IOP Publishing Quantum Sci. Technol. 11 (2026) 015060 https://doi.org/10.1088/2058-9565/ae3fc8

Quantum Science and Technology

OPEN ACCESS

RECEIVED
4 June 2025

REVISED
16 January 2026

ACCEPTED FOR PUBLICATION
29 January 2026

PUBLISHED
10 February 2026

PAPER

Weighted approximate quantum natural gradient for variational quantum eigensolver

Chenyu Shi[1,2], **Vedran Dunjko**[1,2] and **Hao Wang**[1,2,*]

[1] $\langle aQa^L \rangle$Applied Quantum Algorithms, Universiteit Leiden, 2333 CA Leiden, The Netherlands
[2] LIACS, Universiteit Leiden, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
[*] Author to whom any correspondence should be addressed.

E-mail: h.wang@liacs.leidenuniv.nl

Keywords: variational quantum eigensolver, quantum natural gradient descent, quantum optimization

Abstract

The variational quantum eigensolver (VQE) is one of the most prominent algorithms using near-term quantum devices, designed to find the ground state of a Hamiltonian. In VQE, a classical optimizer iteratively updates the parameters in the quantum circuit. Among various optimization methods, the quantum natural gradient descent (QNG) stands out as a promising optimization approach for VQE. However, standard QNG only leverages the quantum Fisher information of the entire system and treats each subsystem equally in the optimization process, without accounting for the different weights and contributions of each subsystem corresponding to each local term in the Hamiltonian. To address this limitation, we propose a Weighted Approximate QNG (WA-QNG) method tailored for k-local Hamiltonians. In this paper, we theoretically analyze the potential advantages of WA-QNG compared to QNG from three distinct perspectives and reveal its connection with the Gauss–Newton method. We also show it outperforms the standard QNG descent in the numerical simulations for seeking the ground state of the Hamiltonian.

1. Introduction

Quantum computing is widely regarded as having potential advances in numerous fields. However, due to the limitations of the noise and scale of current Noisy intermediate-scale (NISQ) quantum devices [1], algorithms such as Shor's [2] and Grover's [3] still remain beyond practical implementation. A computational paradigm well-suited for current NISQ quantum devices is that of the variational quantum algorithms, which is a kind of variational hybrid approach [4]. These algorithms leverage a feedback loop between classical and quantum computers. In this paradigm, the quantum computer evaluates an objective function formulated by a parameterized quantum circuit, while the classical computer employs an optimizer to iteratively update the circuit parameters to seek the optimal value [5].

Variational quantum algorithms have drawn significant attention across various fields, including quantum physics and quantum chemistry [6, 7], optimization [8, 9], and machine learning [10, 11]. Among these, one of the most well-known variational quantum algorithms is the variational quantum eigensolver (VQE) [7]. VQE is designed to find the ground state of a given quantum system. In this algorithm, a variational quantum circuit is employed to estimate the expectation value of the system with respect to a given Hamiltonian. Additionally, a classical optimizer iteratively updates the parameters in the quantum circuit to minimize the expectation value. Through this optimization process, the algorithm is expected to converge to a solution that closely approximates the ground state. Despite the high potential in this field, VQEs also face several challenges during optimization, such as barren plateaus [12–15]. Therefore, developing efficient optimization methods for VQEs is of great importance.

The optimizer in VQE plays a crucial role in determining the algorithm's performance. In addition to the most basic gradient descent method (referred to as vanilla gradient descent in this context), more advanced variants such as stochastic gradient descent [16] and Adam [17] are widely adopted. Among these, the quantum natural gradient descent (QNG) [18] emerges as a promising approach. It is the

quantum analog of natural gradient descent [19, 20] in its classical counterpart. QNG leverages the quantum Fisher information matrix [21] of the quantum system. It captures the geometric information and is expected to obtain better convergence performance in the optimization process.

In the standard formulation of QNG, the quantum Fisher information used in the optimization step corresponds to the entire quantum system. However, in VQE, particularly in quantum chemistry, the Hamiltonian H is often expressed as a summation of several local terms H_m with maximum locality k, where $H = \sum_m h_m H_m$ with the corresponding output of the quantum circuit $\mathrm{tr}(\rho H) = \sum_m h_m \mathrm{tr}(\rho_m H_m)$. Recent studies [15, 22] have shown that leveraging the locality structure of k-local Hamiltonian can improve the performance of finding ground state.

For a k-local Hamiltonian, each subsystem ρ_m corresponding to each local term H_m contributes differently to the final output of the quantum circuit due to the different weights h_m. Therefore, a potential improvement for standard QNG is to assign these subsystems with different weights during the training, rather than only using the quantum Fisher information matrix of the entire system, where all subsystems are treated equally. Hence, here we propose a new approach, the Weighted Approximate QNG (WA-QNG), which takes the different weights and contributions of the subsystem corresponding to each local observable term into account.

In WA-QNG, we replace the quantum Fisher information matrix of the entire quantum system with the weighted summation of the Hilbert–Schmidt metric tensors of the subsystems corresponding to each local term in the optimization step. We theoretically analyze the potential advantages of WA-QNG compared to QNG from three distinct perspectives and reveal its connection to the Gauss–Newton method. Our method displays efficient convergence speed compared to standard QNG in the numerical simulations. Furthermore, we show that the Hilbert–Schmidt metric tensor required for WA-QNG can be efficiently estimated using the classical shadow method [23] for k-local Hamiltonians.

The remainder of the paper will be structured as follows. Section 2 introduces the preliminary background knowledge, including the VQE and the QNG descent. Section 3 formulates the WA-QNG method and theoretically analyzes its potential advantages over standard QNG from three different perspectives. Section 4 explores the connection between WA-QNG and the Gauss–Newton method. Section 5 discusses the computational costs of WA-QNG compared to standard QNG. Section 6 presents the numerical results to support our theoretical analysis. Finally, section 7 concludes the paper.

2. Background

In this section, we provide a brief overview of the foundational background relevant to this paper. First, we briefly introduce VQE and explain its working principles. Then, we give the definition of the QNG descent and discuss its relation with quantum Fisher information.

2.1. Variational quantum eigensolver

The goal of the VQE, initially introduced by [7], is to approximately seek the ground state ρ_{GS} with respect to a Hamiltonian H. A variational quantum circuit is used to prepare a variational quantum state ρ_θ, where $\rho_\theta = U(\theta)|0\rangle\langle 0|U^\dagger(\theta)$. The expectation value with respect to the Hamiltonian H is evaluated by the quantum computer. An illustration of variational quantum circuit in VQE is given in figure 1. Hence, the variation quantum circuit realizes the following function f:

$$f(\theta) = \mathrm{tr}\left(\rho_\theta H\right). \tag{1}$$

According to the definition, the ground state ρ_{GS} is the lowest energy state of the given Hamiltonian H. Hence, seeking the ground state by adjusting the parameter θ is equivalent to minimizing the function $f(\theta)$.

The value of function $f(\theta)$ is fed to a classical computer in the optimization process. The classical computer employs an optimizer, where $f(\theta)$ is the objective function, to iteratively update the parameters in the quantum circuit. The most common optimization method is the vanilla gradient descent. In each optimization step, the parameters are updated by:

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla f\left(\theta^{(k)}\right) \tag{2}$$

where ∇f is the gradient of the objective function and η is the learning rate. In general, the gradient can be approximated using a naive finite difference method. In VQE, however, the most common approach to obtain the exact gradient is through the parameter-shift rule [24, 25], up to finite sampling errors.

The general working principle of VQE is also illustrated in figure 1. After sufficient training with an expressive variational circuit, VQE is expected to produce a good approximation of the ground state.

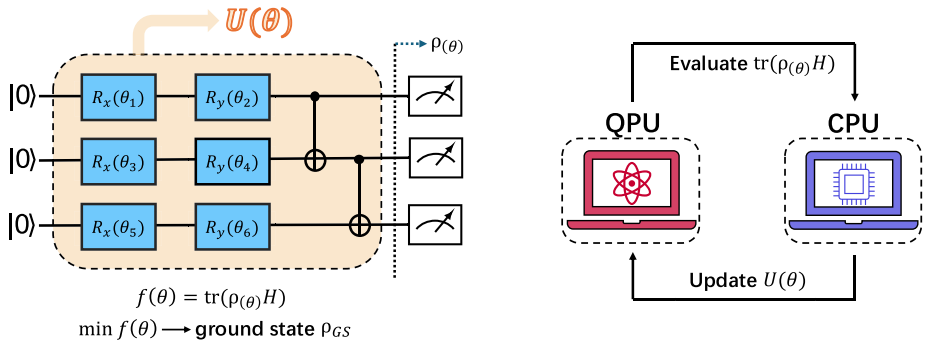

Figure 1. Left: An illustration of the variational quantum circuit in VQE. A parameterized quantum circuit $U(\theta)$ is used to prepare a variational quantum state ρ_θ. By adjusting the parameter θ, the quantum circuit aims to prepare the state ρ_θ that approximates the ground state ρ_{GS} by minimizing the objective function $f(\theta)$. Right: The general working principle of VQE. The quantum computer evaluates the expectation value $f(\theta) = \mathrm{tr}(\rho_\theta H)$, while the classical computer employs an optimizer to iteratively update the parameters θ to minimize the objective function $f(\theta)$.

2.2. QNG

Each optimization update step in vanilla gradient descent can be formulated as the following constrained optimization problem [20, 26] for a small ε with a small change δ in parameter space:

$$\begin{aligned} \min_{\delta} \quad & f(\theta+\delta) \\ \text{s.t.} \quad & \|\delta\|_2^2 \leqslant \varepsilon \end{aligned} \tag{3}$$

This constrained optimization problem seeks to minimize the objective function f within a local neighborhood of $\theta^{(k)}$. Note that solving this problem by applying a first-order Taylor approximation to f leads to the derivation of the vanilla gradient descent step in equation (2).

A limitation of this method is that each step is inherently tied to the Euclidean geometry of the parameter space, as the Euclidean distance is used to define the local neighborhood in the constrained optimization problem. However, the distances in the optimization landscape can be distorted in reparameterization—for example, directions that were equally steep may become scaled differently, potentially leading to inefficiencies in the optimization process [18, 19].

An illustration is shown in figure 2. The parameter space is a Euclidean space $[0,\pi]\times[0,2\pi]$, where each coordinate corresponds to the polar angle and azimuthal angle of a sphere. The parametrization maps each point in the original parameter space to a point on the surface of a unit sphere using the coordinate transformation $x=\sin(\theta)\cos(\phi)$, $y=\sin(\theta)\sin(\phi)$, $z=\cos(\theta)$. In the original parameter space, the distances between points A and B (red line) and between points C and D (purple line) are the same in terms of Euclidean distance. However, after parametrization onto the sphere, their distances differ significantly [3]. An intuitive example is as follows: Suppose the original parameter space consists of latitude and longitude pairs, and the parametrization maps each pair to a point on the surface of the Earth. It is more natural to describe distances using the great-circle path distance defined directly on the Earth's surface (i.e. distance after parametrization) rather than the Euclidean distance between latitude and longitude pairs (i.e. distance in the parameter space).

