npj | quantum information

Article

Published in partnership with The University of New South Wales

https://doi.org/10.1038/s41534-024-00819-8

Analyzing variational quantum landscapes with information content

Check for updates

Adrián Pérez-Salinas[1,2], Hao Wang[1,3] & Xavier Bonet-Monroig[1,2]

The parameters of the quantum circuit in a variational quantum algorithm induce a landscape that contains the relevant information regarding its optimization hardness. In this work, we investigate such landscapes through the lens of information content, a measure of the variability between points in parameter space. Our major contribution connects the information content to the average norm of the gradient, for which we provide robust analytical bounds on its estimators. This result holds for any (classical or quantum) variational landscape. We validate the analytical understating by numerically studying the scaling of the gradient in an instance of the barren plateau problem. In such instance, we are able to estimate the scaling pre-factors in the gradient. Our work provides a way to analyze variational quantum algorithms in a data-driven fashion well-suited for near-term quantum computers.

Variational quantum algorithms (VQAs) have been marked as a promising path towards quantum advantage in pre-fault-tolerant quantum hardware. In nearly a decade of research since its original proposal[1], the field of VQAs has seen significant progress both theoretically and experimentally[2,3]. It is yet to be seen if noisy intermediate-scale quantum (NISQ)[4] devices are able to reach unambiguous quantum advantage through VQA. Issues such as vanishing gradients or barren plateaus (BP)[5–8], the expressivity of the quantum circuits[9–11] or difficulties optimizing a noisy cost function[12] are only a few examples of the hurdles faced by VQA which reduce the hope of quantum advantage in the near-term.

From a computer science point-of-view VQAs are a fascinating object of study. They can be considered classical cost functions with classic input/output. Yet the cost function might not be classically accessible in general. So far, there is no clear evidence that optimizing a VQA is feasible with standard optimization methods[12]. Some researchers have attempted to close this gap by developing new optimizers tailored to quantum circuits[13,14] or using machine learning techniques to assist during the optimization[15,16] with inconclusive results. More recently, the authors of ref. 17 introduced a method to visualize variational quantum landscapes through dimensionality reduction. Yet landscape analysis tools from classical optimization have not been widely used to characterize quantum landscapes.

Landscape analysis aims at characterizing the landscape of cost functions by efficiently sampling the parameter space to understand the "hardness" of the optimization problem[18–23]. For a VQA this implies only classical post-processing of data from a quantum device. In contrast, the optimization step of a VQA involves constant interaction between quantum and classical resources. In NISQ hardware such interactions might come with a large overhead. To the best of our knowledge, no prior work on data-driven landscape analysis exists in the context of VQA.

In this work, we aim to close the gap between VQA and landscape analysis through the information content (IC)[24] of the quantum landscape. We demonstrate the connection between IC and the average norm of the gradient of the landscape. We derive robust lower/upper bounds of this quantity which provides a crucial understanding of the landscape (e.g. complexity of optimizing the cost function). We apply our results to numerically study the BP problem for local and global cost functions from ref. 6, showing excellent agreement with theoretical asymptotic scaling in the size of the gradient. Also, we demonstrate how to calculate pre-factors of the asymptotic scaling, which are in practice more relevant for implementing algorithms. As far as we know, this is the first work where scaling pre-factors are calculated in the context of VQAs and BPs.

The manuscript is organized as follows. In the section "Results" we give background on VQAs and IC, connect the average norm of the gradient with IC, followed by a numerical diagnosis of BP using IC, and their estimation of scaling pre-factors. In the section "Discussion" we present the implications of our results and point out future directions. In the section "Methods" we prove the theorems and lemmas.

Results

Parameterized quantum circuits

In a variational quantum algorithm, one aims at exploring the space of quantum states by minimizing a cost function with respect to a set of tunable real-valued parameters $\vec{\theta} \in [0, 2\pi)^m$ of a parametrized quantum circuit (PQC). A PQC evolves an initial quantum state $|\psi_0\rangle$

[1]⟨aQaL⟩ Applied Quantum Algorithms, Universiteit Leiden, Leiden, Netherlands. [2]Instituut-Lorentz, Universiteit Leiden, Niels Bohrweg 2, 2333 CA Leiden, Netherlands. [3]LIACS, Universiteit Leiden, Niels Bohrweg 1, 2333 CA Leiden, Netherlands. e-mail: perezsalinas@lorentz.leidenuniv.nl; h.wang@liacs.leidenuniv.nl; bonet@lorentz.leidenuniv.nl

to generate a parametrized state

$$|\psi(\vec{\theta})\rangle = U(\vec{\theta})|\psi_0\rangle, \tag{1}$$

where $U(\vec{\theta})$ is a unitary matrix of the form

$$U(\vec{\theta}) = \prod_{i=1}^{m} U_i(\rightarrow \theta_i) W_i, \tag{2}$$

with

$$U_i(\vec{\theta}_i) = \exp\left(-i\vec{\theta}_i V_i\right). \tag{3}$$

Here W_i are fixed unitary operations and V_i are hermitian matrices. In a VQA these parameters are driven by (classically) minimizing a cost function $C(\vec{\theta})$, built as the expectation value of a quantum observable $\hat{O}$,

$$C(\vec{\theta}) = \langle\psi_0|U^\dagger(\vec{\theta})\hat{O}U(\vec{\theta})|\psi_0\rangle. \tag{4}$$

A successful optimization reaches an approximation to the lowest eigenvalue of $\hat{O}$, and the optimal parameters represent an approximation to its ground-state[2,3]. Our object of study is the manifold defined by a PQC and $\hat{O}$ which we call a variational quantum landscape.

In order to measure Eq. (4) in a real device, one must prepare and measure the observable with multiple copies of $U(\vec{\theta})|\psi_0\rangle$. Therefore, the real cost function becomes

$$\bar{C}(\vec{\theta}, R) = \langle\hat{O}\rangle + \kappa(R), \tag{5}$$

where $\kappa(R)$ introduces the uncertainty of sampling the observable with R repetitions. For a sufficiently large number of them, $\kappa(R)$ can be drawn from a Gaussian distribution with variance $\sigma^2 \sim 1/R$[12]. Throughout the rest of the text, we assume access to the exact value of the cost function, $C(\vec{\theta})$, unless otherwise stated.

Exploratory landscape analysis

The goal of exploratory landscape analysis (ELA)[25–27] is to characterize a real-valued function by numerically estimating a set of features from its landscape. Examples of such features are multi-modality (the number of local minima), ruggedness, or curvature of the level set of a given landscape (see ref. 27 for a comprehensive description of ELA features). ELA becomes particularly useful when studying functions with unknown analytical forms or exceedingly difficult for mathematical analysis where one can gain insight by comparing their features to those of known analytical functions[20,21,28]. Similarly, the classical machine learning/optimization community has made use of ELA features to select the most suitable optimization algorithms to optimize unknown cost functions[22,29,30]. A key aspect of landscape analysis is the fact that the number of evaluations of the cost function is significantly smaller than optimizing it; otherwise, it is more efficient to optimize directly.

Among the ELA features, information content (IC)[24,31] is of particular interest as it characterizes the ruggedness of the landscape which is related to its trainability. For example, high information content indicates a trainable landscape, whereas low information content indicates a flat landscape[24,27]. We demonstrate that for any classical or quantum variational landscape, its trainability can be quantified with information content.

