Critical Points in Hamiltonian Agnostic Variational Quantum Algorithms

Eric R. Anschuetz

MIT Center for Theoretical Physics, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

One of the most important properties of classical neural networks is the clustering of local minima of the network near the global minimum, enabling efficient training. This has been observed not only numerically, but also has begun to be analytically understood through the lens of random matrix theory. Inspired by these results in classical machine learning, we show that a certain randomized class of variational quantum algorithms can be mapped to Wishart random fields on the hypertorus. Then, using the statistical properties of such random processes, we analytically find the expected distribution of critical points. Unlike the case for deep neural networks, we show the existence of a transition in the quality of local minima at a number of parameters exponentially large in the problem size. Below this transition, all local minima are concentrated far from the global minimum; above, all local minima are concentrated near the global minimum. This is consistent with previously observed numerical results on the landscape behavior of Hamiltonian agnostic variational quantum algorithms. We give a heuristic explanation as to why ansatze that depend on the problem Hamiltonian might not suffer from these scaling issues. We also verify that our analytic results hold experimentally even at modest system sizes.

I. INTRODUCTION

A. Motivation

Variational quantum algorithms (VQAs) are one of the most promising meta-algorithms for utilizing noisy intermediate-scale quantum (NISQ) computers [1] for real-world computational problems, such as finding the potential landscape of molecules [2] and combinatorial optimization [3]. Furthermore, certain quantum generative modeling algorithms can be expressed in the same framework [4]. In these algorithms, a quantum computer prepares parameterized trial wavefunctions \( |\theta\rangle \) and measures an objective observable \( O \); the observable measurement is then fed into a classical computer, which then tunes the parameters of the trial wavefunction via a classical optimization algorithm to minimize the loss function

\[
F_O(\theta) = \langle \theta | O | \theta \rangle .
\]

This class of algorithms is promising for NISQ devices as the only quantum computational resources needed are enough qubits to store the ansatz wavefunction \( |\theta\rangle \) and enough coherence to prepare \( |\theta\rangle \); all other computations are performed on a classical computer. To minimize the coherence requirements of the quantum device in preparing \( |\theta\rangle \), often a class of ansatze known as hardware-efficient ansatze [5] are used. This class of ansatze are independent of the problem observable \( O \), and are motivated strictly by the hardware constraints of the NISQ device the algorithm is run on. Although in principle these ansatze are expressive enough (assuming large enough circuit depth) to accurately prepare the ground state of many observables \( O \), the optimization landscapes induced by such ansatze are typically characterized by so-called barren plateaus [6–8]. That is, the optimization landscape is generically very flat at high depth (or when using a global objective observable \( O \)), and most known optimization algorithms will struggle to find a local minimum of \( F_O \).

Numerically, a separate infeasibility result has been observed in the training of VQAs. Namely, in certain shallow instances, the distribution of local minima of the objective function is concentrated far away from the global minimum. This makes training using gradient-based algorithms infeasible, as the training algorithm will generally halt far away from the true minimum of the loss function. This behavior was first observed in the learning of unitaries via a certain randomized ansatz, where numerically it was shown that training failed unless the depth was on the order of the dimension of the objective unitary [9]. Similar numeric results [10][12] for VQAs have also shown a “training phase transition,” where at shallow depth almost all local minima are far from the global minimum even for local Hamiltonians, but after some cutoff depth, all local minima are near the global optimum. These numerical results are in stark contrast with the case in classical machine learning, where it is known that even shallow (but wide) neural networks have local minima that well-approximate the global minimum [13][14]. In the classical case, these results
are shown by mapping the class of neural networks under consideration to a type of Gaussian random field on the hypersphere, and then using the Kac–Rice formula on manifolds [15] to analytically find the expected distribution of critical points of the loss function in the limit of a large number of parameters [16].

### B. Summary of Results

In this work, we analytically show the presence of a “phase transition” in the training of VQAs for a certain class of randomized ansatz inspired by the hardware-efficient class of ansatzes, qualitatively similar to the behavior previously observed numerically [9][12]. In particular, we analytically study a class of ansatz constructed via random Pauli rotations, and by mapping the distribution of VQAs $F_O$ over this ansatz (with fixed $O$) to a class of random fields on the hypertorus $\mathbb{S}^p$ we are able to analytically calculate the distribution of critical points of a specified index. Asymptotically, we are able to show that there is a “phase transition” from an underparameterized regime—where all local minima are exponentially concentrated near half the mean eigenvalue of $O$—and an overparameterized regime—where all local minima are exponentially concentrated near the global minimum of $O$. Numerically, we also observe this behavior at a modest depth $p$. Additionally, in the process we prove novel results in the distribution of local minima for this class of random fields, which may be of independent interest.

Our proof techniques are similar to those used in the classical machine learning literature on the distribution of local minima for classes of machine learning models [13][14]. First, we map our models of interest—in our case, a certain class of ansatzes inspired by hardware-efficient ansatzes—to a well-known class of random fields on a compact manifold. Then, we use Morse theory to calculate the expected distribution of critical points of the given random field. Finally, we utilize large deviations theory on this class of random fields to give closed-form solutions for the distribution of critical points in various limits. We give an abbreviated overview in the remainder of this section.

#### 1. Mapping Variational Quantum Algorithms to Random Fields

Previous results in the classical machine learning literature [13][14] on the distribution of critical points in the loss landscape begin by mapping their class of machine learning models of interest to a class of random fields known as Gaussian hyperspherical random fields. These are random fields of the form:

$$F_{\text{GHRF}}(\theta) \propto \sum_{i_1,...,i_r,i'_1,...,i'_r=1}^{A} \sigma_{i_1} \cdots \sigma_{i_r} J_{i_1,...,i_r,i'_1,...,i'_r} \sigma_{i'_1} \cdots \sigma_{i'_r},$$ (2)

where $\sigma$ is a point on the hypersphere $S^A$ parameterized by $\theta$. Here, each $J_{i_1,...,i_r,i'_1,...,i'_r}$ is an i.i.d. random Gaussian variable. These mappings rely on nonlinearities in the model providing the effective Gaussian interaction in Eq. (2), and the fact that the linear transformations in the model are completely parameterized gives the product of $\sigma$, in Eq. (2). Known results on the loss landscape of this class of random fields [16][18] were then used to infer the loss landscape of the studied machine learning models; namely, that the local minima of the landscape concentrate near the global minimum.

Our mapping of Eq. (1) with a class of hardware-efficient inspired ansatzes differs from the classical machine learning construction in two major ways. First, the variational ansatz is linear; the nonlinearity of the classical construction was crucial in giving the effective Gaussian couplings $J$. Second, the unitaries in variational ansatzes are (usually) parameterized rotations by simple Pauli strings; this means the product of unitaries in the ansatz do not give a simple product of the ansatz parameters as in Eq. (2), and each layer is vastly underparameterized. We will later find that this latter fact gives rise to the qualitative differences in the loss landscapes of deep neural networks and variational quantum algorithms.

Thus, our mapping relies on a different strategy. To make the mapping tractable, we consider ansatzes built from rotations by uniformly random Pauli strings. This has the added benefit of making the parameters of our ansatz naturally described as points on the hypertorus. We then consider the path integral expansion of Eq. (1), where each path is weighted by the parameters of the ansatz and various matrix elements of $O$. Then, we show at fixed $O$, these matrix elements are approximately a sum of (shifted) Wishart random variables. This is just a simple multivariate generalization of the gamma distribution. Finally, we express this sum of Wishart random variables as a single Wishart random variable, which is close in distribution under reasonable assumptions on $O$. This means that, unlike

---

1 The $p$-dimensional hypertorus is the product space $(S^1)^\times p$, where $S^1$ is the circle.
The classical case, the natural random field to study for VQAs is the \textit{Wishart hypertoroidal random field} (WHRF). These results are summarized in the following (informal) theorem:

**Theorem 1** (VQAs as WHRFs, informal). \textit{Consider the class of ansatzes}

\[ |\theta\rangle = \prod_{i=1}^{q} U_i(\theta) |\psi_0\rangle = \prod_{i=1}^{q} e^{-i\theta_i Q_i} |\psi_0\rangle, \tag{3} \]

where each \( Q_i \) is drawn uniformly from the Pauli group \( \mathbb{P}_n \) and \( |\psi_0\rangle \) is a uniformly random stabilizer state. Let \( p \) be the number of distinct \( \theta_i \), and let \( r = q/p \). Under reasonable assumptions on the eigenvalues of \( H \) (with ground state energy \( \lambda_1 \)), the random variational objective function

\[ F_H(\theta) = \frac{\langle \theta | H | \theta \rangle - \lambda_1}{2^{-n} \| H - \lambda_1 \|_*} \tag{4} \]

converges in distribution to the random field

\[ F_{\text{WHRF}}(\theta) = \sum_{i_1,\ldots,i_r,w_1',\ldots,w_r'} w_{i_1} \cdots w_{i_r} J_{i_1,\ldots,i_r,w_1',\ldots,w_r'} \tag{5} \]

where \( w \) are points on the hypertorus \( (S^1)^{\times p} \) parameterized by \( \theta \) and \( J \) is a complex Wishart random matrix normalized by its number of degrees of freedom. \( \| \cdot \|_* \) denotes the nuclear norm of \( \cdot \).