Similarly, in the parametrization from the parameter θ to the variational quantum state ρ_θ, the distance distortion can also occur. Changes with the same Euclidean distance in the parameter space can result in different changes in the variational quantum state ρ_θ. Therefore, it is more natural to use a distance metric defined directly for the variational quantum state ρ_θ to reformulate the constrained optimization problem, rather than relying on the Euclidean distance defined for θ in the original parameter space:

$$\begin{aligned} \min_{\delta} \quad & f(\theta+\delta) \\ \text{s.t.} \quad & D_F(\rho_\theta, \rho_{\theta+\delta}) \leqslant \varepsilon \end{aligned} \tag{4}$$

where the distance metric $D_F(\rho,\sigma) = 1-(\mathrm{tr}(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}))^2$ is the fidelity distance. For pure states $\rho=|\phi\rangle\langle\phi|$ and $\sigma=|\psi\rangle\langle\psi|$, the fidelity distance can also be formulated as $D_F(|\phi\rangle,|\psi\rangle) = 1-|\langle\phi|\psi\rangle|^2$.

[3] In figure 2, we use the chord length as the distance metric between two points on the sphere for simplicity. Note that the chord length is always proportional to the great-circle distance, which represents the true geodesic distance on the sphere.

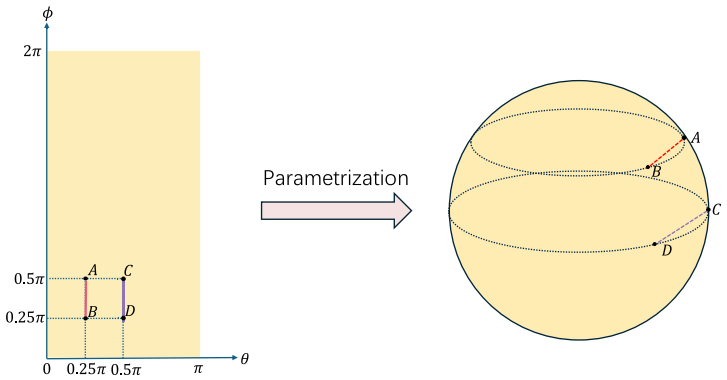

Figure 2. An illustration of distance distortion in the parametrization. The original parameter space is an Euclidean space. The parametrization maps the point in the parameter space to a point on the sphere with radius $r = 1$ by the coordinate transformation $x = \sin(\theta)\cos(\phi)$, $y = \sin(\theta)\sin(\phi)$, $z = \cos(\theta)$. The distances from A to B (red line) and from C to D (purple line) are the same in the original parameter space. However, after parametrization to the sphere, the distances are distorted and become different.

Solving the above optimization problem derives the QNG update step:

$$\theta^{(k+1)} = \theta^{(k)} - \eta F^{+} \nabla f\left(\theta^{(k)}\right) \tag{5}$$

where F^{+} is the pseudo-inverse of the quantum Fisher information matrix F at $\theta^{(k)}$, and η is the learning rate. For a pure state $|\psi(\theta)\rangle$, the quantum Fisher information matrix F is given by:

$$F_{ij} = 4\,\mathrm{Re}\left\{\langle\frac{\partial\psi}{\partial\theta_i}|\frac{\partial\psi}{\partial\theta_j}\rangle - \langle\frac{\partial\psi}{\partial\theta_i}|\psi\rangle\langle\psi|\frac{\partial\psi}{\partial\theta_j}\rangle\right\} \tag{6}$$

where θ_i denotes the i-th element of the parameter θ, and F_{ij} represents the (i,j)-th entry of the quantum Fisher information matrix F. For details on the derivation, please refer to [18] and [27]. A detailed discussion of the derivation is also provided in appendix D. QNG has been shown to achieve better performance compared to vanilla gradient descent in previous studies [18, 19, 28].

3. WA-QNG

In this section, we present the formulation of the WA-QNG method and discuss its potential advantages. First, we highlight the limitation of QNG, where the weights of subsystems corresponding to each local term are not considered. To address this issue, we introduce WA-QNG, which leverages the weighted sum of the Hilbert–Schmidt metric tensors of each subsystem in the optimization step. Additionally, we demonstrate WA-QNG's potential advantages from three different perspectives.

3.1. Limitation of QNG

The quantum Fisher information matrix of ρ_θ in equation (5) represents the quantum Fisher information matrix of the entire quantum system. A key limitation is that this quantum Fisher information does not incorporate any information about the Hamiltonian. Consequently, QNG utilizes the same quantum Fisher information matrix F in the update formula for two different Hamiltonians, which captures the sensitivity of the quantum state with respect to parameter changes. However, in VQE, we care more about how the sensitivity of the final objective function value with respect to parameter changes, rather than the sensitivity of the quantum state with respect to parameter changes itself. Therefore, information about the Hamiltonian should also be taken into account during the optimization process.

Consider a k-local Hamiltonian $H = \sum_m h_m H_m$ and its expectation value $\mathrm{tr}(\rho_\theta H) = \sum_m h_m \mathrm{tr}(\rho_m H_m)$, where each H_m is a Pauli string that acts non-trivially on a subsystem state ρ_m of k qubits. Intuitively, the contributions of each subsystem ρ_m are different due to the associated weights h_m. An illustration (shown in figure 3) is the 3-qubit toy Hamiltonian $H = Z_1Z_2 + 2Z_2Z_3 + X_1 + 0.5X_2 + 1.5X_3$. The contributions of the five subsystems, corresponding to the five terms in the Hamiltonian, are weighted by their respective coefficients. Therefore, in the optimization process, a potential improvement is to also assign appropriate weights to each subsystem, rather than treating them equally as in standard QNG. In addition, as the total system size increases, the parameter sensitivity within each subsystem may differ significantly from that of the entire system. In such cases, the quantum Fisher information matrix F

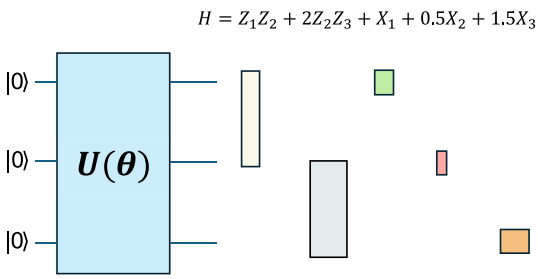

Figure 3. An illustration when the Hamiltonian is $H = Z_1Z_2 + 2Z_1Z_2 + X_1 + 0.5X_2 + 1.5X_3$. The small rectangles indicate that the corresponding Hamiltonian term acts on that subsystem. And the width of rectangles reflects the magnitude of each coefficient h_m. Each subsystem contributes differently to the output due to the coefficients, suggesting that they should be assigned different weights in the optimization process.

of the entire system may struggle to capture the parameter sensitivity of each subsystem. To address all these aspects that are overlooked in QNG, we propose the WA-QNG method.

3.2. Method formulation

In this subsection, we formalize the WA-QNG method. Suppose the target Hamiltonian to solve is $H = \sum_m h_m H_m$. Then the update formula of WA-QNG is then given by:

$$\theta^{(k+1)} = \theta^{(k)} - \eta W^{+} \nabla f\left(\theta^{(k)}\right) \tag{7}$$

where $W = \frac{2}{\sum_m h_m^2} \sum_m h_m^2 T_m$, and T_m represents the Hilbert–Schmidt metric tensor [29, 30] of the m-th subsystem corresponding to the observable term H_m. The i-th row j-th column element of each T_m is that $(T_m)_{ij} = \mathrm{tr}(\partial_i \rho_m \partial_j \rho_m)$. Because the coefficient h_m can be negative, we use square-weighted summation instead of direct-weighted summation. The prefactor $\frac{2}{\sum_m h_m^2}$ is to make WA-QNG consistent with QNG: If each term H_m in the Hamiltonian acts globally on the whole quantum system, then WA-QNG will reduce to QNG and $W = 2T = F$. The proof of this equivalence is detailed in appendix A. From this perspective, WA-QNG can be regarded as a more generalized form of the original QNG.

In the following, we further motivate and explain why the matrix W, namely the weighted average of the Hilbert–Schmidt metric tensors of each subsystem, is chosen in the updated formula of WA-QNG, and also why it is expected to perform well from three different interpretative perspectives.

3.3. Intuitive interpretation

The quantum Fisher information matrix F of the entire system is used in the optimization step of QNG. Mathematically, because F is unrelated to the index m, it can be rewritten as:

$$\begin{aligned} F &= 1 \cdot F \\ &= \frac{1}{\sum_m h_m^2} \sum_m h_m^2 F \end{aligned} \tag{8}$$

where h_m is the m-th coefficient of the k-local Hamiltonian $H = \sum_m h_m H_m$. However, since the m-th observable term H_m and coefficient h_m are only associated with the subsystem ρ_m, an intuitive way to address the different contributions of each subsystem is to replace F in equation (8) with F_m, the quantum Fisher information matrix of the corresponding subsystem ρ_m. Therefore, we define a new matrix $\hat{F}$:

$$\hat{F} = \frac{1}{\sum_m h_m^2} \sum_m h_m^2 F_m. \tag{9}$$

Compared to equations (8) and (9) incorporates a weighted sum of the quantum Fisher information matrices of individual subsystems. Here, h_m^2 serves as a weight to quantify the influence of each subsystem. By taking a weighted sum over the quantum Fisher information matrix of each subsystem ρ_m corresponding to H_m, this formulation explicitly accounts for the different weights and contributions of each subsystem.

However, since each subsystem state will be a mixed state in general, and as noted in [21, 29, 31], the estimation of the quantum Fisher information for a mixed state is generally a more computationally demanding task than for a pure state. Consequently, estimating the matrix F_m required in equation (9)

is computationally demanding. To deal with this problem, [29, 30] propose using the Hilbert–Schmidt metric tensor T as an approximation of the quantum Fisher information F in QNG for a mixed state, where $F \approx 2T$ when the mixed state is close to being pure and does not change dramatically with parameters. We also demonstrate this approximation relation in appendix C. Therefore, we can approximate each F_m in equation (9) using the Hilbert–Schmidt metric tensor T_m:

$$W = \frac{2}{\sum_m h_m^2} \sum_m h_m^2 T_m \tag{10}$$

here we exactly obtain the matrix W used in equation (7) of the update rule of WA-QNG. Equation (9) introduces a weighted average to account for the relative importance of each subsystem. From equation (9) to equation (10), the Hilbert–Schmidt metric tensor is employed to approximate the quantum Fisher information matrix of each subsystem state. Hence, this is where the name of our method WA-QNG: *QNG* comes from.