Information content

Definition 1. (Information content (IC)) Given a finite symbolic sequence $\phi = \{-, \odot, +\}^S$ of length S and let p_{ab}, $a \neq b \in \{-, \odot, +\}$ denote the probability that ab occurs in the consecutive pairs of ϕ. The information content is defined as

$$H = \sum_{a\neq b} h(p_{ab}), \tag{6}$$

with

$$h(x) = -x\log_6 x. \tag{7}$$

In this definition pairs of the same symbols are excluded, leaving only six combinations

$$p_{ab} = \{p_{+-}, p_{-+}, p_{+\odot}, p_{\odot+}, p_{-\odot}, p_{\odot-}\}. \tag{8}$$

The $\log_6$ is necessary to ensure $H \leq 1$.

To compute the IC, we use the algorithm given in ref. 24:

(1) Sample $M(m) \in \mathcal{O}(m)$ points of the parameter space $\Theta = \{\vec{\theta}_1, \ldots, \vec{\theta}_M\} \in [0, 2\pi)^m$.
(2) Measure $C(\vec{\theta}_i)$ on a quantum computer (this is the only step where it is needed).
(3) Generate a random walk W of $S + 1 < M(m)$ steps over Θ, and compute the finite-size approximation of the gradient at each step i

$$\Delta C_i = \frac{C(\vec{\theta}_{i+1}) - C(\vec{\theta}_i)}{\left\|\vec{\theta}_{i+1} - \vec{\theta}_i\right\|}. \tag{9}$$

(4) Create a sequence $\phi(\epsilon)$ by mapping ΔC_i onto a symbol in $\{-,\odot,+\}$ with the rule

$$\phi(\epsilon) = \begin{cases} - & \text{if } \Delta C_i < -\epsilon \\ \odot & \text{if } |\Delta C_i| \leq \epsilon \\ + & \text{if } \Delta C_i > \epsilon \end{cases} \tag{10}$$

(5) Compute the empirical IC (denoted as $H(\epsilon)$ henceforth) by applying Definition 1 to $\phi(\epsilon)$.
(6) Repeat these steps for several values of ϵ.

From this algorithm, we are interested in the regimes of high and low $H(\epsilon)$[24,31]. The maximum IC (MIC) is defined as

$$H_M = \max_{\epsilon} H(\epsilon), \tag{11}$$

at ϵ_M as its corresponding ϵ. The MIC occurs when the variability in the symbols of ϕ is maximum.

The other case of interest occurs when $H(\epsilon) \leq \eta$ with $\eta \ll 1$. This defines the sensitivity IC (SIC)[24,31],

$$H_S = \min\{\epsilon > 0 | H(\epsilon) \leq \eta\}, \tag{12}$$

with $\epsilon = \epsilon_S$. The SIC identifies the ϵ at which (almost) all symbols in ϕ are $\odot$. All symbols become exactly $\odot$ at $\eta = 0$.

Both MIC and SIC are calculated on a classical computer after collecting $(\vec{\theta}, C(\vec{\theta}))$. The values $\epsilon_{M,S}$ can be found by sweeping over multiple values of ϵ with logarithmic scaling, and computing $H(\epsilon)$ with Eq. (6) at each of those values. (In the original work[27], it is suggested to take 1000 values of $\epsilon \in [0, 10^{15}]$). The cost is dominated by computing all quantities ΔC needed to construct $\phi(\epsilon)$ using Eq. (10). Since only $\mathcal{O}(m)$ samples are available, at most $\mathcal{O}(m^2)$ differences of the form ΔC, can be computed. Thus, MIC and SIC can be obtained with at most $\mathcal{O}(m^2)$ classical operations.

Next, we present the relation between IC and the average norm of the gradient, from now denoted as $\|\nabla C\|$. We take advantage of the fact that each step is isotropically random in W. This allows us to derive the

underlying probability of ΔC_i. Additionally, we use IC to bind the probability of pairs of symbols appearing along W, which allows us to estimate $\|\nabla C\|$. Although we demonstrate our results for a variational quantum landscape, they extend to any optimization landscape.

Estimation of the norm of the gradient

The random walks W over Θ satisfy

$$\left\|\vec{\theta}_{i+1} - \vec{\theta}_i\right\| \le d \tag{13}$$

$$\frac{\vec{\theta}_{i+1} - \vec{\theta}_i}{\left\|\vec{\theta}_{i+1} - \vec{\theta}_i\right\|} = \vec{\delta}_i, \tag{14}$$

where $\vec{\delta}_i$ is drawn from the isotropic distribution and d is fixed before starting the walk but might be varied. By Taylor expanding Eq. (9) and the mean-value theorem, the finite-size gradient can be written as

$$\Delta C_i = \nabla C((1-t)\vec{\theta}_i + t\vec{\theta}_{i+1}) \cdot \vec{\delta}_i, \tag{15}$$

with $t \in (0, 1)$. Since the sampled points Θ are chosen randomly, we can assume that ΔC and $\nabla C(\vec{\theta}) \cdot \vec{\delta}$ are drawn from the same probability distribution, given a sufficiently large Θ.

The isotropic condition of W allows us to calculate the probability distribution of $\nabla C(\vec{\theta}) \cdot \vec{\delta}$:

Lemma 1. *Let $C(\vec{\theta})$ be a differentiable function for all $\vec{\theta} \in [0, 2\pi)^m$. Let $\vec{\delta} \in \mathbb{R}, \|\vec{\delta}\| = 1$ be drawn from the isotropic distribution. Then $(\nabla C(\vec{\theta}) \cdot \vec{\delta})^2$ is a random variable with a beta probability distribution ($\mathcal{B}$)[32] such that*

$$\left(\nabla C(\vec{\theta}) \cdot \vec{\delta}\right)^2 \sim \left\|\nabla C(\vec{\theta})\right\|^2 \mathcal{B}\left(\frac{1}{2}, \frac{m-1}{2}\right). \tag{16}$$

The proof can be found in the section "Methods".

We can use Lemma 1 to bound the probability of $\nabla C(\vec{\theta})$ from the $\mathcal{B}$ cumulative distribution function (CDF).

Theorem 1. (CDF of $\nabla C(\vec{\theta}) \cdot \vec{\delta}$) *Let $C(\vec{\theta})$ be a differentiable function at every $\vec{\theta} \in [0, 2\pi)^m$. Let $\vec{\delta} \in \mathbb{R}, \|\vec{\delta}\| = 1$ be drawn from the isotropic distribution. Then $\nabla C(\vec{\theta}) \cdot \vec{\delta}$ is a random variable with cumulative density function*

$$\mathrm{Prob}\left(\nabla C(\vec{\theta}) \cdot \vec{\delta} \le \epsilon\right) = \frac{1}{2}\left(1 + \mathrm{sgn}(\epsilon)\mathcal{I}\left(\frac{\epsilon^2}{\|C(\vec{\theta})\|^2}; \frac{1}{2}, \frac{m-1}{2}\right)\right), \tag{17}$$

where $\mathcal{I}(x; \alpha, \beta)$ is the regularized incomplete beta function with parameters α and β.

The proof of this theorem can be found in the section "Discussion".

It is known that the beta distribution (with the parametrization in Lemma 1) rapidly converges to a normal distribution[33]:

$$\lim_{m\to\infty} \nabla C(\vec{\theta}) \cdot \vec{\delta} \sim \mathcal{N}\left(0, \frac{\|C(\vec{\theta})\|^2}{m}\right). \tag{18}$$

This approximation is accurate even for reasonably small values of m. We can naturally interpret the functionality of $\sqrt{m}$ as dimensionality normalization in Lemma 1.