For the class of ansatzes we consider and for physically relevant \( H \), \( J \) has \( m = \Theta(2^n) \) degrees of freedom. We believe that ansatzes allowed to depend on the problem Hamiltonian are able to bring the effective degrees of freedom \( m \) to grow polynomially with the number of qubits \( n \). We give a heuristic argument for this in Sec. III C and numerical evidence in Sec. IV B and leave a detailed study of this for future work. A more detailed treatment of this mapping is given in Sec. IV B with a discussion of the various assumptions involved given in Sec. III C.

2. **The Loss Landscape of Wishart Hypertoroidal Random Fields**

Once mapped to a standard random field on a compact manifold, we can utilize results from Morse theory to find the expected distribution of critical points of the loss function Eq. (5) (and therefore those of Eq. (4)) by studying the joint distribution of the loss function, its gradient, and its Hessian. In the classical machine learning case, these random fields are all Gaussian. This follows from the fact that sums of the independent Gaussian coefficients in Eq. (2) are also Gaussian. However, the entries of \( J \) in Eq. (5) are not independent—this is easily seen as \( J \) is positive definite. Luckily, through explicit calculation we are able to show that the Hessian (conditioned on being at a critical point) takes the simple form of the sum of a Wishart matrix and an independent Gaussian matrix. The loss and gradient distributions take on similarly simple forms. These calculations are performed in Sec. III B.

With the joint distribution of the loss function and its derivatives in hand, we are able to use results from Morse theory to find the expected distribution of critical points of the loss function Eq. (5) at various energies \( E \) (in units of the mean eigenvalue \( 2^{-n} \| H - \lambda_1 \|_* \)). Though the results are unwieldy—involving an expectation over the eigenvalues of the sum \( C \) of independent Wishart and Gaussian matrices—they are exact. Furthermore, the exact formula allows us to probe the expected distribution of critical points of various indices \( k \), where \( k \) labels the number of negative eigenvalues of the Hessian at the critical point (i.e. \( k = 0 \) probes local minima). The informal result is as follows:

**Theorem 2** (WHRF eigenvalue distribution, informal). \textit{Let \( \mu_{C(E)} \) be the eigenvalue distribution of a random matrix \( C(E) \) drawn from a certain distribution of random matrices dependent on \( E \). Then, the expected number of critical points of index \( k \) at an energy \( E \) is}

\[ E[\text{Crt}_k (E)] = \left( \frac{\pi}{r} \right)^{\frac{r}{2}} \Gamma(m)^{-1} m^{(1+\gamma)m} \mathbb{E}_{C(E)} \left[ e^{p \ln(|\lambda - 2rE|)} \mathbb{1} \left\{ \lambda_{k+1}^C \geq 2rE \right\} \right] E^{(1-\gamma)m-1} e^{-mE}, \tag{6} \]

where

\[ \gamma = \frac{p}{2m} \tag{7} \]

and \( \lambda_i^C \) is the \( i \)th smallest eigenvalue of \( C \).
The precise statement of this theorem is given in Sec. III C along with the distribution from which $C(E)$ is drawn. We call $\gamma$ in Eq. (7) the overparameterization factor; it is the (scaled) ratio of the number of independent parameters of the loss function to the number of degrees of freedom $m$ of the Wishart random field of Eq. (5). As discussed in Sec. III B, $m$ is generically exponential in $n$, so unless the ansatz has exponentially many parameters (or unless the ansatz is structured, i.e. not the generic randomized ansatz we consider here) $\gamma$ is very small.

3. Asymptotic Limits of the Critical Point Distribution

Though Eq. (6) gives the exact distribution of critical points, it is difficult to use in practice. As mentioned in Sec. III A this difficulty comes from the expectation over eigenvalues of the sum of independent Wishart and Gaussian matrices. Surprisingly, however, the eigenvalues of both Wishart and Gaussian orthogonal matrices converge in distribution to fixed distributions. Roughly, asymptotically in the size of the matrix, the eigenvalue distribution of all normalized Wishart matrices are the same (given by the Marchenko–Pastur distribution) and the eigenvalue distribution of all Gaussian orthogonal matrices are the same (given by the Wigner semicircle distribution). Putting aside deviations from this convergence for the moment, an asymptotic treatment of Eq. (6) can be given when considering the asymptotic behavior of the eigenvalue distribution of the sum of these matrices.

Luckily, we can characterize the asymptotic distribution of eigenvalues of the sum well using the tools of free probability theory. Roughly, free probability theory is the probability theory of noncommutative random variables (e.g. random matrices). As the distribution of the sum of two random variables in commutative probability theory can be described by the convolution of the distributions of the two independent random variables, so can the free convolution of the distributions of two freely independent noncommutative random variables. Using the asymptotic free independence of Wishart and Gaussian orthogonal random variables, we are able to show that asymptotically the eigenvalue distribution of their sum weakly converges to the free convolution of a Marchenko–Pastur distribution with a semicircle distribution.

However, weak convergence is not enough; due to the exponential factor in the expectation in Eq. (6), any large deviations from the asymptotic convergence—even if they occur with exponentially vanishing probability—can potentially cause large deviations from the naive application of free probability theory. In Sec. III D 3 we are able to bound the probability of these large deviations, and show that unlike the Gaussian case [16] to (logarithmic) leading order they do not contribute to the final result. This is due to the contribution to the expectation from the deviations being dominated by what is predicted by free probability theory. These results are given in Sec. III D 3.

Armed with an asymptotic expression for the distribution of critical points, we specialize to two limits: $p \geq 2m$ (i.e. the overparameterized regime) and $p \ll m$ (i.e. the underparameterized regime). In the former, we show to leading multiplicative order in $p \gg 1$ that all local minima are located at the global minimum. Though the ansatz differs, we believe a similar phenomenon gives rise to the phase transition in trainability observed in [9][12]; [10] also numerically supports our heuristic and numeric arguments that particular ansatzes specialized to the problem objective observable $O$ have $m$ grow at worst polynomially in the number of qubits $n$. In the underparameterized regime, we show that the density of local minima approximately follows a compound confluent hypergeometric (CCH) distribution [19]

$$f_{CCH}(E | \gamma, m) \propto \left(e^{-E}E^{1-\gamma}(1-2E)^{\gamma}\right)^m,$$

which has a width $\sim m^{-1}$ and is centered near $E = \frac{1}{4} - \gamma$ (i.e. near half of the mean eigenvalue of the objective observable when units are restored). These results are given in Sec. III E. Finally, we are able to numerically confirm our results, and also numerically show that these results hold not only in expectation over ansatzes, but also for each ansatz individually; these results are given in Sec. IV.

C. Implications of Results

For a problem Hamiltonian on $n$ qubits, typically the number of distinct parameters $p$ of the VQA is $O(\text{poly}(n))$ whereas $m$ as defined in Theorem 1 is typically $\sim \exp(n)$. Returning the overall multiplicative factor of $2^{-n}\| H - \lambda_1 \|_*$, Eq. (5) tells us that in this limit the distribution of local minima are exponentially concentrated near half the middle of the spectrum of the problem Hamiltonian $H - \lambda_1$. That is, even if the number of parameters were to grow polynomially with the number of qubits of the problem instance, almost all local minima (in expectation) are far in energy from the global minimum.

---

2 I.e. up to shifts on the order of $\sqrt{\gamma}$ and subleading multiplicative corrections; see Sec. III E 2 for more details.
Though these results are pessimistic, we emphasize here that our results only apply to our specially constructed random class of ansatzes. In Sec. II C we give a heuristic explanation as to why purpose-built ansatzes (i.e. ansatzes that depend on the Hamiltonian) may effectively make \( m = \text{O} (\text{poly} (n)) \) and thereby \( \gamma \) large even for modest \( p \), explaining the good scaling behavior observed numerically in \cite{10} and in our Sec. IV B when looking at the HVA.

We proceed as follows. In Sec. II we give some background on VQAs and show how a certain class of VQAs with randomized ansatzes can be mapped to Wishart random fields on the hypertorus (WHRFs). Then, in Sec. III we give some background on the Kac–Rice formula, and analytically find the distribution of critical points for WHRFs. As the formula at finite \( p \) is unwieldy, we also compute the distribution of local minima when \( p \gg 1 \) and discuss various limits. In Sec. IV we numerically test our analytic predictions, and finally in Sec. V we conclude and discuss possible future directions of research.

II. VARIATIONAL QUANTUM ALGORITHMS AS RANDOM FIELDS

A. Variational Quantum Algorithms

Variational quantum algorithms (VQAs) are a class of algorithms where one expresses the solution of some problem as the smallest eigenvalue of an objective observable \( H \). Equivalently, assuming an ansatz

\[ |\theta\rangle = \prod_{i=1}^{q} U_i (\theta_i) |\psi_0\rangle \tag{9} \]

that for some choice \( \theta \) closely approximates the ground state of \( H \), the solution is encoded as the minimum of the objective function

\[ \tilde{F} (\theta) = \langle \theta | H | \theta \rangle \tag{10} \]

which can be computed on a quantum computer. For simplicity of analysis, throughout this paper we will consider the objective function

\[ F (\theta) = \langle \theta | H | \theta \rangle - \lambda_1, \tag{11} \]

where \( \lambda_1 \) is the ground state energy of \( H \); this has the same loss landscape as Eq. (10), but is minimized at \( F = 0 \) (assuming a sufficiently expressive \( |\theta\rangle \)). In Eq. (9), \( q \) is referred to as the depth of the circuit, and the initial state \( |\psi_0\rangle \) is fixed (and usually is assumed to be known classically, e.g. is the ground state of a mean-field approximation of \( H \)). Different choices of \( U_i \) constitute different choices of ansatz for the ground state of \( H \).

