qubits required after LCC, we consider the case where the size of the subgraph is at its maximum. This happens when the observable nodes stretch out for a distance to each side and overlap at their ends, similar to the subcircuit in Fig. 2(c) in the

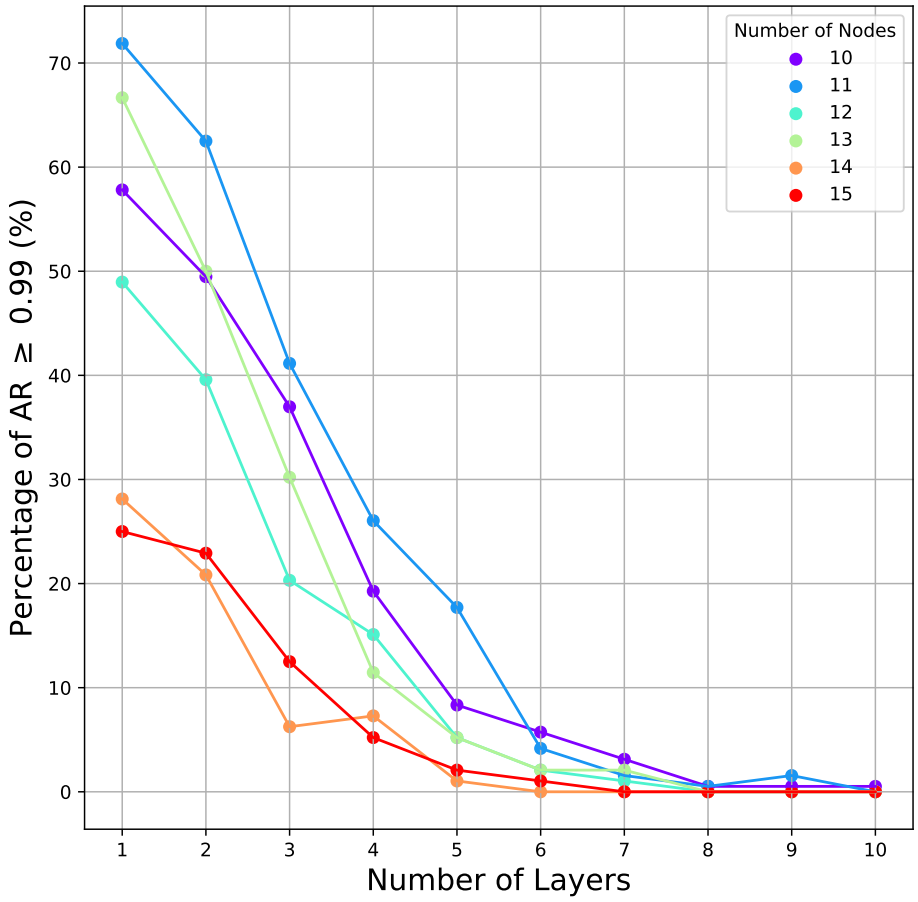

Figure 4. Percentage of AR ≥ 0.99 in different number of layers of the ansatz. For $n = 10, 11, 12$, each point shows the percentage calculated from 8 graph instances, each with 24 different random initial parameters, with a total of 192 trials. For $n = 13, 14, 15$, each point shows the percentage for 4 graph instances, each with 24 trials, with a total of 96 trials. Each trial represents a set of random initial parameters, converged to the given AR after an optimization run.

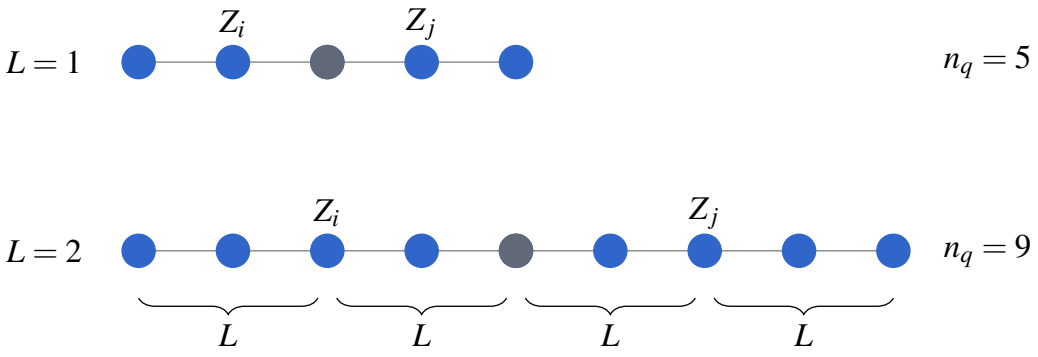

Figure 5. The entanglement map after LCC when the number of ansatz layers L increases. Each node represents a qubit and each edge represents an entanglement. The figure shows the maximum number of qubits required for the linear or circular entanglement after LCC. The number of qubits required for LCC also depends on the number of qubits of the original ansatz, and also the distance between the observables. The maximum number of qubits is achieved when the observables are exactly $2L$ qubits apart of each other. To achieve maximum number of qubits, each of the observables stretches out a distance of L qubits (blue nodes) for two sides, and they overlap at the grey node, forming an inseparable entanglement.