Lemma 1 implies, in the Gaussian limit

$$\mathbb{E}_W\left(C(\vec{\theta}) \cdot \vec{\delta}\right) \sim \mathcal{N}\left(0, \frac{\|\nabla C\|_W^2}{m}\right) \tag{19}$$

where $\mathbb{E}_W$ denotes the expectation taken over the points in random walk W, and

$$\|\nabla C\|_W^2 = \mathbb{E}_W\left(\|\nabla C(\vec{\theta})\|^2\right). \tag{20}$$

As an immediate consequence, we give the CDF of $\nabla C(\vec{\theta}) \cdot \vec{\delta}$ averaged over a random walk W:

Corollary 1. (CDF of average norm of gradients) *Let $C(\vec{\theta})$ be a differentiable function at every $\vec{\theta} \in [0, 2\pi)^m$. Let $\vec{\delta} \in \mathbb{R}, \|\vec{\delta}\| = 1$ be drawn from the isotropic distribution. Then $\nabla C(\vec{\theta}) \cdot \vec{\delta}$ is a random variable with cumulative density function*

$$\mathrm{Prob}\left(\mathbb{E}_W\left(\nabla C(\vec{\theta}) \cdot \vec{\delta}\right) \le \epsilon\right) = \Phi_G\left(\frac{\epsilon\sqrt{m}}{\|\nabla C\|}\right), \tag{21}$$

where $\Phi_G(\cdot)$ is the CDF of the standard normal distribution.

Note that $\|\nabla C\|_W$ converges to the average norm of the gradient over $[0, 2\pi)^m$, i.e., $\|\nabla C\|^2 := \mathbb{E}||\nabla C(\vec{\theta})||^2$ at a rate[34]:

$$|\|\nabla C\| - \|\nabla C\|_W| \in \mathcal{O}_p(M^{-1/2}), \tag{22}$$

allowing us to approximate the interesting $\|\nabla C\|$ with the accessible $\|\nabla C\|_W$ with small error.

Probability concentration for information content

Our next goal is to bound the probability of pairs of symbols appearing in ϕ in the high and low IC regimes.

High information content. If one interprets IC as a partial entropy of the landscape, high H necessarily implies approximately equal probabilities p_{ab}. Therefore, a minimal concentration of probabilities must exist such that a high value of H can be reached. One can formally define this statement:

Lemma 2. *Let $H > 2h(1/2)$ be the IC of a given sequence ϕ. Consider the probabilities in Eq. (8) such that the sum of any four of them (p_4) is bounded by*

$$p_4 \ge 4q, \tag{23}$$

with q the solution of $H = 4h(x) + 2h(1/2 - 2x)$.

The proof can be found in the section "Methods". The bound on p_4 in the above lemma gets tighter as q increases, and so does H, which by definition $H \le 1$. The tightest possible bound is achieved for the MIC defined in Eq. (11).

Low information content. For a low IC to occur, all p_{ab} must be small, and their values can be upper-bounded.

Lemma 3. *Let H be the IC with bound $H \le \eta \le 1/6$ of a given sequence ϕ. Then the probability of consecutive $\odot$ steps during a random walk are close to 1. The expected norm of the gradient is bounded by*

$$p_{\odot\odot} \ge 1 - 3\eta, \tag{24}$$

The proof can be found in the section "Methods". As in the previous case, a tight bound is attained with the SIC defined in Eq. (12).

High and low IC provides insight into the hardness to optimize the cost function that generated the landscape. As shown in this section, low IC implies a flat landscape, which clearly imposes hardness in optimization. In contrast, high IC is a necessary but not sufficient condition for optimization. In this case, there is a guarantee that the landscape can be, at the very least, easy to optimize locally. However, IC does not provide any information on the quality of the accessible minimum (e.g. global or local) or the multimodality (number of local minima) of the landscape. The reason is the random walk over Θ, which allows for high IC both for convex or multimodal landscapes.

Information content to estimate the norm of the gradient

We are ready to show the main result of this work. We make use of the results in the section "Results" and section "Results" to prove that $H(\epsilon)$ estimates the average norm of the gradient $\|\nabla C\|$ for any classical or quantum landscape. To the best of our knowledge, this is the first time, including the field of classical optimization, where such bounds are calculated.

First, we relate H_{M} to the norm of the gradient. High values of IC guarantee a minimal probability for individual steps to increase and decrease and thus bounds the compatible values of $\epsilon/\|\nabla C\|_W$.

Theorem 2. (H_{M} bounds $\|\nabla C\|_W$) *Let H_{M} be the empirical MIC of a given function $C(\vec{\theta})$, and ϵ_{M} its corresponding ϵ. Let q be the solution to the equation $H_{\mathrm{M}} = 4h(x) + 2h(1/2-2x)$. Then,*

$$\frac{\epsilon_{\mathrm{M}}\sqrt{m}}{\Phi_G^{-1}(1-2q)} \leq \|\nabla C\|_W \leq \frac{\epsilon_{\mathrm{M}}\sqrt{m}}{\Phi_G^{-1}\left(\frac{1+2q}{2}\right)}. \tag{25}$$

The proof is given in the section "Methods".

The second result connects SIC to the bounds in the norm of the gradient. Small values of IC imply a large probability of consecutive $\odot$ steps in ϕ_i or equivalently small probabilities for p_{ab}. When this occurs, then ϵ_{S} is used to upper bound $\|\nabla C\|_W$.

Theorem 3. (H_{S} upper bounds $\|\nabla C\|_W$) *Let H_{S} be the empirical SIC of a given function $C(\vec{\theta})$, and ϵ_{S} its corresponding ϵ. Then*

$$\|\nabla C\|_W \leq \frac{\epsilon_{\mathrm{S}}\sqrt{m}}{\Phi_G^{-1}(1-3\eta/2)}. \tag{26}$$

The proof is given in the section "Methods".

From Eq. (22), it is obvious that $\|\nabla C\|_W$ will approximate $\|\nabla C\|$ for a large M. Hence, we can use Theorems 2 and 3 to bound $\|\nabla C\|$ with a long sequence ϕ.

Theorems 2 and 3 provide confidence intervals of $\|\nabla C\|$ without prior knowledge of the variational landscape. The former gives both an upper and lower bound of $\|\nabla C\|$ but can only be applied when H has a minimum value, while the latter is always applicable with an arbitrary small η for an upper bound.

Finally, we discuss the tightness of the bounds in Theorems 2 and 3. Firstly, the bound in Corollary 1 might become loose if the approximation from the Beta to a Gaussian distribution is not accurate. However, the error between these distributions is negligible already at $m \sim 10$. Moreover, as the value of m increases, so does the distribution, and the bound gets tighter. Another possibility might come from the entropy arguments of IC (see Lemmas 2 and 3). The bound saturates at $H_{\mathrm{M}} = 1$ but becomes looser as H_{M} decreases. When $H \leq 2h(1/2)$, they are no longer trustworthy. Nonetheless, for common values of $H_{\mathrm{M}} \sim 0.8$, the bounds are only lost by small factors. Similarly, for H_{S}, the bound tightens as η decreases but becomes uncontrolled with $\eta > 1/6$.

Information content under sampling noise

In the presence of sampling noise, symbols in $\phi(\epsilon)$ might be flipped due to the uncertainty of the cost function. The amount of flips within the sequence depends on the number of repetitions R. The following condition ensures that symbols in the sequence will not be flipped with high probability:

$$R \geq \epsilon^{-2}\left\|\vec{\theta}_{i+1} - \vec{\theta}_i\right\|^{-2} \geq \epsilon^{-2}d^{-2}. \tag{27}$$

It induces a lower bound $\epsilon \geq 1/d\sqrt{R}$. In this scenario, the bound in Corollary 1 and all subsequent results transform

$$\epsilon \to \epsilon \pm \kappa(R). \tag{28}$$

The bounds become slightly looser from the sampling noise dependence.