Ansatz design choice generally falls in one of two categories: Hamiltonian informed ansatzes, and Hamiltonian agnostic ansatzes. Examples of Hamiltonian informed ansatzes include the chemistry-inspired unitary coupled cluster ansatz \cite{2} and the adiabatically inspired quantum approximate optimization algorithm (QAOA) ansatz \cite{3}, known outside of the context of combinatorial optimization as the Hamiltonian variational ansatz (HVA) \cite{20}. These ansatzes depend solely on the problem objective Hamiltonian \( H \), and are usually physically motivated ansatzes which, in some limit, have convergence guarantees. Hamiltonian agnostic ansatzes, conversely, depend solely on the hardware the VQA is run on, and not at all on the problem objective \( H \). This class of ansatzes includes the hardware-efficient ansatz \cite{5}. These ansatzes are designed to eke out as much depth as possible in the objective ansatz \( |\theta\rangle \) by using \( U_i \) that can be easily performed on the given quantum device.

Though hardware-efficient ansatzes generally can be run at larger depth \( q \) than Hamiltonian informed ansatzes, the very generic nature of the ansatz circuit means this class of ansatz is more difficult to train, often encountering barren plateaus in the optimization landscape that are difficult to escape from \cite{6–8}. Heuristically, this can be understood as Hamiltonian agnostic objective functions being so expressive that it must explore essentially all of Hilbert space to find a local minimum, exponentially suppressing the gradients of the loss function \cite{21}.

In this work we consider a class of ansatzes that, like the hardware-efficient ansatz, is independent of the problem instance. In particular, we consider random parameterized ansatzes of the form:

\[ U_i \equiv e^{-i \theta_i Q_i} \tag{12} \]

for Pauli operators \( Q_i \), where each \( Q_i \) is drawn uniformly and independently from the \( n \)-qubit Pauli operators. Throughout this paper, we will use \( q \) to denote the total number of Pauli rotations in \( |\theta\rangle \) as in Eq. (9), \( p \) to denote the total number of independent parameters \( \theta_i \), and \( r_i \) to denote the number of Pauli rotations governed by a single independent parameter \( \theta_i \). For simplicity, we will assume \( r_i = r_j \equiv r \) for all \( i, j \), and thus take

\[ r \equiv \frac{q}{p} \tag{13} \]
to be a natural number.

B. Mapping Variational Quantum Algorithms to Random Wishart Fields

With the background of VQAs in place, we will now show the asymptotic (weak) equivalence of VQAs with the random choice of ansatz described in II A to Wishart random fields. Throughout this section, we will consider a problem Hamiltonian $H$ on $n$ qubits, with ground state energy $\lambda_1$ and mean eigenvalue $\bar{\lambda}$. We also define the degrees of freedom parameter

$$m \equiv \frac{\|H - \lambda_1\|_F^2}{\|H - \bar{\lambda}\|_F^2},$$

(14)

whose interpretation will be discussed in Sec. II C. Twice the degrees of freedom parameter $m$ will turn out to govern the location of the transition from the underparameterized to the overparameterized regime (see Sec. III), and for physically relevant Hamiltonians is expected to be exponential in $n$ (see Sec. II C). We will also consider the Pauli decomposition of the nontrivial part of $H$:

$$H - \bar{\lambda} = \sum_{i=1}^{A} \alpha_i R_i,$$

(15)

where $A$ is the number of terms in the Pauli decomposition and $\alpha$ the Pauli coefficients.

We begin by showing the convergence of a class of randomized VQAs to a weighted sum of Wishart random fields at a rate $\gg n$; the seemingly arbitrary shifts by the mean eigenvalue $\bar{\lambda}$ and the ground state energy $\lambda_1$ here will aid in future discussion, when we approximate the weighted sum of Wishart random fields with a single random field. The wide variety of assumptions will be discussed in detail in Sec. II C.

Theorem 3 (VQAs as RFs). Let $|\psi_0\rangle$ be an arbitrary stabilizer state on $n$ qubits. Fix a sequence of $q$ angles $\theta_i \in [-\pi, \pi]$ such that each $\theta_i$ is present $r$ times in the sequence. We let $p = \frac{q}{r}$ denote the number of distinct parameters. Select an ansatz

$$|\theta\rangle \equiv \prod_{i=1}^{q} U_i(\theta_i) |\psi_0\rangle \equiv \prod_{i=1}^{q} e^{\pm i \theta_i Q_i C} |\psi_0\rangle$$

(16)

by independently at random drawing each $\pm Q_i$ uniformly from the $n$-qubit Pauli group $\mathbb{P}_n$ and $C$ from the $n$-qubit Clifford group $\mathbb{C}_n$. Consider the scaled and shifted

$$\tilde{H} \equiv \frac{H - \bar{\lambda}}{\bar{\lambda} - \lambda_1} = \frac{H - \bar{\lambda}}{2^{-n} \|H - \lambda_1\|_*},$$

(17)

where $\|\cdot\|_*$ is the nuclear norm. Then, the random variational objective function

$$F_{VQA}(\theta) = \frac{\langle \theta | H | \theta \rangle - \lambda_1}{\bar{\lambda} - \lambda_1} = \frac{\langle \theta | H | \theta \rangle - \lambda_1}{2^{-n} \|H - \lambda_1\|_*},$$

(18)

has first two moments exponentially close in $n$ as $n \to \infty$ to those of

$$F_{XHX}(w) = 2^{-n} \left( \bigotimes_{i=1}^{p} w_i \right) \bigotimes_r X \cdot \tilde{H} \cdot X^\dagger \cdot \left( \bigotimes_{i=p+1}^{1} w_i \right) + 1,$$

(19)

where $w_i$ are points on the circle parameterized by $\theta_i$ and $X$ is a matrix of i.i.d. complex standard jointly normal random variables. Furthermore, assuming

$$\frac{\|\alpha\|_\infty}{\bar{\lambda} - \lambda_1} \leq f(n)^{-1}$$

(20)

for some $f(n) = \Omega(1)$, their distributions are bounded in Lévy distance by $\tilde{O} \left( \left( \frac{\log(A)f(n)n}{\bar{\lambda}} \right)^{-1} \right)$. 
Proof. The Feynman path integral representation of the objective function Eq. (18) is of the form

$$F_{\text{VQA}} = \sum_{\gamma, \gamma' \in \{0, 1\}^q} w_{\gamma} w_{\gamma'} \langle \psi_0 | C^\dagger Q_{\gamma} H Q_{\gamma'} C | \psi_0 \rangle + 1,$$

where $\gamma$ labels a term in the path integral expansion of $U$,

$$w_{\gamma} = \prod_{i=1}^{q} \begin{cases} \cos (\theta_i), & \text{if } \gamma_i = 0 \\ \sin (\theta_i), & \text{if } \gamma_i = 1 \end{cases}$$

is the amplitude, and

$$Q_{\gamma} = (-1)^{\| \gamma \|_0} \prod_{i=1}^{q} Q_i.$$

We can rewrite the Feynman path integral as

$$F_{\text{VQA}} = \left( \bigotimes_{i=1}^{p} w_i^T \right)^{\otimes r} \cdot \tilde{X} \cdot \tilde{H} \cdot \tilde{X}^\dagger \cdot \left( \bigotimes_{i=p+1}^{1} w_i \right)^{\otimes r} + 1,$$

where

$$w_i = \begin{pmatrix} \cos (\theta_i) \\ \sin (\theta_i) \end{pmatrix}$$

and $\tilde{X} \in \mathbb{C}^{2 \times 2^n}$ is a random matrix with rows

$$\left\langle \tilde{X} \right\rangle_{\gamma} \equiv \langle \psi_0 | C^\dagger Q_{\gamma} C | \psi_0 \rangle.$$

We will proceed as follows. First, we will bound the difference in the first two moments of Eq. (24) and its i.i.d. Haar equivalent, where the rows of $\tilde{X}$ are i.i.d. Haar random, to be exponentially small in $n$. As Haar random vectors have first three moments matching those of random Gaussian vectors (scaled by $2^{-\frac{1}{2}}$), this gives the desired convergence through second moments. Then, we will show that the characteristic functions at $x$ of Eq. (24) and its i.i.d. Haar random equivalent converge exponentially quickly in $n$ for small enough $x$, giving a convergence in distribution at a rate $\tilde{\Omega} \left( \frac{\lg (A) f(n \beta)}{\Lambda} \right)$ by [22]. Finally, convergence in distribution to Eq. (19) will follow as the error in higher-order moments between Haar random and scaled Gaussian vectors exponentially decays in $n$ by Borel’s lemma [23].