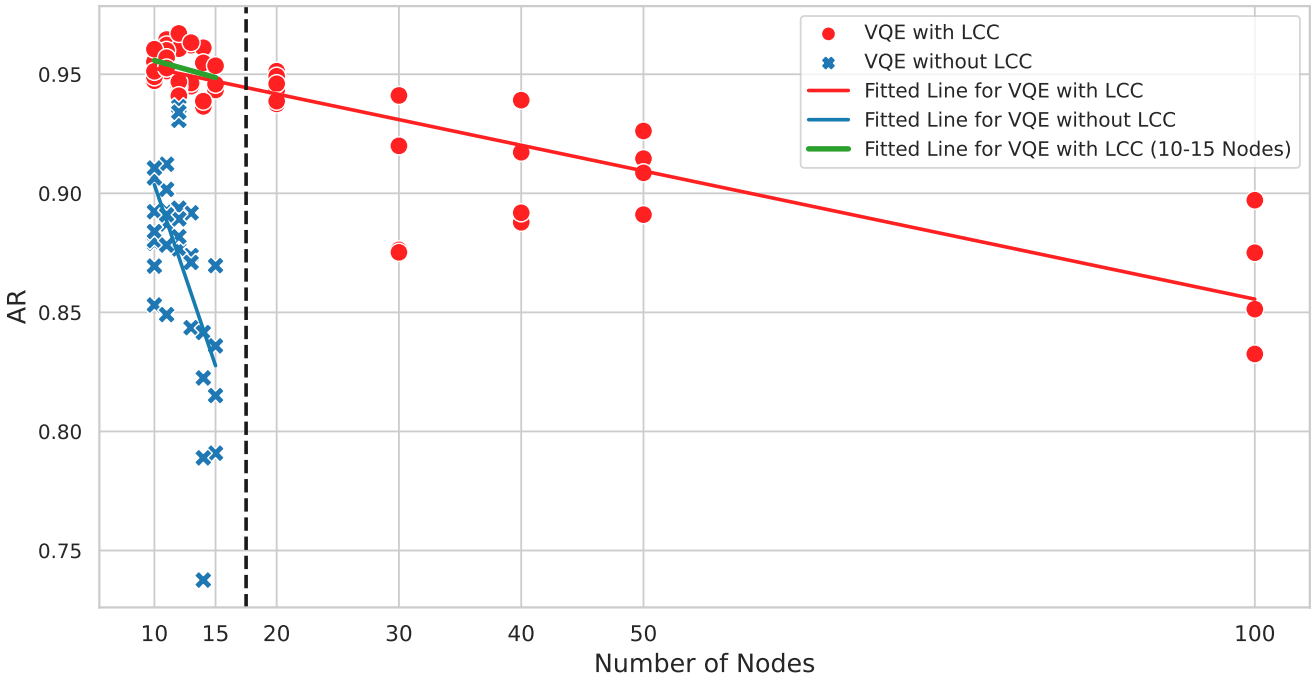

Figure 6. Comparison of the approximation ratio (AR) for the VQE solved with a 7-qubit fake backend `FakeCasablanca` (with LCC) and a 27-qubit fake backend `FakeParis` (without LCC). Each point shows the best AR (highest) chosen out of 24 trials. The lines are linear fits of their respective data. Meanwhile, the gradients of the red line, blue line, and green line are -0.0011, -0.0152 and -0.0018, respectively.

case of $L = 1$. The grey nodes in Fig. 5 show the overlapping nodes for each observable. Thus, it is not difficult to establish a relation between the maximum number of qubits required n_q with the number of layers L:

$$n_q = 4L + 1, \tag{17}$$

for 2-local observables like the Max-Cut Hamiltonian. For k-local observables, the maximum number of qubits is

$$n_q = 2kL + 1. \tag{18}$$

Note that n_q is the maximal case where $n \geq n_q$ and the observables are exactly $2L$ qubits apart from each other. The maximum number of qubits required will still be bounded by the original graph size n, and also when the observables are nearer or farther from each other. Eq. (16) and (17) are the special cases for Eq. (18) when $L = 1$ and $k = 2$ respectively, in the case where the ansatz only has one layer, or the observables are 2-local.