Sampling noise posses the same limitation to both IC and optimization. If $\|\nabla C\|$ decays exponentially, then exponentially many repetitions are required to resolve the gradient. Yet, IC is capable of signaling non-resolvable (flat) landscapes with less evaluations than a full optimization.

Information content to diagnose barren plateaus

Our goal is now to apply the previous results to study the problem of barren plateaus (BP)[5,6] We chose this problem because there exist analytical results on the scaling of the $\|\nabla C\|$. This allows us to directly verify that $H(\epsilon)$ can be used as a proxy to $\|\nabla C\|$.

BPs are characterized by the following conditions[5]:

$$\mathbb{E}\left(\partial_k C(\vec{\theta})\right) = 0, \tag{29}$$

$$\mathrm{Var}\left(\partial_k C(\vec{\theta})\right) \in \mathcal{O}\left(\exp(-n)\right), \tag{30}$$

where n is the number of qubits. BP implies exponentially vanishing variances of the derivatives. Similarly, BP can be understood as having a flat optimization landscape. These two concepts are connected by Chebyshev's inequality[35].

The IC allows us to calculate

$$\|\nabla C\| \approx \mathbb{E}_W\left(\sum_{k=1}^{m}(\partial_k C(\vec{\theta}))^2\right) = \sum_{k=1}^{m}\mathrm{Var}_W(\partial_k C(\vec{\theta})), \tag{31}$$

where Var_W computes the variance over points of the random walk W. Hence, IC is a proxy of the average variance of each partial derivative in the parameter space. To showcase IC as a tool to analyze the landscape of a VQA, we perform a numerical study of the BP problem as described by Cerezo et al.[6]. Here, the authors analytically derive the scaling of $\mathrm{Var}(\partial C(\vec{\theta}))$ in two different scenarios. Such scaling depends both on the qubit size and circuit depth of the PQC. If the cost function is computed from global observables (e.g. non-trivial support on all qubits), BPs exist irrespective of the depth of the PQC. In the case of local observables (e.g. non-trivial support on a few qubits), one can train shallow PQCs, but BPs gradually appear as the circuit depth increases. These results hold for alternating layered ansatz composed of blocks of 2-local operations (Fig. 4 in ref. 6).

Numerical experiments

In our numerical experiments, we use circuits from 2 to 14 qubits, each of them going from 4 to 16 layers. We calculate the cost function from

$$\hat{O}_{\mathrm{Local}} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - 1) \tag{32}$$

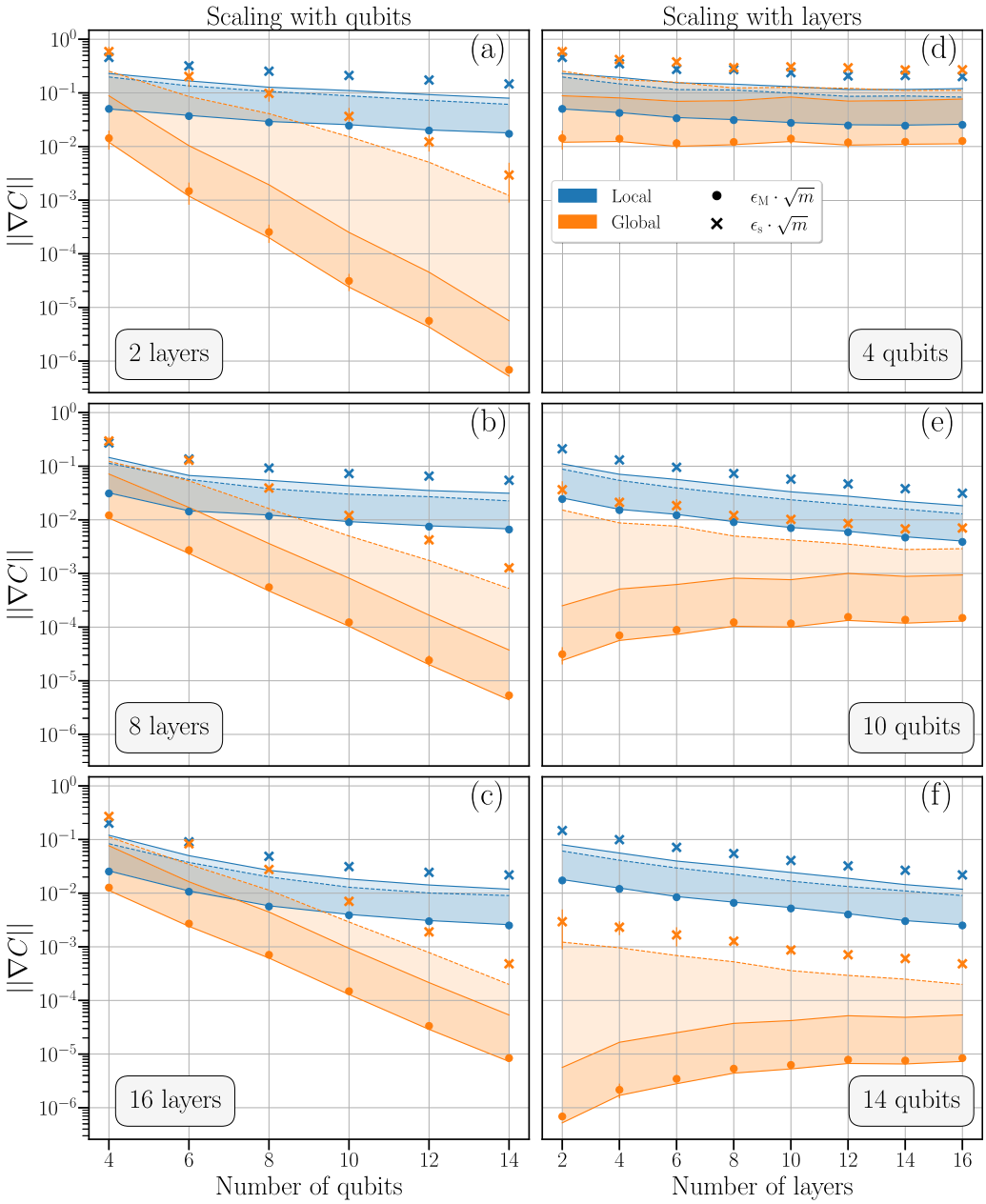

Fig. 1 | Scaling of the average gradient $\|\nabla C\|$. Panels **a**–**c** shows the scaling with with respect to qubits, and panels **d**–**f** the scaling with respect to layers. The solid lines show the bounds from Eq. (25) (solid lines) and Eq. (26) (dashed lines). $\|\nabla C\|$ can take values within the shadow areas in between these lines. The markers refer to the values of $\epsilon_M \cdot \sqrt{m}$ (dots) and $\epsilon_S \cdot \sqrt{m}$ (crosses). They are calculated from the median of five independent runs, with their standard deviation as the error bars. The colors represent the results for the local (blue) and global (orange) cost functions.

$$\hat{O}_{\text{Global}} = \bigotimes_{i=1}^{n} |0\rangle\langle 0|^{\otimes n}, \tag{33}$$

without sampling noise. Further details of the numerical experiments are given in the Supplementary Notes.

The results of the BP problem using IC are shown in Fig. 1. In all plots we compute the bounds on the $\|\nabla C\|$ from Theorem 2 (solid lines) and Theorem 3 (dashed lines), for the local (blue) and global (orange) cost functions. Additionally, we show the value of $\epsilon_M \cdot \sqrt{m}$ (dots) and $\epsilon_S \cdot \sqrt{m}$ (crosses).