Obviously the first moment of Eq. (24) matches that of the i.i.d. Haar random case; off-diagonal entries in the path integral average to zero, and the diagonal entries are correct as $C$ is drawn from a unitary 2-design [24]. Let us now consider the second moments of the nontrivial parts of both, where we are concerned with terms of the form:

$$c_{\alpha \beta \mu \nu} = E \left[ \langle \psi_0 | C^\dagger Q_{\alpha}^\dagger R Q_{\gamma} C | \psi_0 \rangle \langle \psi_0 | C^\dagger Q_{\mu}^\dagger H Q_{\gamma} C | \psi_0 \rangle \right],$$

and how they differ from the i.i.d. Haar random equivalent

$$h_{\alpha \beta \mu \nu} = E \left[ \langle \psi_0 | U_{\alpha}^\dagger H U_{\beta} | \psi_0 \rangle \langle \psi_0 | U_{\mu}^\dagger H U_{\nu} | \psi_0 \rangle \right].$$

First, assume $\alpha = \beta = \mu = \nu$; as $C$ is drawn from a unitary 2-design [24], the terms are equal. Similarly, if

$$\gamma_{\alpha} \oplus \gamma_{\beta} \oplus \gamma_{\mu} \oplus \gamma_{\nu} \neq 0,$$

then both expectations are equal to zero; this is because $c_{\alpha \beta \mu \nu}$ must have an odd number of some $Q$, and $h_{\alpha \beta \mu \nu}$ an odd number of some $U$ (or $U^\dagger$).

Let us now consider when the above conditions are not satisfied. We consider simultaneously terms of the form

$$\left( \langle \psi_0 | C^\dagger Q_{\alpha} R Q_{\beta} C | \psi_0 \rangle \gamma_{\alpha j} \leftrightarrow \gamma_{\beta j} \right) \left( \langle \psi_0 | C^\dagger Q_{\mu} R Q_{\nu} C | \psi_0 \rangle \gamma_{\mu j} \leftrightarrow \gamma_{\nu j} \right),$$

i.e. all terms summed where unequal components of $\gamma_{\alpha}$ and $\gamma_{\beta}$ (and $\gamma_{\mu}$ and $\gamma_{\nu}$) are swapped. Note that the parity of the permutation determines the sign of the term in Eq. (21) (and thus in Eq. (30)). Here, $R$ and $R'$ are terms in
the Pauli expansion of $\tilde{H}$. Consider the largest $j$ where $\gamma_\alpha$ and $\gamma_\beta$ differ; consider the sum of each pair of terms in Eq. (30) that have component $j$ permuted, but are equal at all $k < j$. Each pair of terms is of the form (with relative signs made explicit)

$$\langle \psi_0 | C^i A Q_j A'R B' B C | \psi_0 \rangle - \langle \psi_0 | C^i A A'R'Q_j B C | \psi_0 \rangle = 2 \langle \psi_0 | C^i A Q_j A'R B' B C | \psi_0 \rangle 1_{(Q_j,A'R B') \neq 0}$$

(31)

for some $A, A', B, B'$. For all $Q_j$ that commute with $A'R B'$, the two terms cancel. In particular, $Q_j A'R B'$ cannot be proportional to the identity. As $\tilde{H}$ is traceless, both $R$ and $R'$ are also not proportional to the identity. This can be done inductively for all $j$ where $\gamma_\alpha$ and $\gamma_\beta$ differ.

Consider the case where $\gamma_\alpha + \gamma_\beta \neq \gamma_\mu + \gamma_\nu$; we must have that $\gamma_\alpha$ and $\gamma_\beta$ have a coordinate $i$ where they are both one, and where $\gamma_\mu$ and $\gamma_\nu$ are both zero (assuming Eq. (29) is not satisfied). By Eq. (31), WLOG we can consider the product of Pauli observables between the two $Q_i$ as being not proportional to the identity. Then, averaging over $Q_i$ will yield zero. This is the same as the i.i.d. Haar random case, as every term in the expansion of Eq. (30) must have only one of some unitary when $\gamma_\alpha + \gamma_\beta \neq \gamma_\mu + \gamma_\nu$.

Finally, consider the case where $\gamma_\alpha + \gamma_\beta = \gamma_\mu + \gamma_\nu$. Under this constraint, we must have the same number of terms in each sum in Eq. (50); we call this number of terms $2^n$. In the Pauli case, every time we combine terms as in Eq. (31) introduces an overall factor of 4, and we average only over the anticommuting Pauli operators. As the value of the expectation over $C$ is independent of the (nonidentity) Pauli in the expectation value, this introduces a factor of $\frac{1}{2}$ every time we combine terms. This gives

$$2^n E_{C \sim \mathcal{C}_n} \left[ \langle \psi_0 | C^i S | \psi_0 \rangle \langle \psi_0 | C^i S' | \psi_0 \rangle \right],$$

(32)

for some $S$ and $S'$ that are equal if and only if $R = R'$. Similarly, in the i.i.d. Haar random case, only products of terms with $\gamma_\alpha = \gamma_\mu$ and $\gamma_\beta = \gamma_\nu$ are homogeneous in their unitaries and give nonzero expectations, yielding

$$2^n E_{U \sim \mathcal{C}_n} \left[ \langle \psi_0 | U^i_\alpha U^\dagger R U_\beta | \psi_0 \rangle \langle \psi_0 | U^\dagger R' U_\alpha | \psi_0 \rangle \right].$$

(33)

If $R \neq R'$ (and $S \neq S'$), these are both zero. If $R = R'$ (and $S = S'$), the latter is equal to $2^{2n}$ and the former to $2^{2n} (1 + O(2^{-n}))$. Putting everything together and explicitly writing the overall factor of $\frac{2^n}{\|H - \lambda_1\|_2}$, we have that the error in the second moment is on the order of

$$\epsilon_2 = \frac{2^{2n}}{\|H - \lambda_1\|_2^2} \left( \sum_{i=1}^{A} \alpha_i^2 \right)^2,$$

(34)

where $\alpha_i$ are the coefficients of the Pauli expansion of $H - \tilde{X}$. We also have that

$$\sum_{i=1}^{A} \alpha_i^2 = 2^{-n} \|H - \tilde{X}\|_p^2 = m^{-1} 2^{-n} \|H - \lambda_1\|_2^2,$$

(35)

where $m$ is defined as in Eq. (14). Thus,

$$\epsilon_2 = 2^{-(2n + \lg(m))}.$$

(36)

Let us now consider the $t$th moment for $t \geq 3$. We will bound the higher moments of both models, and show that their characteristic functions have infinite radii of convergence. Then, by showing that the difference in these characteristic functions vanishes exponentially in $n$ for all $x \geq 0$ bounded below $\frac{\lg(A)f(n)n}{A}$, we will show that the two models converge in distribution at a rate $\tilde{\Omega} \left( \frac{\lg(A)f(n)n}{A} \right)$.

By grouping terms as in Eq. (50), it is sufficient to only bound

$$b_t = (\tilde{X} - \lambda_1)^{-t} E_{C \sim \mathcal{C}_n} \left[ \prod_{i=1}^{t} \left( \sum_{j=1}^{A} \alpha_j \langle \psi_0 | C^i S_{ij} C | \psi_0 \rangle \right) \right],$$

(37)

where $S_{ij}$ is not proportional to the identity, $S_{ij} \neq S_{i'j}$ for all $i \neq i'$, and $A$ is the number of terms in the Pauli decomposition of $\tilde{H}$. If a term in the expansion of Eq. (37) contains two $S_{ij}$ that anticommute, the contribution to the moment from that term is zero as $C |\psi_0\rangle$ is a stabilizer state for all $C$. Generally, the contribution to the moment
is maximized when the $S_{ij}$ are “maximally dependent”—that is, for $d$ distinct $S_{ij}$ in a term, the contribution to the moment is maximized when the $S_{ij}$ are generated by a cardinality $\lfloor \log (d) + 1 \rfloor$ subset of them. Thus, the contribution to the moment is bounded by $2^{-\epsilon \lfloor \log (d) + 1 \rfloor} n$ for some constant $c$ \[25\]. Note that this also bounds the i.i.d. Haar random case. Putting everything together and using the multinomial theorem, the $t$th moment of the nontrivial part of both distributions is bounded by

$$b_t \leq \sum_{\text{cyc.}} 2^{-c \lfloor \log (\|k\|_0) + 1 \rfloor} n \left( \frac{\alpha_t}{\lambda - \lambda_1} \right)^k.$$ \hfill (38)

This corresponds to the case where $S_{ij} = S_{ij'} \equiv S_i$, i.e. when there is maximal dependence between the matrix elements. Here, $k_i$ indexes how many times $S_i$ appears in a term in Eq. (37), and $\|\cdot\|_0$ denotes the number of nonzero coordinates of $\cdot$. By Eq. (20), as $t \to \infty$ for any given $A$ and $n$,

$$\frac{b_t}{t!} \leq (1 + o(1)) 2^{-\epsilon \log (t) - \frac{1}{2} \log (2\pi t) + \epsilon \log (\frac{\alpha_n}{\lambda - \lambda_1})} \cdot c \log (A)n.$$ \hfill (39)