3 Results

3.1 LCC under noisy conditions

We solve the Max-Cut problem using VQE with a two-local ansatz (single layer R_y and CZ entanglement). We employ the COBYLA optimizer[31], along with the AerSimulator provided by Qiskit[32], for all the simulations. The demonstrations are designed to compare the performance of the VQE with LCC and that of VQE without LCC. We measure the performance using the approximation ratio (AR), which indicates how close the result given by VQE is to the optimal solution. To make sure that the optimizer does not converge to a good minimum by chance, we perform 24 trials with random initial parameters for every instance, i.e., random initialization. The demonstrations are performed under noisy conditions so that we can observe the effect of the reduction in the number of qubits and the number of gates on the amount of noise in the circuit. We use two different fake noisy devices provided by Qiskit[33]: `FakeCasablanca` (7 qubits) and `FakeParis` (27 qubits). These fake devices simulate the same noise settings in their respective real quantum devices. The specifications of the two noisy devices are stated in the Supplementary Materials. We did two different comparisons to show the noise mitigation of LCC resulted by two different factors: 1) we compare the results for the 7-qubit vs. the 27-qubit devices to show that LCC allows us to run the subcircuits on a smaller device, which results in noise mitigation; 2) we then compare the execution of LCC on the same 27-qubit device to show the noise mitigation due to the reduced number of gates.

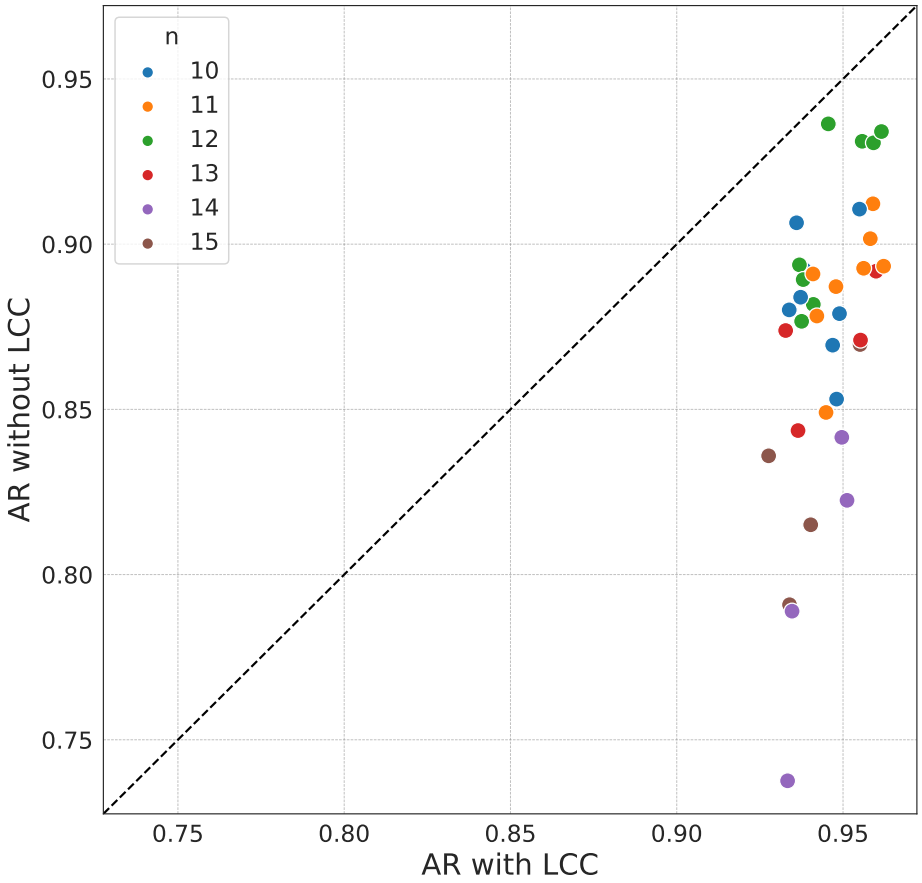

Figure 7. Comparison of the AR for the VQE with LCC and without LCC, using the same `FakeParis` backend (27 qubits). The diagonal dashed line shows where the AR of both methods are equal. All the points are in the lower triangle and represent higher AR with LCC.