3.4. Optimization interpretation

The constrained optimization problem defined in equation (4) uses the fidelity distance D_{F} as the distance metric. As discussed in section 2.2, using D_{F} instead of the Euclidean distance in the parameter space leads to the derivation of the QNG update formula. However, the distance metric D_{F} for the entire quantum system still does not account for the different weights of each subsystem with respect to the observable terms. To capture this characteristic, we introduce the following distance:

$$D_{\mathrm{W}}\left(\rho\left(\theta+\delta\right), \rho\left(\theta\right)\right) = \frac{2}{\sum_m h_m^2} \sum_m h_m^2 \| \rho_m\left(\theta+\delta\right) - \rho_m\left(\theta\right) \|_2^2 \tag{11}$$

where h_m is the m-th coefficient of the k-local Hamiltonian $H = \sum_m h_m H_m$, and ρ_m represents the subsystem state of $\rho(\theta)$ corresponding to the m-th subsystem. Additionally, the coefficient h_m^2 serves as a weight in the weighted averaging process of the 2-norm distance, thereby accounting for the different contributions of each subsystem. Hence, similar to equation (4), we define the following constrained optimization problem for each update step:

$$\begin{aligned} \min_{\delta} \quad & f(\theta+\delta) \\ \text{s.t.} \quad & D_{\mathrm{W}}\left(\rho\left(\theta\right), \rho\left(\theta+\delta\right)\right) \leqslant \varepsilon \end{aligned} \tag{12}$$

From the constrained optimization problem defined above, we can derive the same update rule of WA-QNG as given in equation (7). Thus, we establish WA-QNG from the optimization perspective. The detailed derivation from this optimization problem to WA-QNG is provided in appendix D.

3.5. Geometric interpretation

The interpretation in the previous subsection can be further explained from the perspective of Riemannian geometry. In general, consider a function $F: \Theta \to \mathcal{M}$ from the parameter space $\Theta \subseteq \mathbb{R}^A$ to a Riemannian manifold $\mathcal{M}$ equipped with a Riemannian metric $g_{\mathcal{M}}$. A pullback metric [4] g on Θ is induced by function F, which is defined as [32]:

$$\begin{aligned} g_{ij} &= \left(F^* g_{\mathcal{M}}\right)\left(d\theta_i, d\theta_j\right) \\ &= g_{\mathcal{M}}\left(\partial_i F(\theta), \partial_j F(\theta)\right). \end{aligned} \tag{13}$$

Then, the Riemannian gradient descent [18, 19, 33] with pullback metric can be defined for $\mathcal{M}$:

$$\theta^{(k+1)} = \theta^{(k)} - \eta g(\theta)^{+} \nabla f\left(\theta^{(k)}\right). \tag{14}$$

Now, consider a quantum state $\rho(\theta)$ prepared by a parameterized circuit. For QNG, the quantum circuit defines a mapping function $F: \mathbb{R}^A \to \mathbb{H}^N$, where $\mathbb{H}^N$ represents the $N \times N$ Hermitian matrix

[4] Strictly, only when F is an immersion, the pullback metric defined in equation (13) is guaranteed actually to be a Riemannian metric. However, the pullback metric defined in equation (13) is always well-defined and ensures that the update formula in equation (14) works in general.

space. In this view, QNG can be considered to work with the pullback metric of the Frobenius inner product $g_{\mathbb{H}}(\rho,\sigma)=\mathrm{tr}(\rho\sigma)$ defined on Hermitian matrix space:

$$\begin{aligned} g_{ij}(\theta) &= \left(F^{*}g_{\mathbb{H}}\right)\left(d\theta_i, d\theta_j\right) \\ &= g_{\mathbb{H}}\left(\partial_i\rho(\theta), \partial_j\rho(\theta)\right) \\ &= \mathrm{tr}\left(\partial_i\rho(\theta)\,\partial_j\rho(\theta)\right) \end{aligned} \tag{15}$$

which is consistent with the update rule of QNG in equation (5), up to a constant factor 2. In WA-QNG, to address the different weights and contributions of each subsystem, we view the mapping function F as:

$$\theta \xrightarrow{\mathrm{F}} q(\theta) = \frac{1}{\sqrt{\sum_m h_m^2}}\left[h_1\rho_1(\theta), \cdots, h_M\rho_M(\theta)\right] \tag{16}$$

where the $q(\theta)$ is a point in the product space $\mathbb{H}_1\times\cdots\times\mathbb{H}_M$ where each subsystem $\rho_m\in\mathbb{H}_m$. This product space is equipped with an inner product $\langle(\rho_1,\ldots,\rho_M),(\sigma_1,\ldots,\sigma_M)\rangle=\sum_{m=1}^{M}\mathrm{tr}(\rho_m\sigma_m)$. Similarly, the pullback metric can also be obtained as:

$$\begin{aligned} g_{ij}(\theta) &= \langle\partial_i q(\theta), \partial_j q(\theta)\rangle \\ &= \frac{1}{\sum_m h_m^2}\sum_m h_m^2\,\mathrm{tr}\left(\partial_i\rho_m(\theta)\,\partial_j\rho_m(\theta)\right) \end{aligned} \tag{17}$$

which is consistent with the update rule of WA-QNG in equation (7), also up to a constant factor 2. Compared to QNG, since each subsystem is explicitly represented with its corresponding coefficient in the direct product space, the pullback Riemannian metric tensor is more likely to account for the different weights of each subsystem.

4. Relation with Gauss–Newton method

As illustrated in [20], second-order optimization methods that leverage Fisher information can be interpreted as a generalized Gauss–Newton method. Here, we also demonstrate that the objective function equation (1) can be approximately transferred into a weighted non-linear least squares problem when each subsystem is close to being pure and does not change significantly with respect to parameters. Under this condition, we prove that WA-QNG is equivalent to the Gauss–Newton method for this non-linear least squares problem.

Let $\tilde{H}_m=-H_m$ and $\hat{H}_m=\frac{\tilde{H}_m}{h_m}$ for simplicity in the derivation. Also note that all constant factors, such as $\frac{1}{\sum_m h_m^2}$, do not affect the optimization formulation, as they can ultimately be absorbed into the learning rate. For simplicity, we use the symbol $\Leftrightarrow$ to represent two minimization problem are equivalent up to a constant factor. Starting from the optimization problem in equation (1), we can perform the following transformation:

$$\begin{aligned} &\min_\theta \mathrm{tr}\left(\rho(\theta)H\right) \\ &\Leftrightarrow \min_\theta \frac{1}{\sum_m h_m^2}\sum_m \mathrm{tr}\left(h_m\rho_m(\theta)H_m\right) \\ &\Leftrightarrow \min_\theta \frac{1}{\sum_m h_m^2}\sum_m -\mathrm{tr}\left(h_m\rho_m(\theta)\tilde{H}_m\right) \\ &\approx \min_\theta \frac{\sum_m\left(\mathrm{tr}\left(h_m^2\rho_m^2(\theta)\right)-2\mathrm{tr}\left(h_m\rho_m(\theta)\tilde{H}_m\right)+\mathrm{tr}\left(\tilde{H}_m^2\right)\right)}{\sum_m h_m^2}. \end{aligned} \tag{18}$$

Note that the third transformation is an approximate one, where two additional terms, $\mathrm{tr}(h_m^2\rho_m^2)$ and $\mathrm{tr}(\tilde{H}_m^2)$, are added into the summation. The latter is a constant so it does not affect the optimization. For the former, we leverage the assumption that each subsystem is close to being pure and does not change significantly with respect to the parameters. Under this condition, the term $\mathrm{tr}(h_m^2\rho_m^2)$ remains

close to a constant h_m^2, while the term $2\mathrm{tr}(h_m\rho_m\tilde{H}_m)$ dominants the gradient in the optimization problem. When ρ_m is exactly pure, the approximate transformation becomes exact. We then continue the transformation:

$$\begin{aligned}&\min_\theta \frac{\sum_m \left(\mathrm{tr}\left(h_m^2\rho_m^2(\theta)\right) - 2\mathrm{tr}\left(h_m\rho_m(\theta)\tilde{H}_m\right) + \mathrm{tr}\left(\tilde{H}_m^2\right)\right)}{\sum_m h_m^2}\\ \Leftrightarrow &\min_\theta \frac{1}{\sum_m h_m^2}\sum_m \|h_m\rho_m(\theta) - \tilde{H}_m\|_2^2\\ \Leftrightarrow &\min_\theta \frac{1}{\sum_m h_m^2}\sum_m h_m^2 \|\,\mathrm{vec}\left(\rho_m(\theta)\right) - \mathrm{vec}\left(\hat{H}_m\right)\|^2.\end{aligned} \tag{19}$$

In the final expression of equation (19), we observe that the original problem is transformed into a non-linear least squares problem. The update formula of Gauss–Newton method [34] for such a non-linear least squares problem with weights is given by:

$$\theta^{(k+1)} = \theta^{(k)} - \eta\left(J_\mathrm{r}^T J_\mathrm{r}\right)^{-1} J_\mathrm{r}^T\vec{r}\left(\theta^{(k)}\right) \tag{20}$$

$\vec{r}(\theta)$ is often referred to as the residual vector which is defined in equation (21) in our case, and J_r is the Jacobian of the residual with respect to the parameter θ.

$$\vec{r}(\theta) = \frac{\left[h_1\left(\mathrm{vec}\left(\rho_1\right) - \mathrm{vec}\left(\hat{H}_1\right)\right), \cdots, h_M\left(\mathrm{vec}\left(\rho_M\right) - \mathrm{vec}\left(\hat{H}_M\right)\right)\right]^T}{\sqrt{\sum_m h_m^2}} \tag{21}$$

Note that, as the objective function is defined by equation (19), where $f(\theta) = \frac{1}{\sum_m h_m^2}\sum_m h_m^2\|\,\mathrm{vec}(\rho_m) - \mathrm{vec}(\hat{H}_m)\|^2 = \vec{r}^T\vec{r}$, equation (20) can be rewritten into:

$$\theta^{(k+1)} = \theta^{(k)} - \eta\left(2J_\mathrm{r}^T J_\mathrm{r}\right)^{-1}\nabla f\left(\theta^{(k)}\right). \tag{22}$$

One can verify that the matrix W defined in WA-QNG satisfies the relation: $W = 2J_\mathrm{r}^T J_\mathrm{r}$. Hence, the update rule in equation (22) is fully equivalent to the update rule of WA-QNG in equation (7). The details of the relationship between the matrix W and the Gauss–Newton method are provided in appendix E.

Thus, we have demonstrated that WA-QNG is approximately equivalent to the Gauss–Newton method for a nonlinear least squares problem, under the assumption that each subsystem is close to being pure and does not vary significantly with respect to the parameters. Under this condition, WA-QNG is expected to inherit the favorable properties of the Gauss–Newton method and has the potential to outperform ordinary gradient descent.

5. Complexity analysis

In this section, we analyze the computational complexity in each optimization step of WA-QNG in comparison with standard QNG.

For both methods, it is necessary to estimate the gradient and the metric tensor at each optimization step. The computational costs for the gradient ∇f are identical for both methods, so the main comparison lies in the computational cost of obtaining the metric F in equation (5) and the weighted metric W in equation (7). For simplicity, the computational complexity discussed in this section refers particularly to that of computing the metric tensor, as the complexity of computing the gradient is the same for both methods.