The first trend we observe is that the $\|\nabla C\|$ shows two different scalings with respect to qubits (a–c) and layers (d–f). The scaling with qubits (a, b, c panels) shows an $\mathcal{O}(\text{poly}^{-1}(n))$ decay in the local cost function and a remarkable $\mathcal{O}(\exp(-n))$ decay with the global cost function. We emphasize the fact that these results are in perfect agreement with the predictions in ref. 6. On the other hand, the scaling with layers (d–f panels) strongly depends on the number of qubits. For 4 qubits, $\hat{O}_{\text{Global}}$ has a constant value $\|\nabla C\| \approx 10^{-2}$, while $\hat{O}_{\text{Local}}$ shows a small decay with a similar average. For 10 and 14 qubits, we recover the predicted $\mathcal{O}(\text{poly}^{-1}(n))$ decay in the $\hat{O}_{\text{Local}}$. In contrast, $\hat{O}_{\text{Global}}$ has $\|\nabla C\| \approx 10^{-4} - 10^{-6}$ $\hat{O}_{\text{Global}}$ which is already close to float precision. Finally, we observe that $\epsilon_M \cdot \sqrt{m}$ is close to the lower bound, thus making it a robust proxy for $\|\nabla C\|$.

In Fig. 2 we show a heatmap of the values of $\epsilon_M \cdot \sqrt{m}$ when increasing the number of qubits (x-axis) and layers (y-axis) for both local (panel a) and global (panel b) cost functions. The values for $\hat{O}_{\text{Local}}$ (panel a) show a rich variety of features: for 2–6 layers $\epsilon_M \cdot \sqrt{m}$ has a very slow decay, but for 8 or

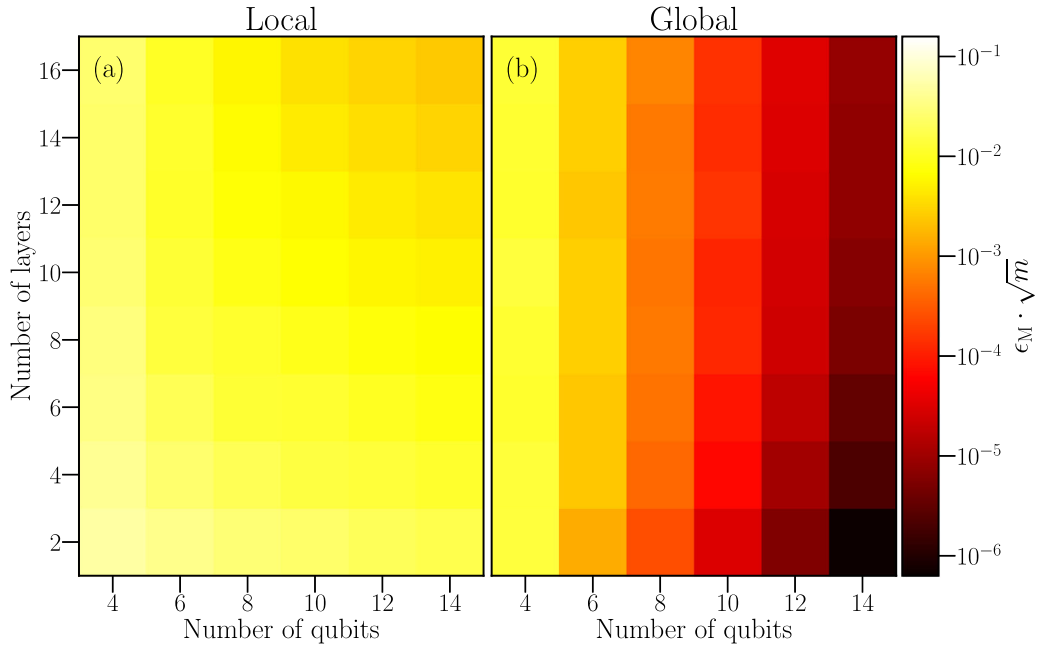

Fig. 2 | Heatmap of $\epsilon_M \cdot \sqrt{m}$. Panel **a** shows the values for the local cost function, and panel **b** the global cost function. The number of qubits is depicted on the x-axis, while the number of layers is shown on the y-axis.

more layers the $\epsilon_M \cdot \sqrt{m}$ decay quickly sharpens. This is exactly as expected for local cost functions: BPs appear gradually as the circuit depth increases. We speculate that the color change at the top right corner of the panel (a) in Fig. 2 corresponds to a transition regime. With regard to the global cost function (panel b), the expected exponential decay (in the number of qubits) is observed.

Surprisingly, both Figs. 1 and 2 show an increase in $\epsilon_M \cdot \sqrt{m}$ (or equivalently $\|\nabla C\|$) at a fixed number of qubits as the number of layers grows for the global cost function. We have not been able to find an explanation for this behavior either analytically or in the literature. However, this is an example of how data-driven methods might provide useful insight for deeper understanding.

Estimation of scaling pre-factor

Thus far the numerical results have just confirmed the asymptotic theoretical predictions of the considered BP problem. Our methodology can be used beyond the asymptotic scaling to compute actual pre-factors by fitting $\epsilon_M \cdot \sqrt{m}$ to its predicted functional form, including bounds on them from Eq. (25). Obtaining such pre-factors is challenging analytically. Yet they are relevant when studying the complexity of an algorithm in practice. In this section, we obtain the scaling pre-factors for the global cost function (in the number of qubits) and the local cost function for the number of qubits and layers (see Supplementary Figs. 1–3, and Tables 1–3 for additional details of these fittings).

First, we study the global cost function scaling with qubits for each number of layers in our data. We fit a linear model $f(x) = \alpha x + \beta$ with $x = \log_2(\epsilon_M \cdot \sqrt{m})$. The results of the fit are shown in Table 1. For each of the coefficients, we show the fitting values of the lower bound (LB column) and $\epsilon_M \cdot \sqrt{m}$ (right column). As the number of layers increases, $\alpha \to -1.0$, which is consistent with the exponential decay of the form 2^{-n} predicted in[6]. More importantly, asymptotic scaling is not sensitive to the constant factor β, but it is given by the right column in Table 1. Based on the trend in this column we speculate that the constant factor is $\beta \to -2.0$.

In the case of $\hat{O}_{\text{Local}}$, there are fewer known asymptotic predictions of the gradient norm. In ref. 6 it is shown that there exist three regimes: trainable, BPs, and transition area, depending on the depth with respect to the system size. Due to the small number of qubits and layers of our numerical study, we assume to be in the trainable regime, where theory predicts an $\|\nabla C\|$ scaling in $\mathcal{O}(\text{poly}^{-1}(n))$. We use a second-order polynomial model $f(n) = \alpha n^2 + \beta n + \gamma$ to fit $(\epsilon_M \cdot \sqrt{m})^{-1}$. The results are given in Table 2. The first observation is the small value of the quadratic coefficient for all layers. This might lead to thinking that a linear function will be better suited. To discard this possibility we perform a linear fit (see Supplementary Fig. 1) leading to comparable values of the slope and intercept, with slightly better fitting statistics for the second-degree polynomial. The β coefficient shows a 10-fold increase as the number of layers grows. In contrast, γ gets increasingly more negative with the number of layers. Note that we can extract the degree of the polynomial, which is impossible from the theory.