Figure 6 shows the comparison between Max-Cut instances solved with LCC on the 7-qubit fake backend, and those solved without LCC on the 27-qubit fake backend. As only 5 qubits are required to simulate the subcircuits for LCC, the VQE simulation can be run on a device with 7 qubits. On the other hand, simulation with full number of qubits is required for those without LCC. The main purpose of this figure is to show the possibility of running LCC in a smaller device with less noise, while Figure 7 shows the comparison of the effect of noise on the same device, attributed to the reduced number of gates after LCC. In both settings, we solve the Max-Cut for 36 non-isomorphic instances, ranging from number of vertices $n = 10$ to $n = 15$. Additonally, 24 non-isomorphic instances are solved on the 7-qubit fake backend (with LCC) for $n = 20, 30, 40, 50$ and 100. The dataset for the demonstrations is shown in the Supplementary Materials. Each point in the plot represents the best AR out of 24 trials for a single problem instance. The red points plot the ARs for the instances solved with LCC on the 7-qubit fake backend; the blue points plot the ARs for the instances solved without LCC on the 27-qubit fake backend. The red line is a linear fit through the red points (with LCC) for $n = 10$ to $n = 100$. The blue line linearly fits through the blue points (without LCC) for $n = 10$ to $n = 15$. The green line is a fit for the red points (with LCC) from $n = 10$ to $n = 15$.

There are a few observations worth noting. LCC enabled the simulation of large problems up until $n = 100$, only with quantum circuits with at most 5 qubits. From the red and blue fitted lines, we can see that the ARs for problems with LCC are generally higher than those without LCC. Although with different environments (7-qubit and 27-qubit fake backends), the error rate is generally lower on a smaller device, so the AR is not so much deteriorated on a 7-qubit device. Moreover, with less number of gates, the effect of noise on the AR is also reduced. It can also be observed that the AR decreases as the problem size (number of the graph vertices) increases. Another interesting point to note is that the green line has similar slope as the red line, which means the decreasing trend is similar for microscopic ($n = 10$ to $n = 15$) and macroscopic ($n = 10$ to $n = 100$) number of nodes. Also, we can observe from the blue line that if without LCC, the AR decreases faster due to more noises in the circuits. This implies the possibility that in the case without LCC, the AR would decrease faster than the case with LCC, if the blue line is extrapolated to larger problem size.

Figure 7 shows the comparison for $n = 10$ to $n = 15$ problems solved with and without LCC, using the same 27-qubit