For standard QNG, it is necessary to estimate the elements of the matrix F, so the number of elements to be estimated scales with v^2, where v is the number of parameters in the circuit. For WA-QNG, each metric tensor of the subsystem in the weighted summation must be estimated, so the total number of elements to be estimated scales with mv^2, where m is the number of local terms in the Hamiltonian. For both methods, each element in the corresponding matrices can be estimated using the Hadamard test or the Swap test [22, 29], both of which require additional ancillary circuits. In this case, WA-QNG requires a larger number of tests in each optimization step compared to standard QNG.

In scenarios with limited circuit scale, where additional ancillary circuits are not feasible, the classical shadow technique [23] can be employed to estimate the required elements of the matrices for both

methods. We show that the complexity of obtaining each element in the metric tensor is $\mathcal{O}(2^k)$, where k is the locality of the corresponding subsystem. The detailed derivation is provided in appendix B. For standard QNG, the quantum Fisher information matrix corresponds to the entire system, involving all qubits, i.e. $k = n$. In contrast, for WA-QNG, each metric tensor T_m in the weighted summation is associated only with a k-local subsystem. Therefore, the time complexity for estimating F in QNG is $\mathcal{O}(v^2 \cdot 2^n)$, while that for estimating W in WA-QNG is $\mathcal{O}(mv^2 \cdot 2^k)$.

In typical applications, the number of Hamiltonian terms m does not grow exponentially with the number of qubits n. For example, in molecular systems, $m = \mathcal{O}(n^4)$ [29]. Therefore, in scenarios with limited circuit scale, where the classical shadow technique is used to estimate metric tensors, WA-QNG is more shot-efficient than standard QNG in each optimization step. This is because the cost of WA-QNG scales exponentially with the locality k, while that of standard QNG scales exponentially with the total number of qubits n.

However, in both scenarios, each optimization step of WA-QNG remains more effective than that of standard QNG in terms of optimization performance. We will show this numerically in the next section.

6. Numerical results

In this section, we present the results of the numerical simulations. First, we compare the overall performance of standard QNG and WA-QNG for the 1D Ising model and Heisenberg model in section 6.1. Additionally, to better evaluate whether the design of WA-QNG to capture the different weights of subsystems effectively improves upon standard QNG, we design numerics to examine the effects of subsystem weights and the number of qubits in sections 6.2 and 6.3 respectively.

6.1. Performance comparison

To evaluate the performance of WA-QNG, we test it alongside standard QNG on the VQE for a 1D Ising model and Heisenberg model. Their Hamiltonians are given as follows, respectively:

$$H = \sum_{\langle ij \rangle} Z_i Z_j + \sum_i X_i \tag{23}$$

$$H = \sum_{\langle ij \rangle} \left(X_i X_j + Y_i Y_j + Z_i Z_j \right) \tag{24}$$

$\langle ij \rangle$ in the Hamiltonian denotes that the i-th and j-th qubits are the nearest neighbors. The variational quantum circuit used in our numerics is the widely used EfficientSU2, illustrated in figure 4 for the 4-qubit, single-layer example. In our numerics, we evaluate QNG and WA-QNG for these two Hamiltonians on 10-, 12-, and 14-qubit cases with a single-layer circuit. In addition, for the 10-qubit case, we also evaluate the methods with multiple EfficientSU2 layers to examine the influence of the number of layers.

In our numerics, standard QNG and vanilla gradient descent are used as baseline methods for comparison with WA-QNG. Both WA-QNG and QNG require additional shots to estimate the metric tensor, whereas vanilla gradient descent does not. For a detailed comparison between standard QNG and vanilla gradient descent, please refer to [18]. As discussed in section 4, WA-QNG has a solid theoretical connection to the Gauss–Newton method when each subsystem is nearly pure. To achieve this, we propose initializing the variational circuit with small angles to ensure low entanglement at the start. Moreover, small-angle initialization is believed to help mitigate issues such as the Barren Plateau problem [13, 35]. In our numerics, each parameter is uniformly randomly selected from the intervals $[-1, 1]$, which limits the angle magnitude to be relatively small while maintaining adequate randomness to evaluate the methods. To ensure a fair comparison, the learning rate for all methods is set to 0.02, and the parameters are initialized identically across all methods. Each configuration is independently run 50 times, and the learning curves presented in the following sections are averaged over these runs with standard deviation shading. All numerics are conducted on classical simulators, where both expectation values and metric tensors can be tracked precisely.

The learning curves of three methods for the 10-, 12-, and 14-qubit Ising and Heisenberg models with a single-layer circuit are shown in figures 5 and 6, respectively. Across all instances, WA-QNG exhibits a markedly faster convergence than both QNG and vanilla gradient descent, with the latter being the slowest and failing to converge within 500 optimization steps. These findings suggest that WA-QNG possesses a potential advantage in convergence speed compared to the other two methods. In the Ising model cases, WA-QNG attains an explicitly better average final convergence value than QNG, while in

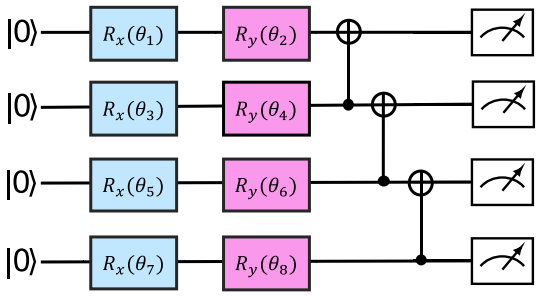

Figure 4. An example of a 4-qubit EfficientSU2 circuit. It consists of single-qubit rotation gates R_x and R_y, followed by a series of CNOT gates to enhance entanglement.

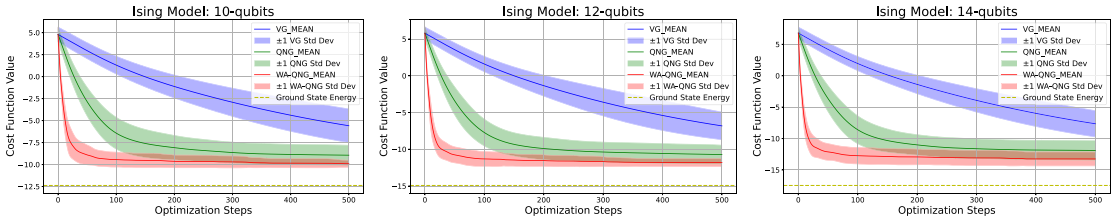

Figure 5. The learning curves of the three methods for Ising model of 10, 12, 14 qubits on 1-layer EfficientSU2 circuit.

the Heisenberg model cases, the two methods yield almost the same final average convergence values, with WA-QNG still maintaining a slight advantage. This indicates that WA-QNG can provide superior, or at least competitive, performance compared to standard QNG in terms of the ability to escape local minima.

One may note that all three methods fail to converge to the true ground-state energy. This can be attributed to two reasons. First, the EfficientSU2 circuit used in the numerics is not guaranteed to include the ground state within its expressible space. Second, both WA-QNG and standard QNG may become trapped in local minima during the optimization process. Hence, the numerics in this work mainly aim to compare the relative performance between WA-QNG and standard QNG as gradient-based local search optimization methods. The absolute performance, related to global convergence to the true minimum, remains a great challenge in the optimization field.

The learning curves of the three methods for the 10-qubit Ising and Heisenberg models with multi-layer circuits are shown in figures 7 and 8, respectively. The results indicate that the potential advantages discussed in the single-layer case still hold when the number of layers increases. Furthermore, both WA-QNG and standard QNG reach a sightly better average convergence value compared to the single-layer circuit—although increasing the number of layers makes the optimization landscape more complex, it also enlarges the expressive space of the variational circuit, which may contain better solutions with lower cost function values. This result also suggests that WA-QNG can still perform well in a more complicated landscape.

In addition, we also provide numerical simulations for the four-qubit Ising and Heisenberg models using a four-layer EfficientSU2 circuit, for which the ground state lies within the search space, as a supplement to the numerics discussed above. As shown in figure 9, both WA-QNG and QNG converge to the ground-state energy in this setting, where the circuit is sufficiently expressive, and WA-QNG still exhibits a faster convergence rate. In contrast, vanilla gradient descent remains trapped in a local minimum and fails to reach the ground state. This result further highlights the importance of the choice of optimizer in VQE: even when the ansatz is sufficiently expressive, a suboptimal optimizer may lead the optimization process to local minima and thus limit overall performance.

6.2. Weights of subsystems

To better understand how accounting for the different weights of each subsystem in WA-QNG plays a central role in improving optimization performance compared to standard QNG, we conduct numerics

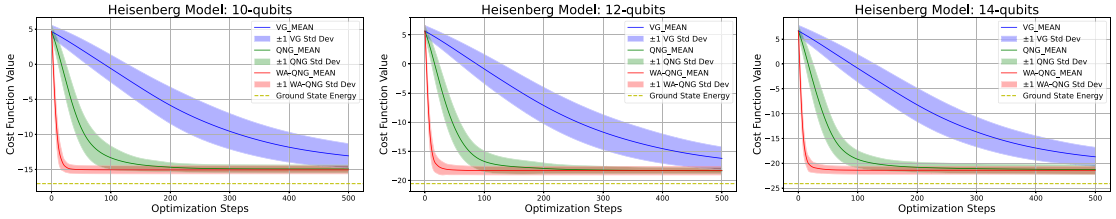

Figure 6. The learning curves of the three methods for Heisenberg model of 10, 12, 14 qubits on 1-layer EfficientSU2 circuit.

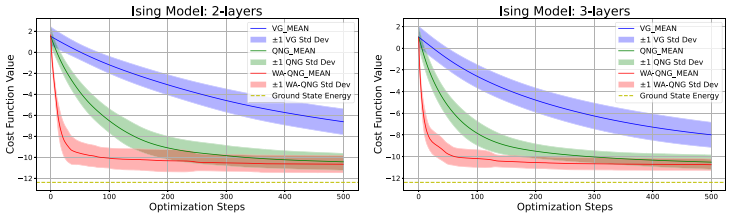

Figure 7. The learning curves of the three methods for Ising model of 10 qubits on 2, 3-layer EfficientSU2 circuit.

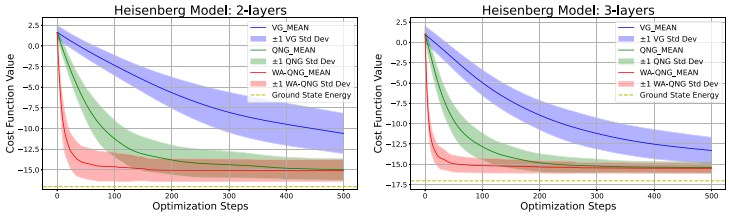

Figure 8. The learning curves of the three methods for Heisenberg model of 10 qubits on 2, 3-layer EfficientSU2 circuit.