Lastly, we estimate the scaling coefficients of the local cost function with respect to layers, where no theoretical scaling is known[6]. We choose a second-order polynomial as the hypothesis functional form to fit the data. A linear model clearly under-fits the data (see Supplementary Fig. 3), thus confirming the intuition of a higher-order polynomial scaling. The results are presented in Table 3. The quadratic coefficient α increases as the number

Table 1 | Estimated qubit scaling pre-factors of the global cost function by linear fitting

Global cost function prefactors				
$f(n) = 2^{\alpha n+\beta}$				
	α		β	
Layers	LB	$\epsilon_M \cdot \sqrt{m}$	LB	$\epsilon_M \cdot \sqrt{m}$
2	−1.43	−1.41	−0.91	−0.68
4	−1.29	−1.27	−1.22	−1.09
6	−1.19	−1.17	−1.85	−1.68
8	−1.13	−1.12	−1.98	−1.85
10	−1.12	−1.12	−1.97	−1.82
12	−1.06	−1.05	−2.41	−2.26
14	−1.07	−1.07	−2.29	−2.14
16	−1.06	−1.06	−2.25	−2.10

α is the slope and β the intercept of the linear model. Each coefficient has two columns showing the results of fitting $\epsilon_M \cdot \sqrt{m}$ and its lower bound (LB).

Table 2 | Estimated qubit scaling pre-factors of the local cost function by fitting a second-degree polynomial

Local cost function scaling with qubits						
$f^{-1}(n) = \alpha n^2 + \beta n + \gamma$						
	α		β		γ	
Layers	LB	$\epsilon_M \cdot \sqrt{m}$	LB	$\epsilon_M \cdot \sqrt{m}$	LB	$\epsilon_M \cdot \sqrt{m}$
2	0.05	0.03	2.80	3.3	8.05	6.37
4	−0.25	−0.24	10.08	10.22	−12.86	−12.96
6	−0.16	−0.16	11.09	11.52	−11.37	−12.16
8	−0.27	−0.30	16.20	16.99	−26.22	−28.13
10	−0.56	−0.55	25.33	25.63	−57.30	−57.50
12	−0.31	− 0.34	25.52	26.77	−56.33	−59.65
14	0.14	0.08	25.99	28.00	−69.42	−75.87
16	0.13	0.12	33.69	34.95	−104.77	−108.92

Each coefficient has two columns showing the results of fitting $\epsilon_M \cdot \sqrt{m}$ and its lower bound (LB).

Table 3 | Estimated layer scaling pre-factors of the local cost function by fitting a second-degree polynomial

Local cost function scaling with layers						
$f^{-1}(l) = \alpha l^2 + \beta l + \gamma$						
	α		β		γ	
Qubits	LB	$\epsilon_M \cdot \sqrt{m}$	LB	$\epsilon_M \cdot \sqrt{m}$	LB	$\epsilon_M \cdot \sqrt{m}$
4	−0.1	−0.11	3.29	3.5	12.74	12.43
6	−0.27	−0.28	9.54	9.94	6.26	6.10
8	0.13	0.13	7.69	7.98	18.80	19.16
10	0.48	0.52	5.98	5.89	27.79	29.72
12	0.81	0.82	4.68	5.04	38.24	38.70
14	1.14	1.18	2.95	3.04	48.93	50.89

Each coefficient has two columns showing the results of fitting $\epsilon_M \cdot \sqrt{m}$ and its lower bound (LB).

of qubits grows at $n \geq 8$, and so does the constant term γ. In contrast, the linear term β remains roughly constant across all system sizes studied. An opposite trend in the coefficients seems to occur between $n = 4$ and $n = 6$, α and γ decrease, while β increases. We have not been able to match this change in the tendency to a change in scaling, leaving a finite-size effect in the fitting as the most possible explanation.

The results presented in this section are a demonstration that data-driven approaches can provide useful insight to complement analytical methods and can be leveraged to get a deeper understanding of a problem.

Discussion

Variational quantum algorithms have been intensively studied as a suitable application for near-term quantum hardware. From a computer science perspective, they are simply an optimization problem with classical input/output, yet their cost function is a quantum object. The parameters of any optimization problem induce a landscape that contains information about its "hardness". Landscape analysis is central to classical optimization but has been somewhat ignored in the VQAs community.

In this work, we investigate the information content features of a variational quantum landscape, which can be calculated efficiently by sampling the parameter space. We prove that for any (classical or quantum) cost function the average norm of its gradient can be rigorously bounded with the information content. We validate our theoretical understating by a numerical experiment, confirming the predicted asymptotic scaling of the gradient in the barren plateau problem. Finally, we apply our results to predict scaling pre-factors of the gradient in a data-driven fashion. To our knowledge, this is the first time that such pre-factors are calculated for a VQA.

The study of optimization landscapes of VQA opens a new avenue to explore their capabilities within the NISQ era. First, landscape analysis does not require constant interaction between quantum and classical hardware. Secondly, only linear (in the number of parameters) queries to a quantum computer suffice to extract the information content instead of polynomially many queries for a standard optimization routine. Finally, information content might be used as an easy and comparable metric between ansatzes.

We envision future research directions with information content such as studying the feasibility of the VQA optimization, estimating the number of shots needed to resolve a gradient, or warm-starting the algorithm from regions of interest in parameter space. Importantly, landscape analysis does not rely on any constraint in the quantum circuit. It can be then used even in the case when analytical approaches are unavailable. Therefore we anticipate that landscape analysis and information content might have a broad range of applications beyond VQAs in the NISQ era.

Methods

Proof. Proof of Lemma 1

The main assumption of Lemma 1 is $\vec{\delta}$ is drawn from the isotropic distribution on the unit sphere in m dimensions. As a first step, we use the spherical symmetry of the parameter space to align the first coordinate of $\vec{\delta}$ with the vector $\nabla C(\vec{\theta})$. Thus,

$$\left(\nabla C(\vec{\theta}) \cdot \vec{\delta}\right)^2 = \left\|\nabla C(\vec{\theta})\right\|^2 \delta_1^2. \tag{34}$$

Now, we redefine the isotropic distribution as the normalized multi-dimensional Gaussian distribution ($\mathcal{N}$),

$$\vec{\delta} = \frac{\vec{x}}{\|\vec{x}\|}; \qquad \text{with } \vec{x} \sim \mathcal{N}(0, \mathbb{I}^m). \tag{35}$$

By definition, each of the coordinates-squared in the multi-dimensional Gaussian distribution follows a χ^2 distribution[36]. In particular, $x_1^2 \sim \chi^2(1)$, and $\sum_{i=2}^m x_i^2 \sim \chi^2(m-1)$. It is well-known[32] that the above quotient follows a beta distribution with parameters 1/2 and $(m-1)/2$, i.e.,

$$\left(\nabla C(\vec{\theta}) \cdot \vec{\delta}\right)^2 = \left\|\nabla C(\vec{\theta})\right\|^2 \frac{x_1^2}{x_1^2 + \sum_{i=2}^m x_i^2} \sim \left\|\nabla C(\vec{\theta})\right\|^2 \mathcal{B}\left(\frac{1}{2}, \frac{m-1}{2}\right), \tag{36}$$

finishing the proof. □

Proof. Proof of Theorem 1 The CDF of a beta distribution is the regularized incomplete beta function $\mathcal{I}$. Thus, in the assumptions of Lemma 1,

$$\Pr\left(\left(\nabla C(\vec{\theta})\cdot\vec{\delta}\right)^2\le\epsilon^2\right)=\mathcal{I}\left(\frac{\epsilon^2}{||\nabla C(\vec{\theta})||^2},\frac{1}{2},\frac{m-1}{2}\right),\tag{37}$$

where ϵ is a realization of $(\nabla C(\vec{\theta})\cdot\vec{\delta})^2$.