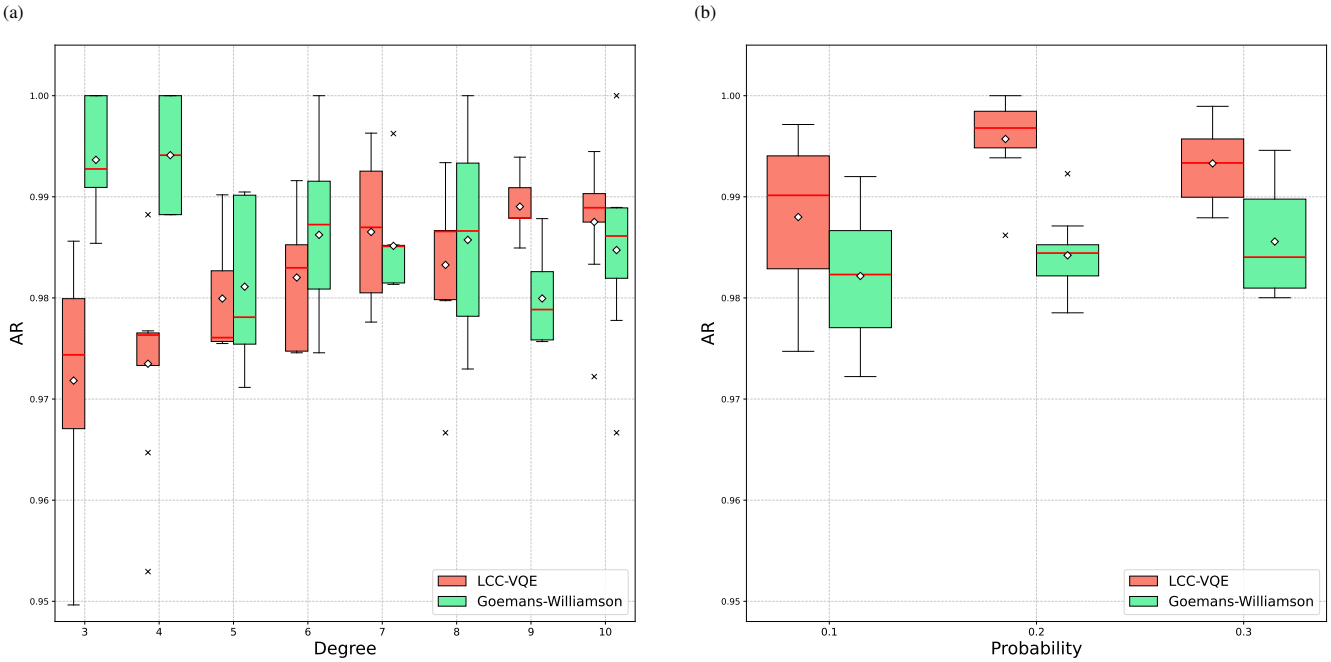

Figure 8. (a) Comparison of the AR for LCC-VQE and GW algorithm on 100-vertex d-regular graph instances, relative to the exact solution found by GUROBI. Each degree corresponds to a set of 8 different instances. The boxplots of both algorithms show the best AR chosen out of 24 trials for each instance, with 8 instances at each degree. (b) Comparison of the AR for LCC-VQE and GW algorithm on 100-vertex $G(n,p)$ graph instances. For each edge probability $p = 0.1, 0.2, 0.3$, we generate 8 graph instances. The boxplots show the best AR for each instance, selected from 24 trials for LCC-VQE and the GW algorithm. The red lines and white diamonds of boxplots represent the medians and means respectively.

backend. The 36 instances used are the same as those in Fig. 6. This figure is to show the effect of noise mitigation after LCC attributed to the reduction in the number of gates. The diagonal dashed line shows where both methods (with and without LCC) have the same AR. All the points are in the lower triangle, which represents that LCC gives higher ARs than without having LCC. Under the same noise conditions, it can be observed that all 36 instances give higher ARs with LCC applied. This result shows the effect of the reduction in the number of gates in a more evident way than the demonstration shown in Fig. 6, as the number of qubits and the error rates are the same for both with and without LCC. It is also observed that problems with larger sizes n benefit more from LCC as their ARs stay away from the diagonal dashed line, and those with smaller sizes stay near the diagonal dashed line. This is because larger circuits generally have more noise, causing their ARs to deteriorate more.

3.2 LCC vs. Goemans-Williamson (GW) algorithm

It is worth noting that under noisy conditions, although the solutions found by LCC-VQE are capable of approximating the true solutions of problems, they consistently fall short of those achieved by the GW algorithm. The GW algorithm serves as a typical benchmark for classical approximation algorithms for the Max-Cut problem. To better evaluate the inherent potential of LCC-VQE without the influence of noise, we therefore conduct further comparative simulations under noiseless conditions. Fundamentally, according to Eq. (15), the comparison is essentially between the VQE algorithm and the GW algorithm, regardless of whether LCC is applied. Previous studies[34] have observed that the performance of the GW algorithm deteriorates as the problem size increases. Furthermore, numerical simulations of XQAOA and the GW algorithm on 128- and 256-vertex regular graphs have shown that the AR of GW tends to decrease as the degree of the graph increases[35].

Motivated by these observations, we seek to establish an effective benchmarking strategy for evaluating the performance of LCC-VQE by comparing it against GW on large size problem instances. Consequently, we construct another dataset where all instances have 100 vertices but different degrees. The subsequent comparative simulations employ LCC in a noiseless environment using the StatevectorEstimator provided by Qiskit, and the circuit remains previous two-local ansatz. To obtain the exact Max-Cut values and compute AR in these simulations, we employ the GUROBI solver[36], which is widely used in industry. Fig. 8(a) shows the comparison of AR for both algorithms on 100-vertex d-regular graph instances with degrees d ranging from 3 to 10, giving a total of 64 instances. Each boxplot represents the best ARs across 8 instances at each degree, where each best AR is selected from 24 trials of each instance. The result shows that the GW algorithm demonstrates consistently strong performance on low-degree regular instances ($d = 3, 4$), with the majority of ARs in the boxplots exceed 0.99 and are

close to 1. However, although the overall performance of GW declines as the degree increases, the trend is neither gradual nor monotonic. In particular, the lowest median AR across all degrees is observed at $d = 5$. In contrast, LCC-VQE exhibits a steadily increasing trend in ARs as the degree increases, and its median values gradually improve as the degree increases. It begins to outperform GW in the upper quartile, median, and lower quartile values on graphs with degree $d \geq 9$.