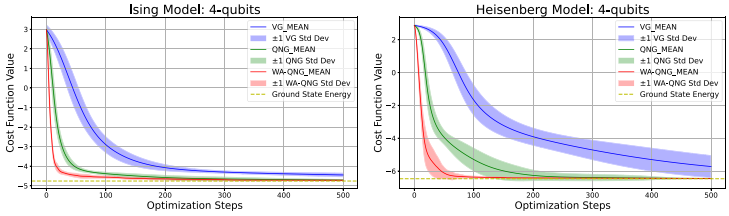

Figure 9. The learning curves of the three methods for Ising and Heisenberg model of 4 qubits on 4-layer EfficientSU2 circuit.

using the following 3-qubit toy Ising model, with α set to 0.8, 0.6, 0.4 and 0.2:

$$H = Z_1Z_2 + Z_2Z_3 + \alpha X_1 + (3 - 2\alpha)X_2 + \alpha X_3. \tag{25}$$

As α decreases from 0.8 to 0.2, the weights of the subsystems become increasingly unbalanced, with the contribution of the subsystem associated with the second qubit to the output growing larger. Therefore, if incorporating subsystem weights into the optimization is really effective, WA-QNG is expected to exhibit increasingly better performance compared to QNG as α decreases.

To intuitively quantify the performance gap between WA-QNG and standard QNG, we use the difference in cost function values at the same optimization step on the learning curves as an indicator. For a fair comparison, this difference is normalized by dividing it by the difference between the initial and converged cost function value. The cost function value gap curve during training and the discrete area under the gap curve (also representing the discrete area between the learning curves of WA-QNG and QNG, namely the AUC) are presented in figure 10.

As α decreases from 0.8 to 0.2, the contribution of the subsystem on the second qubit increases. The numerical results align with the theoretical analysis, as the discrete AUC indeed increases with decreasing

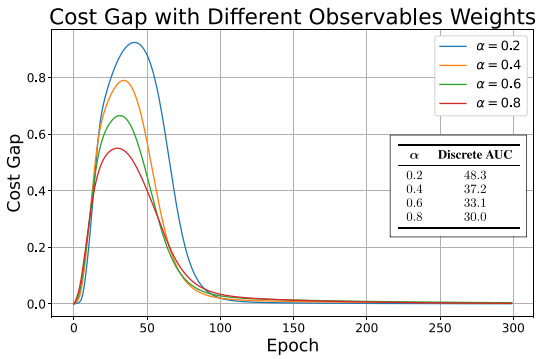

Figure 10. The cost function value gap curves and corresponding discrete AUC between WA-QNG and QNG under different values of α for the Hamiltonian $H = Z_1Z_2 + Z_2Z_3 + \alpha X_1 + (3 - 2\alpha)X_2 + \alpha X_3$. A smaller α increases the weight of the subsystem on the second qubit in the output, leading to a scenario where WA-QNG should outperform QNG more significantly. This figure indicates that the theoretical analysis aligns with the numerical results that WA-QNG indeed achieves greater performance improvement over QNG as α gradually decreases.

α. It indicates that capturing the different weights and contributions of each subsystem in WA-QNG effectively improves optimization performance compared to standard QNG. This result implies that WA-QNG is more suitable for situations where the coefficients of each observable term in the k-local Hamiltonian vary significantly and are unbalanced.

6.3. Locality of Hamiltonian

As mentioned in section 3.1, when the entire system becomes significantly larger than the subsystems that directly contribute to the output, the sensitivity of each parameter in the total system differs considerably from that of each subsystem. Under this condition, WA-QNG is expected to outperform standard QNG to a greater extent. To gain a clearer understanding, we conduct numerics using an n-qubit toy Ising model Hamiltonian, with n varying from 2 to 5, while fixing the maximum locality of the Hamiltonian terms to be 2:

$$H = \sum_{i=1}^{n-1} Z_i Z_{i+1} + \sum_{i=1}^{n} X_i \tag{26}$$

As n increases, the ratio between the locality of Hamiltonian term and the global system size becomes smaller, and the sensitivity of each parameter in each subsystem differs more significantly from that of the entire system, as each observable term in H is at most 2-local. Consequently, WA-QNG is expected to perform increasingly better as n increases compared to standard QNG. Similar to the previous subsection, the cost function value gap curve during training and the discrete AUC are shown in figure 11. For a fair comparison, this difference is also normalized by dividing it by the difference between the initial and converged cost function value.

The numerical results agree with the theoretical analysis. As n increases, the discrete AUC also increases, indicating a more significant performance improvement for WA-QNG. This suggests that WA-QNG is particularly well-suited for scenarios where the total system size n is much larger than the locality factor k for a k-local Hamiltonian. QNG is a special case of WA-QNG when $k = n$.

7. Discussion and conclusion

In this work, we mainly introduce the WA-QNG, which accounts for the different weights and contributions of each subsystem in the optimization process. In particular, we propose using the matrix $W = \frac{2}{\sum_m h_m^2} \sum_m h_m^2 T_m$ instead of the quantum Fisher information matrix of the entire system in each optimization step. We provide three perspectives to explain the effectiveness and potential advantages of WA-QNG. Firstly, the matrix W in WA-QNG can be interpreted as an approximation of the weighted average of the quantum Fisher information matrix of each subsystem contributing to the output. Secondly, from an optimization view, we illustrate that WA-QNG can be derived from a constrained optimization problem where the Euclidean distance in the parameter space is replaced by a weighted sum over the 2-norm distances between density matrices. We further explain that WA-QNG can also be derived from an information geometric perspective, where it emerges as a pullback metric. Additionally, we demonstrate

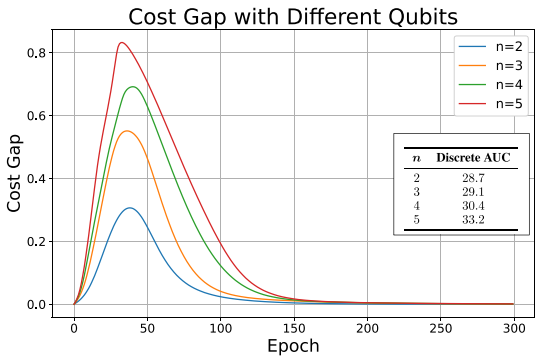

Figure 11. The cost function value gap curves and corresponding discrete AUC between WA-QNG and QNG under different values of n for the Hamiltonian $H = \sum_{i=1}^{n-1} Z_i Z_{i+1} + \sum_{i=1}^{n} X_i$. A larger n decreases the ratio between the locality of Hamiltonian term and the global system size becomes smaller, and increases the difference in parameter sensitivity of each subsystem compared to that of the entire system, leading to a scenario where WA-QNG should outperform QNG more significantly. This figure shows that the theoretical analysis aligns with the numerical results, indicating that WA-QNG achieves greater performance improvement over QNG as n increases with a fixed locality factor k.

that the optimization task can be approximately transformed into a non-linear least squares problem, where WA-QNG is equivalent to the Gauss–Newton method.

To evaluate the performance of WA-QNG, we conduct numerical simulations on the variational eigensolver for the Ising model and Heisenberg model. The results indicate that WA-QNG achieves superior optimization performance compared to standard QNG. Additionally, we perform further numerics to investigate the source of WA-QNG's advantage. Our findings indicate two key factors. The first is accounting for the weights of each subsystem indeed improves optimization performance. The second is the Hilbert–Schmidt metric tensor for each subsystem provides a better representation of parameter sensitivity within subsystems compared to using the quantum Fisher information matrix of the entire system. The numerical results are consistent with the theoretical analysis.

In sections 4 and 3.3, we mention that WA-QNG has a stronger theoretical explanation from both its approximation to quantum Fisher information matrix and its connection with Gauss–Newton method when each subsystem state is close to being pure and does not change significantly with parameters. However, this assumption is not necessary for WA-QNG to perform better than standard QNG. In the numerics, we observe that WA-QNG still outperforms standard QNG, even when the final target is an entanglement state where each subsystem is mixed.

Since the focus of this paper is to introduce the novelty of WA-QNG itself, we track the exact expectation values and metric tensors of the quantum circuit. In practical applications, these quantities can only be estimated through finite shots, which introduces shot noise. Additionally, we assume that the circuit is noise-free, which is not the case in real-world implementations. Investigating the effects of shot noise and circuit noise on WA-QNG is left as a potential direction for future work.

This work primarily focuses on optimization for the quantum eigensolver. However, since QNG can also be applied to optimize other variational quantum algorithms such as the variational quantum classifier in the field of machine learning, we believe WA-QNG is expected to be extendable to these tasks similarly. Given the formulation of WA-QNG, any variational algorithm employing a cost function defined by the expected value of a k-local Hamiltonian could potentially benefit from WA-QNG beyond ground state energy problems. While evaluating the performance of WA-QNG in these broader applications is beyond the scope of this work, it remains a promising direction for future research.

In conclusion, WA-QNG offers a promising and efficient optimization method for VQEs. By accounting for the weights of each subsystem that contributes to the output, WA-QNG presents a potential research direction for optimization in variational quantum algorithms.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Funding

All authors acknowledge the support from the Dutch National Growth Fund (NGF), as part of the Quantum Delta NL programme. V D acknowledges support from the Dutch Research Council

(NWO/OCW), as part of the Quantum Software Consortium programme (Project Number 024.003.03). This project was also co-funded by the European Union (ERC CoG, BeMAIQuantum, 10 112 4342).

Author contributions

This project was conceived by C.S. and H.W.. C.S. and H.W. formulated the theoretical part. V.D. helped C.S. for the design of numerical simulations. C.S. conducted numerical simulations and analyzed the results with assistance from V.D. and H.W. All authors reviewed the paper on both theoretical and numerical parts.