We are however interested in $\nabla C(\vec{\theta})\cdot\vec{\delta}$. From the isotropic condition of $\vec{\delta}$, it is immediate that $\nabla C(\vec{\theta})\cdot\vec{\delta}$ is symmetric with respect to 0. Using this observation,

$$\Pr(|\nabla C(\vec{\theta})\cdot\vec{\delta}|\le\epsilon)=\Pr(-\epsilon\le\nabla C(\vec{\theta})\cdot\vec{\delta}\le 0)+\Pr(0\le\nabla C(\vec{\theta})\cdot\vec{\delta}\le\epsilon)\tag{38}$$

$$=2\Pr(-\epsilon\le\nabla C(\vec{\theta})\cdot\vec{\delta}\le 0)=2\Pr(0\le\nabla C(\vec{\theta})\cdot\vec{\delta}\le\epsilon)\tag{39}$$

$$=\Pr\left(\left(\nabla C(\vec{\theta})\cdot\vec{\delta}\right)^2\le\epsilon^2\right)=\mathcal{I}\left(\frac{\epsilon^2}{||\nabla C(\vec{\theta})||^2},\frac{1}{2},\frac{m-1}{2}\right).\tag{40}$$

The step taken in Eq. (39) allows us to rewrite them as

$$\Pr(\nabla C(\vec{\theta})\cdot\vec{\delta}\le\epsilon)=\frac{1}{2}\left(1+\mathrm{sgn}\,(\epsilon)I\left(\frac{\epsilon^2}{||\nabla C(\vec{\theta})||^2},\frac{1}{2},\frac{m-1}{2}\right)\right),\tag{41}$$

where sgn is the sign function. □

Proof. Proof of Lemma 2

For this proof, we must focus on the regime of large values of the IC. We recall the definition of the IC from Definition 1

$$H=\sum_{a\neq b}h(p_{ab}),\tag{42}$$

with $h(x)=-x\log_6 x$. We define the inverse function h^{-1} to be applied in the domain $x\le 1/e$.

For a given value of the sum of probabilities, the maximum entropy is achieved for uniform distribution. This leads to the expression

$$\sum_{a\neq b}p_{ab}\ge 6h^{-1}(H/6).\tag{43}$$

Note that a given value of H is compatible with probability distributions with larger joint probability but uneven distributions. Completeness of the probability distribution implies $\sum_{a\neq b}p_{ab}\le 1$. The properties of the function $h(x)$ allow maintaining a value H, with one probability p_1 to decrease for a given quantity x, as long as another probability increases some other quantity $f(x)>x$. Hence, a high value of H implies a minimal value on at least some set of probabilities.

We focus on the probability held by only 4 elements in the probability distribution. We first split the IC value into two pieces, the 4 smallest ones and the 2 largest,

$$H=\sum_4 h(p_{ab})+\sum_2 h(p_{ab}),\tag{44}$$

where $\sum_4$ indicates the sum over the smallest terms, and $\sum_2$ stands for the largest terms. To obtain the minimal probability held by the smallest 4 terms, we start in the situation with the smallest possible sum of all 6 probabilities, namely $p_{ab}=h^{-1}(H/6)\equiv q,\forall\, a,b$. Now we subtract probability from these 4 terms and add probability to the remaining terms to keep the IC constant.

$$H=4h(q-x_1-x_2)+h(q+f_1(x_1))+h(q+f_2(x_2)),\tag{45}$$

where $f_{1,2}$ is whatever function needed. Concavity of the function h allows us to bound

$$H\le 4h(q-x)+2h(q+f(x)).\tag{46}$$

This bound holds as long as

$$4(q-x)+2(q+f(x))\le 1,\tag{47}$$

where the equality is satisfied under the limit case. Substituting this equality into Eq. (46), we obtain

$$H\le 4h(q-x)+2h\left(\frac{1}{2}-2q+2x\right),\tag{48}$$

which comprises the values of x compatible with H. This bound also considers probability transferred to elements not relevant to the IC. Recalling p_4, we know

$$\sum_4 p_{ab}\ge 4(q-x),\tag{49}$$

and a more straightforward version of this condition is written as

$$\sum_4 p_{ab}\ge 4q_4,\tag{50}$$

with q_4 being the solution to the equation $H=4h(x)+2h(1/2-2x)$. □

Proof. Proof of Lemma 3

The first step is to observe that in case of sufficiently small IC, there are only two possible scenarios, which are (a) only one of the probabilities p_{ab}, with $a\neq b$ is close to one, and (b) all p_{ab} are close to zero. This scenario (a) is impossible by construction since these probabilities must come at least in pairs. Thus, all of them are small. In particular, since $h(x)\ge x$ for $x\le 1/6$, we bound

$$\sum_{a\neq b}p_{ab}\le H\le\eta\le 1/6.\tag{51}$$

Since all probabilities are small and are also combinations of the probabilities of only one step to be $p_+,p_-,p_\odot$, we can conclude that at least two of those must be small. Since $p_\pm$ are symmetric by construction, the only candidate left is $p_\odot$, which is the event that concentrates all probabilities. This observation allows us to bound

$$p_{++}\le p_{\odot+}\le\eta,\tag{52}$$

and subsequently

$$p_{\odot\odot}=1-p_{++}-p_{--}-\sum_{a\neq b}p_{ab}\ge 1-3\eta.\tag{53}$$

□

Proof. Proof of Theorem 2

For this theorem, we need to bound the probability of one step of the random walk to be $\{-,\odot,+\}$. However, the IC measures only pairs of steps. In this proof, we use the results from Lemma 2 to bound probabilities in only one step, and we connect those to the results in Corollary 1. We take, without

loss of generality p_+,

$$p_+ = p_{++} + \frac{1}{2}\left(p_{+\odot} + p_{\odot+} + p_{-+} + p_{+-}\right) \tag{54}$$

The IC is insensitive to p_{++}, so we discard it. The second term is bounded by Lemma 2, thus

$$p_+ \geq 2q_4, \tag{55}$$

with q_4 being the solution to the equation $H = 4h(x) + 2h(1/2-2x)$. Now, we recall Corollary 1 and bound

$$\Phi_G\left(\frac{\epsilon\sqrt{m}}{\|\nabla C\|_W}\right) \geq 2q_4, \tag{56}$$

which directly leads to

$$\|\nabla C\|_W \geq \frac{-\epsilon\sqrt{m}}{\Phi_G^{-1}(2q_4)}, \tag{57}$$

yielding the desired result, up to a symmetry of the CDF function. The upper bound can be obtained by applying the same reasoning to $p_0 \geq 2q_4$. □

Proof. Proof of Theorem 3

For this theorem, we need to bound the probability of one step of the random walk to be $\odot$. The results from Lemma 3 bound $P_{\odot\odot}$, so

$$p_\odot \geq p_{\odot\odot} \geq 1 - 3\eta \tag{58}$$

Now we connect this bound to the results in Corollary 1.

$$\Phi_m\left(\frac{-\epsilon}{\|\nabla C\|_W}\right) \leq \frac{3}{2}\eta, \tag{59}$$

which yields the result

$$\|\nabla C\|_W \leq \frac{-\epsilon}{\Phi_m^{-1}(3\eta/2)}. \tag{60}$$

□

Data availability

The data to reproduce the results of this work can be found in Zenodo[37].

Code availability

All packages used in this work are open source and available via git repositories online.