Thus, the Taylor series of the characteristic functions of both distributions have infinite radii of convergence, and both are completely determined by their moments. Furthermore, Eq. (39) gives us that the difference in their characteristic functions at 0 are bounded in Lévy distance by

$$A T \frac{\log (\|k\|_0)}{\lambda} \cdot c \log (A)n.$$ \hfill (40)

Theorem 4 (XH RFs as WHRFs). The random field given by Eq. (19) has first two moments equal to the Wishart hypertoroidal random field (WHRF)

$$F_{\text{WHRF}} (\theta) = m^{-1} \sum_{\text{cyc.}} \frac{w_{i_1} \ldots w_{i_r} J_{i_1 \ldots \ldots i_r} J_{i_1' \ldots \ldots i_r'} w_{i_1'} \ldots w_{i_r'}}{w_{i_1} \ldots w_{i_r} J_{i_1 \ldots \ldots i_r} J_{i_1' \ldots \ldots i_r'} w_{i_1'} \ldots w_{i_r'}}.$$ \hfill (41)

where $J \sim \text{CW}_{2n} (m, I_{2n})$ is a complex Wishart random matrix and the effective degrees of freedom defined in Eq. (14) is formally a real number, but can be rounded to the nearest natural number with negligible error. Furthermore, assuming the largest eigenvalue of $H$ as defined in Eq. (17) is at most $2^m$ for some constant $c$ bounded below 1, their distributions are bounded in Lévy distance by $2^{-\Omega(\min(m, \log (m)))}$.\hfill (42)

Proof. By the unitary invariance of random matrices with Gaussian entries, by diagonalizing $\tilde{H}$ we can rewrite $F_{\text{XH}}$ as the random field

$$F_{\text{XH}} (w) = \|H - \lambda_1\|_2^{-1} \left( \bigotimes_{i=1}^p w_i \right) \cdot \sum_{\text{cyc.}} \left( \frac{1}{\lambda - \lambda_1} \right) \cdot \left( \bigotimes_{i=p}^n \right),$$ \hfill (43)

where $X_i$ is the $i$th column of $X$ and $h_i$ are the eigenvalues of $H$. The sum over Kronecker products of columns is just the weighted sum of (at most) $2^n$ independent Wishart random variables, each with a single degree of freedom. It is known \[28, 30\] that the first two moments of this weighted sum of independent Wishart random variables is equal (up to rounding of the degrees of freedom) to that of the single Wishart random variable

$$J \sim \text{CW}_{2n} (m, m^{-1} I_{2n}).$$ \hfill (44)

Let us now consider higher moments of both distributions. A useful property of both $F_{\text{XH}}$ and $F_{\text{WHRF}}$ is that they are invariant under rotations on the hypertorus $w \mapsto O \cdot w$ (for real orthogonal $O \in \text{SO} (2)^\otimes p$) due to the invariance of the Wishart distribution under orthogonal transformations \[31\]. Due to this property, we will often take

$$w = n \equiv (1, 0, \ldots, 0)^T.$$ \hfill (45)
i.e. perform calculations at a fixed point $\theta = 0$ on the hypertorus. For instance, by inspection of the marginal distributions of the elements of $X \otimes X^\dagger$ and $J$ \cite{32,33}, we immediately see that

$$
\left(\bigotimes_{i=1}^{p} w_i\right)^{\otimes r} \cdot \left(X \otimes X^\dagger\right) \cdot \left(\bigotimes_{i=p+1}^{1} w_i\right)^{\otimes r} \sim \Gamma(1, 1)
$$

and

$$
F_{\text{WHRF}}(w) \sim m^{-1} J_{(1,\ldots,1),(1,\ldots,1)} \sim \Gamma(m, m^{-1});
$$

here, $\Gamma(k, \theta)$ is a gamma distributed random variable with shape $k$ and scale $\theta$. We therefore have that the moment-generating function for $F_{\text{XHX}}(w)$ is

$$
M_{\text{XHX}}(x) = e^x \prod_{i=1}^{2^n} \left(1 - \frac{h_i - X}{\|H - \lambda_i\|} x\right)^{-1} = e^x \det \left(1 - 2^{-n} \hat{H} x\right)^{-1}
$$

and for $F_{\text{WHRF}}(w)$ is

$$
M_{\text{WHRF}}(x) = \left(1 - \frac{x}{m}\right)^{-m}.
$$

Assuming the largest eigenvalue of $\hat{H}$ is at most $2^m$, we see that these moment generating functions differ at any given $0 \leq x < 2^{\min((1-c)n, Jg(m))}$ by at most $O(2^{-3(1-c)n x^3 + m^{-3} x^3})$. As the two distributions have equal first and second moments, it can then be shown \cite{[22]} that the error in Lévy distance between the two is bounded by $2^{-\Omega(\min(n, Jg(m)))}$.

Combining the two theorems, we roughly see that under reasonable assumptions on the spectrum of $H$ the random fields induced by the specific class of VQAs we consider can be approximated by WHRFs up to an error on the order of $\tilde{O}\left(\frac{\lg(A)f(n)m}{A}^{-1} + m^{-1}\right)$ as $m, n \to \infty$.

C. Discussion of the Mapping

Let us now briefly discuss the intuition and assumptions behind the results proved in Sec. \ref{sec:results} beginning with the random class of ansatzes we consider. Of course, in practice, VQA ansatzes are not chosen at random. Indeed, VQA ansatzes have a layered structure that precludes any independence between layers even if the layers were randomly chosen. Though this randomness assumption is strong, heuristically deep enough circuits (that are independent of the problem Hamiltonian) will still look roughly uniform over stabilizer states in the Feynman path integral expansion performed in the proof of Theorem \ref{thm:main} giving qualitatively similar results. Furthermore, though throughout this paper we consider results in expectation over this distribution of ansatzes, we find numerically in Sec. \ref{sec:numerics} that our analytic results seem to also hold in distribution; we therefore suspect that our analytic results in Sec. \ref{sec:analytic} hold more generally for individual ansatzes that are independent of the problem Hamiltonian.

Given the randomized class of ansatzes, in Theorem \ref{thm:main} we show that the VQA loss function is close in distribution to that of the random field given in Eq. \ref{eq:random} (the “XHX” model). Intuitively, this just stems from the fact that different paths in the Feynman path integral are matrix elements in uniformly random stabilizer states. We then show that the error induced in higher moments by taking each of these paths to be independent vanishes as $n \to \infty$. To prove this formally, we rely on the boundedness of Eq. \ref{eq:bdd} to bound higher moments of the distribution. Luckily, in practice this bound holds; for extensive Hamiltonians, one expects $f(n) \gtrsim n$.

Theorem \ref{thm:whrf} extends Theorem \ref{thm:main} by showing that the XHX model can be written as a sum of Wishart models weighted by the (scaled and shifted) eigenvalues of $H$, which can then be approximated by a single Wishart model. Heuristically, one can think of complex Wishart matrices as multidimensional generalizations of the gamma distribution; then, the approximation used in Theorem \ref{thm:whrf} is just a multidimensional generalization of the Welch–Satterthwaite approximation \cite{36,37}. This approximation (in both the univariate and multivariate cases) is exact in the first two moments of the distribution when the effective degrees of freedom $m$ given in Eq. \ref{eq:eff} is allowed to be real. In practice, $m$ is rounded to the nearest natural number, inducing a slight error in the approximation. Generally, errors in higher moments in the Welch–Satterthwaite approximation may be large when the moments of the approximated distribution is large, particularly when the coefficients of the sum can have arbitrary sign and are at different scales \cite{36,34}. 
However, for physically relevant Hamiltonians, the spectral radius is much smaller than $2^n$, and the coefficients of the sum are approximately equal. We show that under such conditions, errors in the moment generating functions vanish as $m, n \to \infty$ at the given rate.

The effective degrees of freedom $m$ as in Eq. (14) can be interpreted as roughly a signal-to-noise ratio of the mean eigenvalue of $H - \lambda_1$, and generically is exponentially large in $n$ (for small eigenvalue spacings). We will show in Sec. III E 1 that $m$ sharply dictates the variational loss landscape: for a number of independent parameters $p \geq 2m$, local minima concentrate near the global minimum. Conversely, for $p$ bounded below $2m$, local minima concentrate far away from the global minimum. This would imply that for the class of randomized ansatz we consider here, training large instances is infeasible. However, consider an ansatz that is allowed to depend on the problem instance $H$, such as in the Hamiltonian variational ansatz (HVA) [20]. With a clever enough ansatz, one can in principle “reweigh” the coefficients of Eq. (41) by having a nonuniform distribution over stabilizer states in the Feynman path integral expansion of Eq. (24), effectively making $m$ smaller. This would be consistent with what was numerically investigated in prior work [10] (and in Sec. IV B), where it was shown that even for a modest number of parameters the distribution of local minima concentrate near the global minimum for the HVA. We leave further investigation in this direction to future work.