Appendix A. QNG as a special case of WA-QNG

When a term H_m in the Hamiltonian acts globally on the entire quantum system, the corresponding nontrivial subsystem ρ_m becomes the full system ρ. In this scenario, each T_m no longer depends on the index m, causing the terms $\sum_m h_m^2$ in the weighted summation to cancels out with that in the prefactor. As a consequence, the matrix W exactly reduces to the Hilbert–Schmidt metric tensor of the whole quantum system ρ_θ with a constant factor 2, where the i-th row and j-th column element of the matrix W is as follows:

$$W_{ij} = 2\mathrm{tr}\left(\partial_i \rho_\theta \partial_j \rho_\theta\right) \tag{27}$$

Now we prove the matrix W is equal to the quantum Fisher information matrix F. The variational state ρ_θ on the whole system is a pure state, so we can write the state ρ_θ as $\rho_\theta = |\phi\rangle\langle\phi|$. So we only have to prove the right side of equation (27) is equal to that of equation (6):

$$\begin{aligned}
W_{ij} &= 2\mathrm{tr}\left(\partial_i \rho \partial_j \rho\right) \\
&= 2\mathrm{tr}\left(\partial_i \left(|\phi\rangle\langle\phi|\right) \partial_j \left(|\phi\rangle\langle\phi|\right)\right) \\
&= 2\mathrm{tr}\left(\left(|\partial_i\phi\rangle\langle\phi| + |\phi\rangle\langle\partial_i\phi|\right)\left(|\partial_j\phi\rangle\langle\phi| + |\phi\rangle\langle\partial_j\phi|\right)\right) \\
&= 2\mathrm{tr}\left(|\partial_i\phi\rangle\langle\phi|\partial_j\phi\rangle\langle\phi| + |\partial_i\phi\rangle\langle\partial_j\phi| + |\phi\rangle\langle\partial_i\phi|\partial_j\phi\rangle\langle\phi| + |\phi\rangle\langle\partial_i\phi|\phi\rangle\langle\partial_j\phi|\right) \\
&= 2\langle\phi|\partial_j\phi\rangle\langle\phi|\partial_i\phi\rangle + 2\langle\partial_i\phi|\phi\rangle\langle\partial_j\phi|\phi\rangle + 2\langle\partial_j\phi|\partial_i\phi\rangle + 2\langle\partial_i\phi|\partial_j\phi\rangle \\
&= 2\langle\partial_j\phi|\partial_i\phi\rangle + 2\langle\partial_i\phi|\partial_j\phi\rangle - 2\langle\phi|\partial_j\phi\rangle\langle\partial_i\phi|\phi\rangle - 2\langle\phi|\partial_i\phi\rangle\langle\partial_j\phi|\phi\rangle \\
&= 4\,\mathrm{Re}\left(\langle\partial_i\phi|\partial_j\phi\rangle - \langle\partial_i\phi|\phi\rangle\langle\phi|\partial_j\phi\rangle\right) \\
&= F_{ij}
\end{aligned} \tag{28}$$

Appendix B. Estimate Hilbert–Schmidt metric tensor via classical shadows

In this section, we demonstrate that the Hilbert–Schmidt metric tensor T [5] used in WA-QNG can be efficiently estimated by classical shadow and bound the shots required to obtain the element T_{ij}. To avoid confusion in the derivation, we will use calligraphic font $\mathcal{T}$ to represent the Hilbert–Schmidt metric tensor in the following section.

B.1. Parameter-shift rule

First, the parameter-shift rule can be applied to compute each element of the matrix $\mathcal{T}$:

$$\begin{aligned}
\mathcal{T}_{ij} &= 2\mathrm{tr}\left(\partial_i \rho_\theta \partial_j \rho_\theta\right) \\
&= \frac{1}{2}\mathrm{tr}\left(\left(\rho_{\theta+\frac{\pi}{2}e_i} - \rho_{\theta-\frac{\pi}{2}e_i}\right)\left(\rho_{\theta+\frac{\pi}{2}e_j} - \rho_{\theta-\frac{\pi}{2}e_j}\right)\right) \\
&= \frac{1}{2}\left(\mathrm{tr}\left(\rho_{\theta+\frac{\pi}{2}e_i}\rho_{\theta+\frac{\pi}{2}e_j}\right) - \mathrm{tr}\left(\rho_{\theta+\frac{\pi}{2}e_i}\rho_{\theta-\frac{\pi}{2}e_j}\right) - \mathrm{tr}\left(\rho_{\theta-\frac{\pi}{2}e_i}\rho_{\theta+\frac{\pi}{2}e_j}\right) + \mathrm{tr}\left(\rho_{\theta-\frac{\pi}{2}e_i}\rho_{\theta-\frac{\pi}{2}e_j}\right)\right)
\end{aligned} \tag{29}$$

where e_i represents the unit vector with the i-th element set to one and all other elements set to zero. To estimate $\mathcal{T}_{ij}$, we only need to estimate the four terms in equation (29) respectively. To estimate the entire matrix $\mathcal{T}$, we can estimate each element individually, meaning the total cost scales quadratically with the number of parameters. Thus, if we can bound the cost of estimating the term like $\mathrm{tr}(\rho\sigma)$, we can also bound the total cost. In our case, where the Hamiltonian is k-local, we show the cost of estimating such term $\mathrm{tr}(\rho\sigma)$ via classical shadow is exponential to the subsystem size k rather than the size of the whole system n.

[5] Here for notation simplify, we omit the subscript for each T_m in the definition of the matrix W in 3.2.

B.2. Classical shadow

The classical shadow technique constructs a series of unbiased estimators $\hat{\rho}^{(t)}$ $(1 \leqslant t \leqslant T$, where T is the total number of the classical shadows constructed) for a state ρ, with the property that $\mathrm{E}[\hat{\rho}^{(t)}] = \rho$. Each $\hat{\rho}^{(t)}$ is represented as:

$$\hat{\rho}^{(t)} = \bigotimes_{i=1}^{n} \left(3U_i^{\dagger}|\hat{b}_i\rangle\langle\hat{b}_i|U_i - \mathbb{I}\right) \tag{30}$$

where n is the system size, b is a binary string obtained by measurements, and b_i represents the i-th bit of b (either 0 or 1). U denotes the corresponding random Pauli gate applied to the i-th qubit. For more details on the data acquisition process in the classical shadow technique, please refer to [23] and [14]. Two important properties of each estimator $\hat{\rho}^{(t)}$ are as fellows:

$$\mathrm{E}\left[\mathrm{tr}\left(\hat{\rho}^{(t)}O\right)\right] = \mathrm{tr}\left(\rho O\right) \tag{31}$$

$$\mathrm{Var}\left[\mathrm{tr}\left(\hat{\rho}^{(t)}O\right)\right] \leqslant 2^{w(O)}\mathrm{tr}\left(O^2\right) \tag{32}$$

where $w(O)$ represents the number of qubits on which the observable O acts nontrivially. For the details of derivation of equations (31) and (32), please refer to the paper [36].

To reduce error, an empirical average is taken over all samples to construct the estimator $\hat{\rho}$:

$$\hat{\rho} = \frac{1}{T}\sum_{i}^{T}\hat{\rho}^{(t)}. \tag{33}$$

From equations (31) and (32), the following properties of the estimator $\hat{\rho}$ can be derived:

$$\mathrm{E}\left[\mathrm{tr}\left(\hat{\rho}O\right)\right] = \mathrm{tr}\left(\rho O\right) \tag{34}$$

$$\mathrm{Var}\left[\mathrm{tr}\left(\hat{\rho}O\right)\right] \leqslant \frac{2^{w(O)}\mathrm{tr}\left(O^2\right)}{T}. \tag{35}$$

B.3. Construct estimator for Hilbert–Schmidt metric tensor

As discussed in appendix B.1, estimating the Hilbert–Schmidt metric tensor via the classical shadow technique requires constructing an estimator for terms like $\mathrm{tr}(\rho\sigma)$. Similar to the estimator used for estimating purity in [23] and [36], the following estimator can be constructed for the term like $\mathrm{tr}(\rho\sigma)$. For simplicity, we denote $p = \mathrm{tr}(\rho\sigma)$, then the corresponding estimator $\hat{p}$ is:

$$\hat{p} = \frac{1}{T^2}\sum_{ij}\mathrm{tr}\left(\hat{\rho}^{(i)}\hat{\sigma}^{(j)}\right) \tag{36}$$

where each $\hat{\rho}^{(i)}$ and $\hat{\sigma}^{(j)}$ $(1 \leqslant i,j \leqslant T)$ is obtained using the classical shadow technique as described in equation (30). Because $\hat{\rho}^{(i)}$ and $\hat{\sigma}^{(j)}$ are independent, we have:

$$\begin{aligned}
\mathrm{E}\left[\hat{p}\right] &= \mathrm{E}\left[\frac{1}{T^2}\sum_{ij}\mathrm{tr}\left(\hat{\rho}^{(i)}\hat{\sigma}^{(j)}\right)\right] \\
&= \frac{1}{T^2}\sum_{ij}\mathrm{E}\left[\mathrm{tr}\left(\hat{\rho}^{(i)}\hat{\sigma}^{(j)}\right)\right] \\
&= \frac{1}{T^2}\sum_{ij}\mathrm{tr}\left(\mathrm{E}\left[\hat{\rho}^{(i)}\right]\mathrm{E}\left[\hat{\sigma}^{(j)}\right]\right) \\
&= \mathrm{tr}\left(\rho\sigma\right) \\
&= p.
\end{aligned} \tag{37}$$

Hence, the estimator $\hat{p}$ is also an unbiased estimator for p. To bound the computational cost, we also need to bound the variance of the estimator $\hat{p}$.

B.4. Bounding variance

According to the definition of the variance of a random variable, we have: $\hat{p}$.

$$\begin{aligned}\mathrm{Var}[\hat{p}] &= \mathrm{E}\left[(\hat{p}-p)^2\right] \\ &= \mathrm{E}\left[\left(\frac{1}{T^2}\sum_{ij}\left(\mathrm{tr}\left(\hat{\rho}^{(i)}\hat{\sigma}^{(j)}\right)-\mathrm{tr}(\rho\sigma)\right)\right)^2\right] \\ &= \frac{1}{T^4}\sum_{ij}\sum_{kl}\mathrm{E}\left[\left(\mathrm{tr}\left(\hat{\rho}^{(i)}\hat{\sigma}^{(k)}\right)-\mathrm{tr}(\rho\sigma)\right)\left(\mathrm{tr}\left(\hat{\rho}^{(j)}\hat{\sigma}^{(l)}\right)-\mathrm{tr}(\rho\sigma)\right)\right]. \end{aligned} \tag{38}$$

The summation in equation (38) can be divided into the following three cases. Suppose the density operator ρ and σ are systems of n-qubit.

1. When $i\neq k$ and $j\neq l$: There are $T^2(T-1)^2$ terms. For each term, we have:

$$\begin{aligned}&\mathrm{E}\left[\left(\mathrm{tr}(\hat{\rho}^{(i)}\hat{\sigma}^{(j)}-\mathrm{tr}(\rho\sigma)\right)\left(\mathrm{tr}(\hat{\rho}^{(k)}\hat{\sigma}^{(l)})-\mathrm{tr}(\rho\sigma)\right)\right] \\ &= \left(\mathrm{E}\left[\mathrm{tr}(\hat{\rho}^{(i)}\hat{\sigma}^{(j)})\right]-\mathrm{tr}(\rho\sigma)\right)\left(\mathrm{E}\left[\mathrm{tr}(\hat{\rho}^{(j)}\hat{\sigma}^{(k)})\right]-\mathrm{tr}(\rho\sigma)\right) \\ &= 0. \end{aligned} \tag{39}$$

2. When $i=k$ but $j\neq l$, or $j=l$ but $i\neq k$: There are $2T^2(T-1)$ terms. Without loss of generality, we take the case where $i=k$ but $j\neq l$ as an example. The calculation in another case is the same. For each term, we have:

$$\begin{aligned}&\mathrm{E}\left[\left(\mathrm{tr}(\hat{\rho}^{(i)}\hat{\sigma}^{(j)})-\mathrm{tr}(\rho\sigma)\right)\left(\mathrm{tr}(\hat{\rho}^{(i)}\hat{\sigma}^{(l)})-\mathrm{tr}(\rho\sigma)\right)\right] \\ &= \mathrm{E}\left[\mathrm{tr}(\hat{\rho}^{(i)}\hat{\sigma}^{(j)})\mathrm{tr}(\hat{\rho}^{(i)}\hat{\sigma}^{(l)})\right]-\mathrm{tr}^2(\rho\sigma) \\ &= \mathrm{E}\left[\mathrm{tr}(\hat{\rho}^{(i)\otimes 2}\hat{\sigma}^{(j)}\otimes\hat{\sigma}^{(l)})\right]-\mathrm{tr}^2(\rho\sigma) \\ &= \mathrm{E}\left[\mathrm{tr}^2(\hat{\rho}^{(i)}\sigma)\right]-\mathrm{tr}^2(\rho\sigma) \\ &= \mathrm{Var}\left[\mathrm{tr}(\hat{\rho}^{(i)}\sigma)\right] \\ &\leqslant 2^{w(\sigma)}\mathrm{tr}(\sigma^2) \\ &\leqslant 2^n. \end{aligned} \tag{40}$$

The third equality in equation (40) relies on the property that, when the unbiased estimators $\hat{\rho}$ and $\hat{\sigma}$ are independent, then $\mathrm{E}[\hat{\rho}\otimes\hat{\sigma}]=\rho\otimes\sigma$. The details of this property can be found in [36]. The first inequality is from equation (32).