Then, we conduct the comparative simulation for LCC-VQE and the GW algorithm on 100-vertex Erdös-Rényi (also known as $G(n,p)$) graph instances. Consistent with our previous work, we employ the COBYLA optimizer to optimize the parameters in this simulation. Fig. 8(b) presents the comparison of the AR across graph instances with different edge probabilities p. For each probability $p = 0.1, 0.2, 0.3$, 8 graph instances are generated, giving a total of 24 instances. Each boxplot represents the best ARs across 8 instances at each edge probability, where for each instance the best AR is selected from 24 trials for both LCC-VQE and the GW algorithm.

As shown in the Fig. 8(b), LCC-VQE surpasses GW in terms of the upper quartile, lower quartile, mean, and median ARs at different edge probability p. Notably, the number of edges in the instances at $p = 0.1$ is comparable to that in Fig. 8(a) for graphs with degree $d = 10$. The number of edges in regular graphs is $nd/2$, whereas the average number of edges in $G(n,p)$ graphs is $pn(n-1)/2$. Despite this similarity, the upper quartile and median ARs of LCC-VQE at $p = 0.1$ are higher than that at $d = 10$. As the edge probability increases to $p = 0.2$, the AR of LCC-VQE further increases. Although a slight drop is observed at $p = 0.3$, the majority of the box remains above 0.99.

Fig. 8(a) and (b) show similar trend in the comparison of XQAOA and GW in[35], in which LCC-VQE (or XQAOA) initially shows lower AR than GW at small degrees, but surpasses the performance of GW as the degree increases.

4 Conclusion

In this work, we presented the LCC on VQE and studied what the effect it acts in solving the Max-Cut problem. Our work opens up the possibility of using VQE to solve combinatorial optimization problems, as VQE requires less number of layers than QAOA to achieve the same performance. This allows us to cancel a larger number of qubits when applying LCC, thereby shifting the complexity of circuit simulation from exponential scaling (number of qubits) to polynomial scaling (number of edges in a graph). For the Max-Cut problem with a two-local cost Hamiltonian, only at most five qubits are required to solve the problem of any size. Concerning the implementation of LCC, our preliminary calculations reveal the relationship between the maximum number of qubits n_q with k-local observables in calculating the expectation of Hamiltonian H_C. It is worth noting that LCC can only be implemented under linear and circular entanglement structures, whereas it is not feasible under full entanglement. To look at the precise relation between high-quality solution and layer numbers in addressing Max-Cut problem, we compare the performance of VQE ansatz with different number of layers and conclude that the opportunity to achieve a high-quality solution (AR ≥ 0.99) comes to decrease with the number of layers L increases. This decline is attributed to tendency to become trapped in local minimas due to overparameterization. Meanwhile, the computational cost increases with the number of layers L, and those causes make it necessary for the optimization process to achieve faster convergence rate by setting the ansatz to a single layer. Furthermore, as the number of layers L increases, the size of subcircuits with the maximum number of qubits $n_q = 4L + 1$ also grows. This can significantly undermine the effectiveness of LCC in a noisy environment.

We compare the performance of circuits with and without LCC on a noisy simulator provided by Qiskit. The results show that the circuits with LCC generally yield higher approximation ratios than the circuits without LCC, hence implying that the noise is being mitigated. Furthermore, the performance of LCC-VQE and the GW algorithm is compared under a noiseless condition. The results show that even on denser graphs, LCC-VQE exhibits significant advantages and potential. We are aware that the simulation results from the fake backend might differ from that of the real quantum hardware, as some of the noises, e.g., coherent errors, crosstalk, non-Markovian noises, etc. Benchmarking results show that there exist discrepancies between the noisy simulation from a fake backend and that from a real hardware[37,38]. This means that extra efforts are required to create dedicated noise models that include the effect of these noises in the simulation of noisy backends, leading to an active area of research. Weber et al. adopted a machine learning approach, where they trained a noise model that included crosstalk and state preparation noise to better mimic the behavior of a real device[39]. Maschek et al. proposed a more realistic noise model and applied it in the QAOA simulation of solving the job shop scheduling problem[40]. Although the performance of LCC on a real noisy quantum hardware is not included in this work, we anticipate that it can be further improved using typical error mitigation, such as zero-noise extrapolation (ZNE) or probabilistic error cancellation (PEC). Methods like the Conditional Value-at-Risk (CVaR) is also compatible with LCC and can be utilized to improved the results on COPs. We also noticed that the gate errors improves with better hardware developed by the quantum industries, with the two-qubit gate errors reaching the order of 10^{-3} in IBM Q devices at the time when this work is published[41], compared to the two-qubit gate errors in the order of 10^{-2} for the fake backends used in the experiments. Hence, we leave the simulation on the real hardware for LCC as a future work.