Finally, we note that all of our asymptotic equivalence results so far have been shown to converge at a rate

$$\rho \equiv \log (A) f (n) n / A,$$

which is typically $\gtrsim \log (n)$ for physically relevant (i.e. two-local with arbitrary range, molecular in the plane wave dual basis [35], etc.) Hamiltonians. In Sec. III D we give the loss landscape of WHRFs (Eq. (40)) as $p, m \to \infty$, taking into account large deviations in $p$. If $p$ grows as $\Omega (\log (\rho))$, then in principle uncontrolled large deviations in the convergence of VQAs to WHRFs will dominate the asymptotics of the landscape (Eq. (91)). In particular, with probability $\sim \rho^{-1}$, deviations of the eigenvalues of the Hessian on the order of the eigenvalues themselves can occur, which are then “blown up” by a factor exponentially large in $p$ if all deviations constructively interfere. Thus, though Eq. (91) holds for WHRFs, it does not necessarily hold for VQAs when $p = \Omega (\log (\rho))$. If the deviations of eigenvalues of the Hessian due to the mapping from VQAs to WHRFs are roughly independent between eigenvalues, however, then these deviations are further exponentially suppressed in $p$, and the result holds independently of how $p$ scales with $n$. We believe in practice this is what occurs, and see numerically in Sec. IV A that our analytic results hold well even when $p \gg n$.

III. THE LOSS LANDSCAPE OF WISHART HYPERTOROIDAL RANDOM FIELDS

A. The Kac–Rice Formula

Our strategy for showing the distribution of critical points of Eq. (40) will closely follow that in [16], where similar results were shown for Gaussian spherical random fields. Namely, we will lean heavily on the Kac–Rice formula, which gives the expected number of critical points of a certain index at a given range of energies for random fields on manifolds. In this section we will consider a generic WHRF of the form

$$H_{\text{WHRF}} (w) = m^{-1} \sum_{i_1, \ldots, i_r, i'_1, \ldots, i'_r = 1}^{2^p} w_{i_1} \ldots w_{i_r} J_{i_1, \ldots, i_r, i'_1, \ldots, i'_r} w_{i'_1} \ldots w_{i'_r},$$

(49)

where once again $J \sim \mathcal{CW}_{2p} (m, I_{2p})$. We give an informal description of the Kac–Rice formula here, with the formal version given in Appendix A.

Lemma 1 (Kac–Rice formula, informal). Let $M$ be a compact, oriented manifold. Assume a random field $H (\sigma)$ is sufficiently nice. Then, the number of critical points of index at most $k$ with $H (\sigma) \in B$ for an open set $B \subset \mathbb{R}$ is

$$\mathbb{E} \left[ \text{Crt}_k (B) \right] = \int_M \mathbb{E} \left[ |\det (\nabla^2 H (\sigma))| \right] 1 \{ H (\sigma) \in B \} 1 \{ \iota (\nabla^2 H (\sigma)) \leq k \} \, p_{\sigma} (\nabla H (\sigma) = 0) \, d\sigma,$$

(50)

where $\nabla \cdot$ is the covariant gradient, $\iota (\cdot)$ is the index of $\cdot$, $p_{\sigma}$ is the probability density of $\nabla H (\sigma)$ at $\sigma$, and $d\sigma$ is the volume element on $M$.

From Lemma 1 we see that when the joint distribution of $\nabla^2 H, \nabla H$, and $H$ is known, then the expected number of critical points with energies in an open set $B$ can be calculated. We spend Sec. III B computing this joint distribution for the WHRF $H_{\text{WHRF}}$, and Sec. III C using the Kac–Rice formula to compute the expected number of critical points given this distribution.
B. The Joint Distribution of $H_{ \text{WHRF}}$ and its Derivatives

Now, we calculate the distribution of the Hessian when given the function value and that the covariant gradient is zero. In the course of the proof we will also show the gradient distribution when conditioned on the function value. We will heavily lean on the rotational invariance property of the distribution discussed in the proof of Theorem \[4\] with $\mathbf{n}$ once again the fixed point with all $\theta_i = 0$. Note that for the given embedding of the hypertorus into $\mathbb{R}^{2p}$, the Christoffel symbols are zero (i.e. we are considering the Euclidean hypertorus) and thus for the most part we can ignore the distinction between covariant and normal derivatives. Here, we choose local coordinates $\theta$ such that:

$$w_i = \begin{pmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{pmatrix}.$$ \hfill (51)

Perhaps surprisingly, we will find that conditioned on being at a critical point at a specified energy, the Hessian takes the simple form of a normalized and shifted Wishart matrix summed with a normalized GOE matrix.

**Lemma 2** (Hessian distribution). The scaled Hessian $m \partial_i \partial_j H_{ \text{WHRF}}(w)$ conditioned on $H_{ \text{WHRF}}(w) = x$ and $\partial_k H_{ \text{WHRF}}(w) = 0$ is distributed as

$$m \tilde{C}_{ij}(x) = -2r \delta_{ij} + rW_{ij} + r\sqrt{2m}N_{ij},$$ \hfill (52)

where $W \sim W_p(2m, I_p)$ and $N \sim \text{GOE}_p$ are independent.

**Proof.** Without loss of generality we take $w = \mathbf{n}$. Let $i \in \{1, 2\}^p$ be the vector with the $i$th component equal to 2 and all others equal to 1, $(i, j)$ similar with both the $i$th and $j$th component, and $b$ the vector with all components equal to 1. Taking derivatives explicitly yields

$$m \partial_i H_{ \text{WHRF}}(\mathbf{n}) = 2 \text{Re} \left\{ J_{(i, b, \ldots, b), (b, \ldots, b)} \right\} + \ldots + 2 \text{Re} \left\{ J_{(b, \ldots, b, i), (b, \ldots, b)} \right\}$$ \hfill (53)

and

$$m \partial_i \partial_j H_{ \text{WHRF}}(\mathbf{n}) = -2r \delta_{ij} + 2 \text{Re} \left\{ J_{(i, b, \ldots, b), (j, b, \ldots, b)} \right\} + 2 \text{Re} \left\{ J_{(i, b, \ldots, b), (j, b, \ldots, b)} \right\} + \ldots + 2 \text{Re} \left\{ J_{(b, \ldots, b, i), (b, \ldots, b)} \right\} + 2 \text{Re} \left\{ J_{(i, j, b, \ldots, b), (b, \ldots, b)} \right\} + \ldots + 2 \text{Re} \left\{ J_{(b, \ldots, b, (i, j), b, \ldots, b)} \right\}.$$ \hfill (54)

As $J$ is a Wishart matrix with identity scale matrix, it can be written as $X \cdot X^\dagger$ for $X$ a $2^q \times m$ matrix with i.i.d. standard complex normal entries. By performing an LQ decomposition of $X$, one can then by inspection determine the distributions of the entries of $J$ \[32, 33\]. For ease of notation, we let $\tau : \{1, 2\}^q \rightarrow \{1, \ldots, 2^q\}$ be a mapping between representations of the indices of $J$, with the convention $\tau((i, \ldots, 2)) = 1$. We then find (taking $\tau((i, \ldots, 2)) < \tau((j, \ldots, 2))$ WLOG) that

$$2J_{(b, \ldots, b), (b, \ldots, b)} = 2m H_{ \text{WHRF}}(\mathbf{n}),$$ \hfill (55)

$$2 \text{Re} \left\{ J_{(i, b, \ldots, b), (b, \ldots, b)} \right\} = \sqrt{2m} H_{ \text{WHRF}}(\mathbf{n}) \tilde{M}_{b, \ldots, b};$$ \hfill (56)

and, for $\tau((i, \ldots, b)) \leq m$,

$$\text{Re} \left\{ J_{(i, \ldots, b), (j, \ldots, b)} \right\} = \sqrt{2} \Gamma_{(i, \ldots, b)} \tilde{M}_{(i, \ldots, b), (j, \ldots, b)} + \sum_{\mu=1}^{(i, \ldots, b)} \sum_{\mu=1}^{(i, \ldots, b)} \tilde{M}_{(i, \ldots, b)} \tilde{M}_{(i, \ldots, b)}$$ \hfill (58)

and otherwise

$$\text{Re} \left\{ J_{(i, \ldots, b), (j, \ldots, b)} \right\} = \sum_{\mu=1}^{m} \sum_{\mu=1}^{m} \tilde{M}_{(i, \ldots, b)} \tilde{M}_{(i, \ldots, b)} + \sum_{\mu=1}^{m} \sum_{\mu=1}^{m} \tilde{M}_{(i, \ldots, b)} \tilde{M}_{(i, \ldots, b)}.$$ \hfill (59)

Here, $\tilde{M}$ and $\tilde{M}$ are symmetric with off-diagonal entries i.i.d. drawn from the standard normal distribution, and $\Gamma^{-1}(\mu) \equiv M_{(i, \ldots, b), (j, \ldots, b)}$ has entries i.i.d. drawn from $\Gamma(m - \mu + 1, 1)$. Note that each $\sqrt{2 \tilde{M}}$ is chi-square distributed
with $2(m - \mu + 1)$ degrees of freedom; therefore, Eqs. (58) and (59) can be considered as elements of a real Wishart matrix $\hat{W}$ with $2m$ degrees of freedom. Also, note that Eqs. (58) and (59) are independent of $\partial_k H_{\text{WHRF}}(n)$ when conditioned on $H_{\text{WHRF}}(n) \equiv 0$. If $x \neq 0$, the condition $\partial_k H_{\text{WHRF}}(n) = 0$ is equivalent to taking each sum over the elements of $M$ from $\mu = 2$ instead of $\mu = 1$, which is equivalent to taking the convention $\tau ((b, \ldots, b)) = 2\gamma$ and shifting the indices of $M$ and $\hat{M}$. Therefore, the (scaled) Hessian conditioned on $H_{\text{WHRF}}(n) = x$ and $\partial_k H_{\text{WHRF}}(n) = 0$ is distributed as

$$m\hat{C}_{ij}(x) = -2mx\delta_{ij} + \left(O \cdot \hat{W} \cdot O^T\right)_{ij} + r\sqrt{2mx}N_{ij};$$

(60)

here, $N \sim \text{GOE}_p$ (with the convention that diagonal entries $\sim \mathcal{N}(0,2)$ and off-diagonal entries $\sim \mathcal{N}(0,1)$), and $O$ is a matrix such that $O_{ij} = 1$ if and only if $\tau^{-1}(\mu)$ is of the form $(b, \ldots, i, \ldots, b)$, and is otherwise equal to 0. The invariance of the Wishart distribution under orthogonal transformations and partitioning \cite{31,32} leads to the final result.