3. When $i=k$ and $j=l$: There are T^2 terms. For each term, we have:

$$\begin{aligned}&\mathrm{E}\left[\left(\mathrm{tr}(\hat{\rho}^{(i)}\hat{\sigma}^{(j)})-\mathrm{tr}(\rho\sigma)\right)\left(\mathrm{tr}(\hat{\rho}^{(i)}\hat{\sigma}^{(j)})-\mathrm{tr}(\rho\sigma)\right)\right] \\ &= \mathrm{E}\left[\mathrm{tr}^2(\hat{\rho}^{(i)}\hat{\sigma}^{(j)})\right]-\mathrm{tr}^2(\rho\sigma) \\ &= \mathrm{Var}\left[\mathrm{tr}(\hat{\rho}^{(i)}\hat{\sigma}^{(j)}))\right] \\ &= \mathrm{Var}\left[\mathrm{tr}(S\hat{\rho}^{(i)}\otimes\hat{\sigma}^{(j)}))\right] \\ &\leqslant 2^{w(S)}\mathrm{tr}(S^2) \\ &= 2^{4n}. \end{aligned} \tag{41}$$

The third equality follows from the property $\mathrm{tr}(S\rho\otimes\sigma)=\mathrm{tr}(\rho\sigma)$, where $S\in\mathbb{C}^{2^{2n}\times 2^{2n}}$ is the *SWAP* operator. The details of this property can be found in [37]. The inequality arises from equation (32). The last equality holds because S acts on 2^{2n} qubits and satisfies $S^2=I$.

Hence, the final result of equation (38) shall be upper bounded by:

$$\mathrm{Var}[\hat{p}] = \frac{1}{T^4}\sum_{ij}\sum_{kl}\mathrm{E}\left[\left(\mathrm{tr}\left(\hat{\rho}^{(i)}\hat{\sigma}^{(k)}\right)-\mathrm{tr}(\rho\sigma)\right)\left(\mathrm{tr}\left(\hat{\rho}^{(j)}\hat{\sigma}^{(l)}\right)-\mathrm{tr}(\rho\sigma)\right)\right]$$

$$\leqslant \frac{1}{T^4}\left(2T^2(T-1)2^n + T^2 2^{4n}\right)$$
$$\leqslant \frac{2^{n+1}}{T} + \frac{2^{4n}}{T^2}. \tag{42}$$

B.5. Bounding shots

Using the bound of variance equation (42), we can derive the upper bound of total shots required in the estimation for the term like $\mathrm{tr}(\rho\sigma)$. Following the assumption in the reference paper [14] that, the T can be very large where the expression of equation (42) is dominated by the first term. Following Chebyshev's inequality:

$$\Pr\left[|\hat{p} - p| \geqslant \epsilon\right] \lesssim \frac{2^{n+1}}{T\epsilon^2}. \tag{43}$$

Then, a measurement budget that scales as

$$T \geqslant \frac{2^{n+1}}{\epsilon^2 \delta} \tag{44}$$

with probability $1-\delta$ suffice to control the estimation error below ϵ. Hence, the lower bound of shots required to estimate a term like $\mathrm{tr}(\rho\sigma)$ is $\mathrm{O}(2^{n+1})$, where n is the system size of the quantum density operator ρ and σ.

In our case, there are four terms like $\mathrm{tr}(\rho\sigma)$ in the equation (29) required to estimate for the element $\mathcal{T}_{ij}$, and each term is only k-local. Hence the cost of shots required to estimate each element $\mathcal{T}_{ij}$ is $\mathrm{O}(4\cdot 2^{k+1})$, which is exponential to the subsystem size k instead of the entire system size n.

Appendix C. Approximate quantum Fisher information via Hilbert–Schmidt metric tensor

In this section, we provide a simple proof that the Hilbert–Schmidt metric tensor serves as an approximation to the quantum Fisher information matrix F when the state ρ is close to being pure and does not change significantly with parameters. Moreover, this approximation becomes exact when ρ is pure.

For a state $\rho = \sum_{k=1}^{n} r_k |r_k\rangle\langle r_k|$, suppose the dominant eigenvector [29, 30] is $|r_\mathrm{d}\rangle$ with eigenvalue r_d. When state ρ is close to being pure, the dominant eigenvalue satisfies $r_\mathrm{d} \approx 1$, while all other eigenvalues satisfy $r_k \approx 0$. For the state ρ, we show its Hilbert–Schmidt metric tensor can be computed and simplified as:

$$\begin{aligned}\mathrm{tr}\left(\partial_i\rho\partial_j\rho\right) &= \sum_k \frac{\partial r_k}{\partial\theta_i}\cdot\frac{\partial r_k}{\partial\theta_j} + \sum_k r_k^2\frac{F_k}{2} - \sum_{kl,k\neq l} r_k r_l \cdot 2\,\mathrm{Re}\left[\langle\partial_i r_k|r_l\rangle\langle r_k|\partial_j r_l\rangle\right]\\ &\approx \frac{r_\mathrm{d}^2}{2}\left(F_\mathrm{d}\right)_{ij}\\ &\approx \frac{1}{2}F_{ij}\end{aligned} \tag{45}$$

where $(F_\mathrm{d})_{ij}$ represents the (i,j)-th element of the quantum Fisher information matrix of the dominant state. The derivation of the first equation will be explained in detail in the following discussion. When the state ρ is close to being pure, the other eigenvalues are higher-order infinitesimals compared to the dominant eigenvalues. Consequently, the second and third terms in the first equation can be approximated as $\frac{r_\mathrm{d}^2}{2}(F_\mathrm{d})_{ij}$. If the state does not change drastically with respect to the parameters, the first term will also be small. In practical applications, this can be achieved by initializing the variational quantum circuit with low entanglement and setting the learning rate to a small value. Moreover, because $r_\mathrm{d} \approx 1$ and the state $|r_\mathrm{d}\rangle$ dominants ρ, we can obtain $\frac{r_\mathrm{d}^2}{2}(F_\mathrm{d})_{ij} \approx \frac{1}{2}F_{ij}$. However, this approximation assumption is not necessary for WA-QNG to be well defined or to outperform standard QNG.

Now, we provide a brief derivation of the first equation. For simplicity of notation, we denote the three terms in $\partial_i\rho = \sum_k \partial_i r_k|r_k\rangle\langle r_k| + \sum_k r_k|\partial_i r_k\rangle\langle r_k| + \sum_k r_k|r_k\rangle\langle\partial_i r_k|$ as A_i, B_i, and C_i respectively. Similarly, for $\partial_j\rho$ we can also denote A_j, B_j and C_j for the three terms analogously. Hence, we can express:

$$\begin{aligned}&\mathrm{tr}\left(\partial_i\rho\partial_j\rho\right)\\ &= \mathrm{tr}\left(A_iA_j\right) + \mathrm{tr}\left(A_iB_j\right) + \mathrm{tr}\left(A_iC_j\right) + \mathrm{tr}\left(B_iA_j\right) + \mathrm{tr}\left(B_iB_j\right) + \mathrm{tr}\left(B_iC_j\right) + \mathrm{tr}\left(C_iA_j\right) + \mathrm{tr}\left(C_iB_j\right) + \mathrm{tr}\left(C_iC_j\right).\end{aligned} \tag{46}$$

We can compute these terms separately.
1. $\mathrm{tr}(A_iA_j)$:

$$\begin{aligned}\mathrm{tr}\left(A_iA_j\right) &= \mathrm{tr}\left(\sum_k \frac{\partial r_k}{\partial\theta_i}\frac{\partial r_k}{\partial\theta_j}|r_k\rangle\langle r_k|\right)\\ &= \sum_k \frac{\partial r_k}{\partial\theta_i}\frac{\partial r_k}{\partial\theta_j}\end{aligned} \tag{47}$$

2. $\mathrm{tr}(A_iB_j)$ and $\mathrm{tr}(A_iC_j)$:

$$\begin{aligned}\mathrm{tr}\left(A_iC_j\right)+\mathrm{tr}\left(A_iB_j\right) &= \sum_k \frac{\partial r_k}{\partial\theta_i} r_k\langle\partial_j r_k|r_k\rangle + \sum_k \frac{\partial r_k}{\partial\theta_i} r_k\langle r_k|\partial_j r_k\rangle\\ &= 0\end{aligned} \tag{48}$$

3. $\mathrm{tr}(B_iC_j)$ and $\mathrm{tr}(C_iB_j)$:

$$\begin{aligned}\mathrm{tr}\left(B_iC_j\right)+\mathrm{tr}\left(C_iB_j\right) &= \sum_k r_k^2\langle\partial_j r_k|\partial_i r_k\rangle + \sum_k r_k^2\langle\partial_i r_k|\partial_j r_k\rangle\\ &= \sum_k 2r_k^2\,\mathrm{Re}\left[\langle\partial_i r_k|\partial_j r_k\rangle\right]\end{aligned} \tag{49}$$

4. $\mathrm{tr}(B_iB_j)$ and $\mathrm{tr}(C_iC_j)$:

$$\begin{aligned}\mathrm{tr}\left(B_iB_j\right)+\mathrm{tr}\left(C_iC_j\right) &= \sum_{kl} r_k r_l\left(\langle r_k|\partial_j r_l\rangle\langle r_l|\partial_i r_k\rangle + \langle\partial_i r_k|r_l\rangle\langle\partial_i r_k|r_l\rangle\langle\partial_j r_l|r_k\rangle\right)\\ &= -\sum_k r_k^2\cdot 2\,\mathrm{Re}\left[\langle\partial_i r_k|r_k\rangle\langle r_k|\partial_j r_k\rangle\right] - \sum_{kl,k\neq l} r_k r_l\cdot 2\,\mathrm{Re}\left[\langle\partial_i r_k|r_l\rangle\langle r_k|\partial_j r_l\rangle\right]\end{aligned} \tag{50}$$

5. $\mathrm{tr}(C_iA_j)$ and $\mathrm{tr}(B_iA_j)$:

$$\begin{aligned}\mathrm{tr}\left(C_iA_j\right)+\mathrm{tr}\left(B_iA_j\right) &= r_k^2\langle r_k|\partial_i r_k\rangle + r_k^2\langle\partial_i r_k|r_k\rangle.\\ &= 0\end{aligned} \tag{51}$$

By combining these results, we obtain the first equation in equation (45).