Funding Declaration

This work was supported by JST SPRING, Grant Number JPMJSP2124 and by the National Research Foundation, Singapore under its Quantum Engineering Programme 2.0 (NRF2021-QEP2-02-P01).

Data availability

We provide the code that implements the LCC framework, along with the benchmark and simulation datasets used in this paper. These resources is publicly available at https://github.com/xenoicwyce/lcc[42].

Author contributions

X. Lee contributed to the idea, devised the main experiment and analyzed the results. X. Yan conducted the experiments and prepared the materials (figures and tables) for the manuscript. N. Xie derived the mathematical formulation for LCC. Y. Saito and L. Kurosawa analyzed the results. N. Asai, D. Cai and H.C. LAU supervised the project and revised the manuscript. All authors reviewed the manuscript.

References

1. Cerezo, M. *et al.* Variational quantum algorithms. *Nat. Rev. Phys.* **3**, 625–644 (2021).
2. Farhi, E., Goldstone, J. & Gutmann, S. A quantum approximate optimization algorithm (2014). arXiv:1411.4028.
3. Zhou, L., Wang, S.-T., Choi, S., Pichler, H. & Lukin, M. D. Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices. *Phys. Rev. X* **10**, 021067, DOI: 10.1103/PhysRevX.10.021067 (2020).
4. Cook, J., Eidenbenz, S. & Bärtschi, A. The quantum alternating operator ansatz on maximum k-vertex cover. *2020 IEEE Int. Conf. on Quantum Comput. Eng. (QCE)* 83–92 (2020).
5. Lee, X., Xie, N., Cai, D., Saito, Y. & Asai, N. A depth-progressive initialization strategy for quantum approximate optimization algorithm. *Mathematics* **11**, 2176, DOI: 10.3390/math11092176 (2023).
6. Farhi, E., Goldstone, J., Gutmann, S. & Sipser, M. Quantum computation by adiabatic evolution (2000). arXiv: quant-ph/0001106.
7. Sack, S. H. & Serbyn, M. Quantum annealing initialization of the quantum approximate optimization algorithm. *Quantum* **5**, 491, DOI: 10.22331/q-2021-07-01-491 (2021).
8. Peruzzo, A. *et al.* A variational eigenvalue solver on a photonic quantum processor. *Nat. communications* **5**, 4213 (2014).
9. Sim, S., Johnson, P. D. & Aspuru-Guzik, A. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Adv. Quantum Technol.* **2**, DOI: 10.1002/qute.201900070 (2019).
10. Yan, X., Lee, X., Cai, D. & Asai, N. Comparison between the performances of general two-local ansatzes and qaoa in max-cut problem. *IPSJ SIG Tech. Rep.* **2024-QS-11**, 1–8 (2024).
11. Herrman, R., Lotshaw, P. C., Ostrowski, J., Humble, T. S. & Siopsis, G. Multi-angle quantum approximate optimization algorithm. *Sci. Reports* **12**, 6781 (2022).
12. Preskill, J. Quantum computing in the nisq era and beyond. *Quantum* **2**, 79, DOI: 10.22331/q-2018-08-06-79 (2018).
13. Lowe, A. *et al.* Unified approach to data-driven quantum error mitigation. *Phys. Rev. Res.* **3**, DOI: 10.1103/physrevresearch.3.033098 (2021).
14. Leone, L., Oliviero, S. F., Cincio, L. & Cerezo, M. On the practical usefulness of the hardware efficient ansatz. *Quantum* **8**, 1395 (2024).
15. Pelofske, E., Bärtschi, A. & Eidenbenz, S. Short-depth qaoa circuits and quantum annealing on higher-order ising models. *npj Quantum Inf.* **10**, 30, DOI: 10.1038/s41534-024-00825-w (2024).
16. Huang, C. *et al.* Efficient parallelization of tensor network contraction for simulating quantum computation. *Nat. Comput. Sci.* **1**, 578–587, DOI: 10.1038/s43588-021-00119-7 (2021).
17. Lykov, D., Schutski, R., Galda, A., Vinokur, V. & Alexeev, Y. Tensor network quantum simulator with step-dependent parallelization. In *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)*, 582–593, DOI: 10.1109/QCE53715.2022.00081 (2022).