C. The Exact Distribution of Critical Points

Given the joint distribution of $H_{\text{WHRF}}$, its gradient, and its Hessian, we are now equipped to calculate the expected number of critical points of a given index $k$ using the Kac–Rice formula (Lemma 1).

**Theorem 5** (Distribution of critical points in WHRFs). Let

$$\mu_{C(x)} = \frac{1}{p} \sum_{i=1}^{p} \delta \left(\lambda_{C(x)}^i\right)$$

be the empirical spectral measure of the random matrix

$$C(x) = \frac{r}{m} \left(W + \sqrt{2mx}N\right),$$

where $W \sim \mathcal{W}_p(2m, I_p)$ and $N \sim \text{GOE}_p$ are independent and $\lambda_{C(x)}^i$ is the $i$th smallest eigenvalue of $C(x)$. Then, the distribution of the expected number of critical points of index $k$ at an energy $E > 0$ is given by

$$\mathbb{E}[\text{Crt}_k(E)] = \left(\frac{\pi}{r}\right)^{\frac{p}{2}} \Gamma(m)^{-1} m^{(1+\gamma)m} \mathbb{E}_{\lambda_{C(x)}} \left[ e^{p \int \ln(\lambda - 2\gamma r \lambda) d\mu_{C(x)}} 1 \left\{ \lambda_{k+1}^{C(x)} \geq 2\gamma r \right\} \right] E^{(1-\gamma)m - 1} e^{-mE},$$

(63)

where

$$\gamma = \frac{p}{2m}.$$  

(64)

**Proof.** As discussed in Appendix A, the assumptions of the Kac–Rice formula (i.e. Lemma 7) are satisfied. Furthermore, due to the invariance of the Wishart distribution with respect to rotations on the hypertorus \cite{31,32}, we can integrate out the volume element independently; the volume of $(S^1)^\times p$ is

$$\int_{(S^1)^\times p} dw = (2\pi)^p.$$

(65)

Additionally, we have from Eq. (53) that the probability density of the gradient vector being zero at any $w$ conditioned on $H_{\text{WHRF}}(w) = x$ is

$$p_w(\nabla H_{\text{WHRF}}(w) = 0 \mid H_{\text{WHRF}}(w) = x) = \left(\frac{4\pi r x}{m}\right)^{-\frac{p}{2}}.$$

(66)

Taking the expectation over $x$ via Eq. (45), we have from Lemma 1 that

$$\mathbb{E}[\text{Crt}_k(B = (0,E))] = \left(\frac{\pi}{r}\right)^{\frac{p}{2}} \Gamma(m)^{-1} m^{(1+\gamma)m}$$

$$\times \int_0^E \mathbb{E}_{\lambda_{C(x)}} \left[ e^{p \int \ln(\lambda - 2\gamma r \lambda) d\mu_{C(x)}} 1 \left\{ \lambda_{k+1}^{C(x)} \geq 2\gamma r \right\} \right] x^{(1-\gamma)m-1} e^{-mx} dx.$$

(67)

Taking the derivative of this cumulative distribution with respect to $E$ yields the final result.
D. Logarithmic Asymptotics via Free Probability Theory

Though Eq. (63) is exact, it is difficult to use in practice. Luckily, we are able to use a surprising fact about the eigenvalue distributions of Wishart and GOE matrices; asymptotically, the empirical spectral distributions of these matrices weakly converge to fixed distributions. Concretely, in the limit \( p \to \infty \) where \( \gamma = \frac{2}{m} \) is held constant, the eigenvalue distribution of \( W/2m \) where \( W \sim W_p(2m, I_p) \) weakly converges to the Marchenko–Pastur distribution \([36]\):

\[
d\mu_{\text{M.P.}} = (1 - \gamma^{-1}) \mathbf{1} \{ \gamma > 1 \} \delta(\lambda) d\lambda + \frac{1}{2\pi\gamma^2} \left( (1 + \sqrt{\gamma})^2 - \lambda \right) \left( \lambda - (1 - \sqrt{\gamma})^2 \right) d\lambda. \tag{68}
\]

Similarly, the eigenvalue distribution of \( N/\sqrt{p} \) where \( N \sim \text{GOE}_p \) weakly converges to the Wigner semicircle distribution \([37]\):

\[
d\mu_{\text{s.c.}} = \frac{1}{2\pi} \sqrt{4 - \lambda^2} d\lambda. \tag{69}
\]

Furthermore, by using free probability theory one can find the asymptotic distribution of eigenvalues for a weighted sum of these matrices, given their eigenbases are in “generic position” with respect to each other. We now give a brief review of free probability theory—at least in the context of random matrix theory—here. Later, we will also briefly review large deviations theory, which we use to bound the probability of large deviations from the weak convergence of the eigenvalue distributions of Wishart and GOE matrices. Note that, as we are unable to control large deviations in Theorem 3, in principle large deviations in the weak convergence of VQAs to WHRFs could dominate the large deviations in WHRFs; however, as discussed in Sec. II C, this provably does not occur at shallow enough depths, and there are reasons to believe it does not occur even at large depths.

1. Free Probability Theory

Free probability theory is the study of noncommutative random variables. Specializing to random matrix theory on \( N \times N \) matrices, we define the unital linear functional

\[
\phi(X) \equiv \frac{1}{N} \mathbb{E} [\text{tr}(X)] \tag{70}
\]

as the free analogue of the expectation. Note that the eigenvalues of a matrix \( A \) are completely constrained by the trace of powers \( A^k \)—therefore, one can study the average distribution of the eigenvalues of a random matrix \( A \) via the moments \( \phi(A^k) \). Free independence (or freeness) is a generalization of the notion of independence in commutative probability theory to free probability theory. In the context of random matrix theory, two \( N \times N \) random matrices \( A \) and \( B \) are said to be freely independent if the mixed moments are identically zero; that is,

\[
\phi((A^{m_1} - \phi(A^{m_1}))(B^{n_1} - \phi(B^{n_1})) \ldots (A^{m_k} - \phi(A^{m_k}))(B^{n_k} - \phi(B^{n_k}))) = 0 \tag{71}
\]

for all \( n_i, m_i \in \mathbb{N} \). Roughly, the free independence of two random matrices means that their eigenbases are in “generic position” from one another.

Taking the analogy with commutative probability theory further, the analogue of the moment-generating function associated with the distribution of a random variable is the Stieltjes transform of the measure \( \mu \):

\[
G_\mu(z) = \int \frac{d\mu(t)}{z - t}, \tag{72}
\]

which can be inverted via the Stieltjes inversion formula:

\[
d\mu(t) = -\frac{1}{\pi} \lim_{\epsilon \to 0^+} \text{Im} \{ G_\mu(t + i\epsilon) \} dt. \tag{73}
\]

Similarly, the free analogue of the cumulant-generating function is the \( R \)-transform, which can be defined via the Stieltjes transform as the solution to the implicit equation:

\[
R_\mu(G_\mu(z)) + \frac{1}{G_\mu(z)} = z. \tag{74}
\]
The $R$-transform is important in that, if two random variables $A$ and $B$ are freely independent with probability measures $\mu_A$ and $\mu_B$ respectively, the probability measure $\mu_{A+B}$ of $A+B$ satisfies

$$R_{\mu_{A+B}} = R_{\mu_A} + R_{\mu_B}. \quad (75)$$

This can be interpreted as the free analog of the additivity of cumulants for commutative random variables. The probability measure $\mu_{A+B}$ is called the free convolution of $\mu_A$ and $\mu_B$, and is denoted using the notation

$$\mu_{A+B} = \mu_A \boxplus \mu_B. \quad (76)$$

Thus, given the probability distributions of two free random variables $A$ and $B$, there is a prescription for determining the probability distribution of their sum by taking their free convolution, just as the convolution in commutative probability theory describes the distribution of the sum of random variables.