Appendix D. Deriving WA-QNG optimization step from geometric interpretation

According to the Lagrange multiplier method, the constrained optimization problem in equation (12) can be formulated as:

$$d^* = \underset{d}{\mathrm{argmin}} f(\theta+d) + \lambda\left(\frac{2}{\sum_m h_m^2}\sum_m h_m^2\|\rho_m(\theta+d)-\rho_m(\theta)\|_2^2 - \epsilon\right). \tag{52}$$

For the trace 2-norm term, applying the first-order Taylor expansion to $\rho_m(\theta+d)$, we obtain:

$$\begin{aligned}\|\rho_m(\theta+d)-\rho_m(\theta)\|_2^2 &\approx \|\rho_m(\theta)+\sum_i \partial_i\rho_m(\theta)\,d_i-\rho_m(\theta)\|_2^2\\ &= \|\sum_i \partial_i\rho_m(\theta)\,d_i\|_2^2\\ &= \mathrm{tr}\left(\sum_i\sum_j \partial_i\rho_m\partial_j\rho_m d_i d_j\right)\\ &= \sum_i\sum_j \mathrm{tr}\left(\partial_i\rho_m\partial_j\rho_m\right) d_i d_j.\end{aligned} \tag{53}$$

Substituting equations (53) into (52) and applying the first-order Taylor expansion to $f(\theta+d)$, we obtain:

$$\begin{aligned} d^* &\approx \underset{d}{\operatorname{argmin}} f(\theta) + \nabla f(\theta)^T d + \frac{2\lambda}{\sum_m h_m^2} \sum_m h_m^2 \sum_i \sum_j \operatorname{tr}\left(\partial_i \rho_m \partial_j \rho_m\right) d_i d_j - \lambda\epsilon \\ &= \underset{d}{\operatorname{argmin}} f(\theta) + \nabla f(\theta)^T d + \sum_i \sum_j \left(\frac{2\lambda}{\sum_m h_m^2} \sum_m h_m^2 \operatorname{tr}\left(\partial_i \rho_m \partial_j \rho_m\right) \right) d_i d_j - \lambda\epsilon \\ &= \underset{d}{\operatorname{argmin}} f(\theta) + \nabla f(\theta)^T d + \lambda d^T W d - \lambda\epsilon \end{aligned} \tag{54}$$

where the matrix W is exactly the same matrix defined in WA-QNG in equation (7). Since we are computing the minimum, equation (54) should satisfy the Karush–Kuhn–Tucker conditions [38]. Here, it simply means that the derivative of the right side with respect to d should be zero:

$$\begin{aligned} 0 &= \nabla f(\theta) + 2\lambda W d \\ d &= -\frac{1}{2\lambda} W^{+} \nabla f(\theta). \end{aligned} \tag{55}$$

Equation (55) indicates the optimal update direction in WA-QNG. Since the Lagrange multiplier λ can be absorbed into the learning rate, the above formula can be exactly transformed into the update formula of WA-QNG, as given in equation (7).

Appendix E. Relation with Gauss–Newton method supplement

Here we verify the relation $W = 2J_{\mathrm{r}}^T J_{\mathrm{r}}$. The (i,j)-th element of the Jacobian J_{r} is:

$$\begin{aligned} (J_{\mathrm{r}})_{ij} &= \frac{h_i}{\sqrt{\sum_m h_m^2}} \frac{\partial\left(\operatorname{vec}(\rho_i) - \operatorname{vec}\left(\tilde{H}_i\right)\right)}{\partial\theta_j} \\ &= \frac{h_i}{\sqrt{\sum_m h_m^2}} \frac{\partial \operatorname{vec}(\rho_i)}{\partial\theta_j}. \end{aligned} \tag{56}$$

Hence, the (i,j)-th element of $J_{\mathrm{r}}^T J_{\mathrm{r}}$ should be:

$$\begin{aligned} \left(J_{\mathrm{r}}^T J_{\mathrm{r}}\right)_{ij} &= \sum_k \left(J_{\mathrm{r}}^T\right)_{ik} (J_{\mathrm{r}})_{kj} \\ &= \sum_k \frac{h_k}{\sqrt{\sum_m h_m^2}} \frac{h_k}{\sqrt{\sum_m h_m^2}} \left(\frac{\partial \operatorname{vec}(\rho_k)}{\partial\theta_i} \cdot \frac{\partial \operatorname{vec}(\rho_k)}{\partial\theta_j} \right) \\ &= \frac{1}{\sum_m h_m^2} \sum_k h_k^2 \operatorname{tr}\left(\partial_i \rho_k \partial_j \rho_k\right) \\ &= \frac{1}{\sum_m h_m^2} \sum_m h_m^2 (T_m)_{ij} \\ &= \frac{1}{2} W_{ij}. \end{aligned} \tag{57}$$

Hence, we have proved the relation $W = 2J_{\mathrm{r}}^T J_{\mathrm{r}}$.

ORCID iDs

Chenyu Shi 0009-0006-8691-5499
Vedran Dunjko 0000-0002-2632-7955
Hao Wang 0000-0002-4933-5181

References

[1] Preskill J 2018 Quantum computing in the NISQ era and beyond *Quantum* **2** 79
[2] Shor P W 1999 Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer *SIAM Rev.* **41** 303–32
[3] Grover L K 1996 A fast quantum mechanical algorithm for database search *Proc. 28th Annual ACM Symp. on Theory of Computing* pp 212–9

[4] Bharti K et al 2022 Noisy intermediate-scale quantum algorithms Rev. Mod. Phys. 94 015004
[5] Moll N et al 2018 Quantum optimization using variational algorithms on near-term quantum devices Quantum Sci. Technol. 3 030503
[6] Aspuru-Guzik A, Dutoi A D, Love P J and Head-Gordon M 2005 Simulated quantum computation of molecular energies Science 309 1704–7
[7] Peruzzo A, McClean J, Shadbolt P, Yung M-H, Zhou X-Q, Love P J, Aspuru-Guzik A and O'brien J L 2014 A variational eigenvalue solver on a photonic quantum processor Nat. Commun. 5 4213
[8] Farhi E, Goldstone J and Gutmann S 2014 A quantum approximate optimization algorithm (arXiv:1411.4028)
[9] Grange C, Poss M and Bourreau E 2023 An introduction to variational quantum algorithms for combinatorial optimization problems 4OR 21 363–403
[10] Havlíček V, Córcoles A D, Temme K, Harrow A W, Kandala A, Chow J M and Gambetta J M 2019 Supervised learning with quantum-enhanced feature spaces Nature 567 209–12
[11] Jerbi S, Fiderer L J, Nautrup H P, Kübler J M, Briegel H J and Dunjko V 2023 Quantum machine learning beyond kernel methods Nat. Commun. 14 1–8
[12] Larocca M, Thanasilp S, Wang S, Sharma K, Biamonte J, Coles P J, Cincio L, McClean J R, Holmes Z and Cerezo M 2025 Barren plateaus in variational quantum computing Nat. Rev. Phys. 7 1–16
[13] Larocca M, Thanasilp S, Wang S, Sharma K, Biamonte J, Coles P J, Cincio L, McClean J R, Holmes Z and Cerezo M 2024 A review of barren plateaus in variational quantum computing (arXiv:2405.00781)
[14] Sack S H, Medina R A, Michailidis A A, Kueng R and Serbyn M 2022 Avoiding barren plateaus using classical shadows PRX Quantum 3 020365
[15] Park C-Y and Killoran N 2024 Hamiltonian variational ansatz without barren plateaus Quantum 8 1239
[16] Robbins H and Monro S 1951 A stochastic approximation method Ann. Math. Stat. 22 400–7
[17] Diederik P K 2014 Adam: a method for stochastic optimization (No Title)
[18] Stokes J, Izaac J, Killoran N and Carleo G 2020 Quantum natural gradient Quantum 4 269
[19] Amari S-I 1998 Natural gradient works efficiently in learning Neural Comput. 10 251–76
[20] Martens J 2020 New insights and perspectives on the natural gradient method J. Mach. Learn. Res. 21 1–76
[21] Liu J, Yuan H, Xiao-Ming L and Wang X 2020 Quantum fisher information matrix and multiparameter estimation J. Phys. A: Math. Theor. 53 023001
[22] Anuar A A, Jamet F, Gironella F, SimkovicIV F and Rossi R 2024 Operator-projected variational quantum imaginary time evolution (arXiv:2409.12018)
[23] Huang H-Y, Kueng R and Preskill J 2020 Predicting many properties of a quantum system from very few measurements Nat. Phys. 16 1050–7
[24] Mitarai K, Negoro M, Kitagawa M and Fujii K 2018 Quantum circuit learning Phys. Rev. A 98 032309
[25] Schuld M, Bergholm V, Gogolin C, Izaac J and Killoran N 2019 Evaluating analytic gradients on quantum hardware Phys. Rev. A 99 032331
[26] Ollivier Y, Arnold L, Auger A and Hansen N 2017 Information-geometric optimization algorithms: a unifying picture via invariance principles J. Mach. Learn. Res. 18 pp 1–65
[27] Yao Y, Cussenot P, Wolf R A and Miatto F 2022 Complex natural gradient optimization for optical quantum circuit design Phys. Rev. A 105 052402
[28] Yamamoto N 2019 On the natural gradient for variational quantum eigensolver (arXiv:1909.05074)
[29] Koczor B and Benjamin S C 2022 Quantum natural gradient generalized to noisy and nonunitary circuits Phys. Rev. A 106 062416
[30] Koczor B and Benjamin S C 2022 Quantum analytic descent Phys. Rev. Res. 4 023017
[31] Beckey J L, Cerezo M, Sone A and Coles P J 2022 Variational quantum algorithm for estimating the quantum fisher information Phys. Rev. Res. 4 013083
[32] Lee J M 2018 Introduction to Riemannian Manifolds vol 2 (Springer)
[33] Gacon J, Zoufal C, Carleo G and Woerner S 2021 Simultaneous perturbation stochastic approximation of the quantum fisher information Quantum 5 567
[34] Nocedal J and Wright S J 1999 Numerical Optimization (Springer)
[35] Zhang K, Liu L, Hsieh M-H and Tao D 2022 Escaping from the barren plateau via gaussian initializations in deep variational quantum circuits Advances in Neural Information Processing Systems vol 35 pp 18612–27
[36] Neven A et al 2021 Symmetry-resolved entanglement detection using partial transpose moments npj Quantum Inf. 7 152
[37] Mele A A 2024 Introduction to haar measure tools in quantum information: a beginner's tutorial Quantum 8 1340
[38] Kuhn H W and Tucker A W 1951 Proceedings of 2nd berkeley symposium Proc. 2nd Berkeley Symp. pp 481–92