Received: 24 April 2023; Accepted: 9 February 2024;
Published online: 29 February 2024

References

1. Peruzzo, A. et al. A variational eigenvalue solver on a photonic quantum processor. *Nat. Commun.* **5**, 4213 (2014).
2. Cerezo, M. et al. Variational quantum algorithms. *Nat. Rev. Phys.* **3**, 625–644 (2021).
3. Bharti, K. et al. Noisy intermediate-scale quantum algorithms. *Rev. Mod. Phys.* **94**, 015004 (2022).
4. Preskill, J. Quantum computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018).
5. McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R. & Neven, H. Barren plateaus in quantum neural network training landscapes. *Nat. Commun.* **9**, 4812 (2018).
6. Cerezo, M., Sone, A., Volkoff, T., Cincio, L. & Coles, P. J. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nat. Commun.* **12**, 1791 (2021).
7. Wang, S. et al. Noise-induced barren plateaus in variational quantum algorithms. *Nat. Commun.* **12**, 6961 (2021).
8. Larocca, M. et al. Diagnosing barren plateaus with tools from quantum optimal control. *Quantum* **6**, 824 (2022).
9. Herasymenko, Y. & O'Brien, T. A diagrammatic approach to variational quantum ansatz construction. *Quantum* **5**, 596 (2021).
10. Du, Y., Tu, Z., Yuan, X. & Tao, D. Efficient measure for the expressivity of variational quantum algorithms. *Phys. Rev. Lett.* **128**, 080506 (2022).
11. Holmes, Z., Sharma, K., Cerezo, M. & Coles, P. J. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum* **3**, 010313 (2022).
12. Bonet-Monroig, X. et al. Performance comparison of optimization methods on variational quantum algorithms. *Phys. Rev. A* **107**, 032407 (2023).
13. Nakanishi, K. M., Fujii, K. & Todo, S. Sequential minimal optimization for quantum-classical hybrid algorithms. *Phys. Rev. Res.* **2**, 043158 (2020).
14. Ostaszewski, M., Grant, E. & Benedetti, M. Structure optimization for parameterized quantum circuits. *Quantum* **5**, 391 (2021).
15. Wilson, M. et al. Optimizing quantum heuristics with meta-learning. *Quantum Mach. Intell.* **3**, 13 (2021).
16. Sung, K. J. et al. Using models to improve optimizers for variational quantum algorithms. *Quantum Sci. Technol.* **5**, 044008 (2020).
17. Rudolph, M. S. et al. ORQVIZ: visualizing high-dimensional landscapes in variational quantum algorithms. Preprint at https://arxiv.org/abs/2111.04695 (2021).
18. Kerschke, P. & Trautmann, H. Automated algorithm selection on continuous black-box problems by combining exploratory landscape analysis and machine learning. *Evol. Comput.* **27**, 99–127 (2019).
19. Zou, F. et al. A survey of fitness landscape analysis for optimization. *Neurocomputing* **503**, 129–139 (2022).
20. Kerschke, P., Preuss, M., Wessing, S. & Trautmann, H. Detecting funnel structures by means of exploratory landscape analysis. In *Proc. Genetic and Evolutionary Computation Conference, GECCO 2015*, Madrid, Spain, July 11–15, 2015 (eds. Silva, S. & Esparcia-Alcázar, A. I.), 265–272 (ACM, 2015).
21. Morgan, R. & Gallagher, M. Analysing and characterising optimization problems using length scale. *Soft Comput.* **21**, 1735–1752 (2017).
22. Bischl, B., Mersmann, O., Trautmann, H. & Preuß, M. Algorithm selection based on exploratory landscape analysis and cost-sensitive learning. In *Genetic and Evolutionary Computation Conference, GECCO '12, Philadelphia, PA, USA, July 7–11, 2012* (eds. Soule, T. & Moore, J. H.) 313–320 (ACM, 2012).
23. Kostovska, A. et al. Per-run algorithm selection with warm-starting using trajectory-based features. In *Parallel Problem Solving from Nature–PPSN XVII: 17th International Conference, PPSN 2022, Dortmund, Germany, September 10–14, 2022, Proc., Part I*, (eds. Rudolph, G. et al.) 46–60 (Springer, 2022).
24. Muñoz, M. A., Kirley, M. & Halgamuge, S. K. Exploratory landscape analysis of continuous space optimization problems using information content. *IEEE Trans. Evol. Comput.* **19**, 74–87 (2015).
25. Mersmann, O. et al. Exploratory landscape analysis. In *13th Annual Genetic and Evolutionary Computation Conference, GECCO 2011, Proceedings, Dublin, Ireland, July 12–16, 2011* (eds. Krasnogor, N. & Lanzi, P. L.) 829–836 (ACM, 2011).
26. Kerschke, P. et al. Towards analyzing multimodality of continuous multiobjective landscapes. In *Parallel Problem Solving from Nature—PPSN XIV—14th International Conference, Edinburgh, UK, September 17–21, 2016, Proceedings*, Vol. 9921, *Lecture Notes in Computer Science* (eds Handl, J. et al.) 962–972 (Springer, 2016).

27. Kerschke, P. & Trautmann, H. Comprehensive feature-based landscape analysis of continuous and constrained optimization problems using the r-package flacco. In *Applications in Statistical Computing: from Music Data Analysis to Industrial Quality Improvement* (eds. Bauer, N. et al.) 93–123 (Springer, 2019).
28. Mersmann, O., Preuss, M., Trautmann, H., Bischl, B. & Weihs, C. Analyzing the BBOB results by means of benchmarking concepts. *Evol. Comput.* **23**, 161–185 (2015).
29. Kerschke, P., Kotthoff, L., Bossek, J., Hoos, H. H. & Trautmann, H. Leveraging TSP solver complementarity through machine learning. *Evol. Comput.* **26**, 597–620 (2018).
30. Mersmann, O. et al. A novel feature-based approach to characterize algorithm performance for the traveling salesperson problem. *Ann. Math. Artif. Intell.* **69**, 151–182 (2013).
31. Vassilev, V. K., Fogarty, T. C. & Miller, J. F. Information characteristics and the structure of landscapes. *Evol. Comput.* **8**, 31–60 (2000).
32. Bailey, R. W. Distributional identities of beta and chi-squared variates: a geometrical interpretation. *Am. Stat.* **46**, 117–120 (1992).
33. Temme, N. M. Asymptotic inversion of the incomplete beta function. *J. Comput. Appl. Math.* **41**, 145–157 (1992).
34. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
35. Tchébychef, P. Des valeurs moyennes. *J. Math. Pures Appl.* **12**, 177–184 (1867).
36. Johnson, N. L., Kotz, S. & Balakrishnan, N. *Continuous Univariate Distributions* Vol. 2 (John Wiley & Sons, 1995).
37. Pérez-Salinas, A., Wang, H. & Bonet-Monroig, X. Code and data: analyzing variational quantum landscapes with information content. *Zenodo* https://zenodo.org/records/7760281 (2023).

Acknowledgements

The authors would like to thank Carlo Beenakker, Vedran Dunjko, and Jordi Tura for their support in this project and Patrick Emonts and Yaroslav Herasymenko for useful feedback on the manuscript. The authors would also like to thank all aQa members for fruitful discussions. A.P.S. acknowledges support from 'Quantum Inspire - the Dutch Quantum Computer in the Cloud' project (with project number [NWA.1292.19.194]) of the NWA research program 'Research on Routes by Consortia (ORC)', which is funded by the Netherlands Organization for Scientific Research (NWO). X.B.M. acknowledges funding from the Quantum Software Consortium.

Author contributions

The project was conceived by X.B.M. and H.W. A.P.S. and H.W. proved the theorems, corollaries, and lemmas. X.B.M. performed the numerical simulations with assistance from A.P.S. and H.W. X.B.M. analyzed the data and performed the model fits. All the authors interpreted the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41534-024-00819-8.

Correspondence and requests for materials should be addressed to Adrián Pérez-Salinas, Hao Wang or Xavier Bonet-Monroig.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024