18. Lykov, D. & Alexeev, Y. Importance of diagonal gates in tensor network simulations. In *2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 447–452, DOI: 10.1109/ISVLSI51109.2021.00088 (2021).

19. Vidal, G. Class of quantum many-body states that can be efficiently simulated. *Phys. Rev. Lett.* **101**, DOI: 10.1103/physrevlett.101.110501 (2008).

20. Frías-Pérez, M. & Bañuls, M. C. Light cone tensor network and time evolution. *Phys. Rev. B* **106**, DOI: 10.1103/physrevb.106.115117 (2022).

21. Haghshenas, R., Gray, J., Potter, A. C. & Chan, G. K.-L. Variational power of quantum circuit tensor networks. *Phys. Rev. X* **12**, DOI: 10.1103/physrevx.12.011047 (2022).

22. Suzuki, Y., Sakuma, R. & Kawaguchi, H. Light-cone feature selection for quantum machine learning. *Adv. Quantum Technol.* **8** (2025).

23. Goemans, M. X. & Williamson, D. P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM (JACM)* **42**, 1115–1145 (1995).

24. Karp, R. M. Reducibility among combinatorial problems. In *50 Years of Integer Programming 1958-2008: from the Early Years to the State-of-the-Art*, 219–241 (Springer, 2009).

25. Jagannath, A., Ko, J. & Sen, S. Max κ-cut and the inhomogeneous potts spin glass. *The Annals Appl. Probab.* **28**, 1536–1572 (2018).

26. Benedetti, M., Fiorentini, M. & Lubasch, M. Hardware-efficient variational quantum algorithms for time evolution. *Phys. Rev. Res.* **3**, 033083 (2021).

27. Brandao, F. G. S. L., Broughton, M., Farhi, E., Gutmann, S. & Neven, H. For fixed control parameters the quantum approximate optimization algorithm's objective function value concentrates for typical instances (2018). arXiv:1812.04170.

28. Basso, J., Farhi, E., Marwaha, K., Villalonga, B. & Zhou, L. The quantum approximate optimization algorithm at high depth for maxcut on large-girth regular graphs and the sherrington-kirkpatrick model. DOI: 10.4230/LIPICS.TQC.2022.7 (Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022).

29. Wurtz, J. & Lykov, D. Fixed-angle conjectures for the quantum approximate optimization algorithm on regular maxcut graphs. *Phys. Rev. A* **104**, 052419, DOI: 10.1103/PhysRevA.104.052419 (2021).

30. Galda, A., Liu, X., Lykov, D., Alexeev, Y. & Safro, I. Transferability of optimal qaoa parameters between random graphs. In *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)*, 171–180 (IEEE, 2021).

31. Powell, M. J. D. *A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation*, 51–67 (1994).

32. Javadi-Abhari, A. *et al.* Quantum computing with Qiskit, DOI: 10.48550/arXiv.2405.08810 (2024). 2405.08810.

33. Qiskit contributors. Qiskit: An open-source framework for quantum computing, DOI: 10.5281/zenodo.2573505 (2023).

34. Muñoz-Arias, M. H., Kourtis, S. & Blais, A. Low-depth clifford circuits approximately solve maxcut. *Phys. Rev. Res.* **6**, 023294 (2024).

35. Vijendran, V., Das, A., Koh, D. E., Assad, S. M. & Lam, P. K. An expressive ansatz for low-depth quantum approximate optimisation. *Quantum Sci. Technol.* **9**, 025010 (2024).

36. Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual (2024).

37. Bravo-Montes, J. A., Bastante, M., Botella, G., del Barrio, A. & Garcia-Herrero, F. A methodology to select and adjust quantum noise models through emulators: benchmarking against real backends. *EPJ Quantum Technol.* **11** (2024).

38. Piskor, T. *et al.* Simulation and benchmarking of real quantum hardware (2025). 2508.04483.

39. Weber, T., Borras, K., Jansen, K., Krücker, D. & Riebisch, M. Construction and volumetric benchmarking of quantum computing noise models. *Phys. Scr.* **99**, 065106 (2024).

40. Maschek, S. R., Schwitalla, J., Franz, M. & Mauerer, W. Make some noise! measuring noise model quality in real-world quantum software. In *2025 IEEE International Conference on Quantum Software (QSW)*, 1–11, DOI: 10.1109/qsw67625.2025.00010 (IEEE, 2025).

41. Computing resources | ibm quantum platform. https://quantum.cloud.ibm.com/computers (2025).

42. Lee, X., Saito, Y. & Yan, X. LCC-Dataset (GitHub). https://github.com/xenoicwyce/lcc (2024).