### 2. Large Deviations Theory

In order to bound the probability of large deviations from the weak convergence of the eigenvalue distribution of $C$ to its asymptotic limit we will use results from large deviations theory, which we briefly review here. A sequence of measures $\{\mu_n\}$ is said to satisfy a large deviation principle in the limit $n \to \infty$ with speed $s(n)$ and lower semicontinuous rate function $I$ with codomain $[0, \infty]$ if and only if [38]

$$\inf_{x \in \Gamma^c} I(x) \leq \lim \inf_{n \to \infty} \frac{1}{s(n)} \ln (\mu_n (\Gamma)) \leq \lim \sup_{n \to \infty} \frac{1}{s(n)} \ln (\mu_n (\Gamma)) \leq - \inf_{x \in \Gamma} I(x) \quad (77)$$

for all Borel measurable sets $\Gamma$ that all $\mu_n$ are defined on. Here, $\Gamma^c$ denotes the closure of $\Gamma$ and $\Gamma^c$ the interior of $\Gamma$. The rate function $I$ is said to be good if all level sets of $I$ are compact. Large deviations theory will be useful for us to bound the probabilities of large deviations of the empirical spectral distribution of $\mu_{C(x)}$ as $p \to \infty$, and show that they do not contribute to leading order in the (logarithmic) asymptotic distribution of critical points. We do this using Varadhan’s lemma, which we state now.

**Lemma 3** (Varadhan’s lemma [38]). Suppose $\{\mu_n\}$ satisfies a large deviation principle with speed $s(n)$ and good rate function $I$ and let $\phi$ be a real-valued continuous function. Further assume either the tail condition

$$\lim_{M \to \infty} \lim_{n \to \infty} \sup \frac{1}{s(n)} \ln \left( E_{X_n \sim \mu_n} \left[ e^{s(n)\phi(X_n)} 1 \{ \phi(X_n) \geq M \} \right] \right) = -\infty, \quad (78)$$

or the moment condition for some $\gamma > 1$

$$\lim_{n \to \infty} \sup \frac{1}{s(n)} \ln \left( E_{X_n \sim \mu_n} \left[ e^{s(n)\phi(X_n)} \right] \right) < \infty. \quad (79)$$

Then,

$$\lim_{n \to \infty} \frac{1}{s(n)} \ln \left( E_{X_n \sim \mu_n} \left[ e^{s(n)\phi(X_n)} \right] \right) = \sup_x (\phi(x) - I(x)). \quad (80)$$

### 3. Logarithmic Asymptotics of the Distribution of Critical Points

Equipped with these mathematical tools, we prove our first result on the asymptotic behavior of $\mu_{C(x)}$ which is present in the expectation of Eq. [63].

**Lemma 4** (Asymptotic behavior of $\mu_{C(x)}$). Define $G^*_x (z)$ as the implicit solution of the equation

$$8r^3 \gamma^2 x G^*_x (z)^3 - 2r \gamma (z + 2rx) G^*_x (z)^2 + (z - 2r (1 - \gamma)) G^*_x (z) - 1 = 0 \quad (81)$$

with the smallest imaginary part. Define

$$d\mu^*_x = -\frac{1}{\pi} \Im \{G^*_x\} \, d\lambda. \quad (82)$$

Let $p, m \to \infty$ as $\gamma = \frac{p}{2m}$ is held constant. Then, the empirical spectral measure $\mu_{C(x)}$ satisfies a large deviation principle as $p \to \infty$ with speed $p^2$ with good rate function uniquely minimized at $\mu^*_x$ with a value of 0.
Proof. The empirical spectral measure of the random matrix \( N / \sqrt{p} \) satisfies a large deviation principle at a scale \( p^2 \), with good rate function minimized by Wigner’s semicircle law \([39]\). Similarly, the empirical spectral measure of the random matrix \( W/2m \) satisfies a large deviation principle at a scale \( p^2 \), with good rate function minimized by the Marchenko–Pastur distribution \([40]\). As the \( R \)-transform of the empirical spectral distribution of \( A \) satisfies the scaling property

\[
R_{aA}(z) = aR_A(az),
\]

the \( R \)-transform of the empirical spectral distribution of the weighted GOE term of \( C \) is of the form

\[
R_{\text{GOE}}(z) = 4r^2 \gamma xz
\]

and the \( R \)-transform of the weighted Wishart term is of the form

\[
R_{\text{Wishart}}(z) = \frac{2r}{1 - 2r \gamma z}.
\]

By the asymptotic freeness of independent GOE and Wishart matrices \([41]\), \( \mu_C(x) \) converges weakly to the fixed measure \( \mu^* \) with \( R \)-transform

\[
R_x(z) = R_{\text{Wishart}}(z) + R_{\text{GOE}}(z).
\]

Eqs. \(81\) and \(82\) come from inverting the \( R \)-transform \( R_x \) via Eqs. \(74\) and \(73\), respectively.

We now consider large deviations in the weak convergence \( \mu_C(x) \to \mu^* \). Conditioning on the empirical spectral distribution of \( W/2m \) and using the “strongest growth wins” principle \([38]\), we have that \( \mu_C(x) \) satisfies a large deviation principle with speed \( p^2 \) with rate function given by

\[
I(\mu) = \inf_{\mu_{W/2m}} (J_{\mu_{W/2m}}(\mu) + K(\mu_{W/2m}));
\]

here, \( K \) is the rate function governing convergence of the empirical spectral distribution of the Wishart ensemble \([40]\) and \( J_{\mu_{W/2m}} \) is the rate function governing convergence of the empirical spectral distribution of a fixed matrix with asymptotic eigenvalue distribution \( \mu_{W/2m} \) summed with a GOE matrix \([42]\). This sum is obviously uniquely minimized by \( \mu = \mu^* \), when \( I(\mu^*) = 0 \).

Now, we examine the asymptotic behavior of the smallest eigenvalue \( \lambda^{C(x)}_1 \) of \( C(x) \). Unlike the empirical spectral measure \( \mu_C(x) \) which satisfies a large deviation principle at a speed \( p^2 \), we will see that this eigenvalue satisfies a large deviation principle at a speed \( p \), with deviations at this speed to the left of the asymptotic value \( \lambda^*_{z,1} \).

**Lemma 5** (Asymptotic behavior of \( \lambda^{C(x)}_1 \)). Let \( \lambda^*_{z,1} \) be the infimum of the support of \( \mu^*_z \) as defined in Eq. \(82\). Then, the smallest eigenvalue \( \lambda^{C(x)}_1 \) of \( C(x) \) satisfies a large deviation principle with speed \( p \) with good rate function that is infinite at \( y > \lambda^*_{z,1} \) and is uniquely minimized at \( y = \lambda^*_{z,1} \) with a value of 0.

**Proof.** The limiting smallest eigenvalues \( \lambda^W_1, \lambda^N_1 \) of \( W/2m \) and \( N/\sqrt{p} \) both satisfy large deviation principles with speed \( p \) that are infinite for \( \lambda_1 \) in the bulk of their respective limiting empirical spectral distributions \([43, 44]\). As in the proof of Lemma \(4\) we condition on large deviations of these eigenvalues \([38]\) and therefore have that the rate function governing \( \lambda^{C(x)}_1 \) is

\[
I(y) = \inf_{\lambda^W_1, \lambda^N_1} \left( J_{\lambda^W_1, \lambda^N_1}(y) + K(\lambda^W_1) + L(\lambda^N_1) \right);
\]

here, \( K \) is the rate function governing the convergence of \( \lambda^W_1, L \) that of \( \lambda^N_1 \), and \( J \) that of the smallest eigenvalue \( C(x) \) conditioned on the eigenvalue distributions of \( W \) and \( N \). Using known results on the large deviations of the smallest eigenvalue of the sum of two matrices with fixed eigenvalues (i.e. \( J_{\lambda^W_1, \lambda^N_1} \)) \([45]\), we see that \( I(y) \) is infinite for \( y > \lambda^*_{z,1} \) and is uniquely minimized at \( y = \lambda^*_{z,1} \) with a value of 0. \(\square\)

Using Lemmas \(4\) and \(5\), we can prove the following logarithmic asymptotics on the expectation term in Eq. \(63\). We will find that neither the large deviations in the convergence \( \mu_C(x) \) or \( \lambda^{C(x)}_1 \) will contribute to leading order in the logarithmic asymptotics of \( \text{Crt}_k(E) \), as at a speed \( p \) the only large deviations are \( \lambda^{C(x)}_1 \leq \lambda^*_{z,1} \) which are dominated by \( \lambda^{C(x)}_1 = \lambda^*_{z,1} \) in the expectation.
Lemma 6 (Logarithmic asymptotics of the determinant). Let $d\mu^*_E$ be the spectral measure given in Eq. (82), with $\lambda^*_{E,1}$ the infimum of its support. Let $p, m \gg 1$ with $\frac{p}{2m} = \gamma = O(1)$. Then,

$$\frac{1}{p} \ln \left( \mathbb{E}_{C(E)} \left[ e^{p \int [\ln(\lambda - 2rE)]d\mu^*_E} 1 \left\{ \lambda^*_k \geq 2E \right\} \right] \right) = \int \ln \left( 1 \left\{ \lambda^*_{E,1} \geq 2E \right\} |\lambda - 2E| \right) d\mu^*_E + o(1).$$

Proof. As $\mu^*_C$ satisfies a large deviation principle with speed $p^2$ with rate function minimized at $\mu^*_E$ by Lemma 4 and as $1 \left\{ \lambda^*_C \geq 2E \right\} \leq 1$, we have that the tail condition of Varadhan’s lemma at speed $p$ is satisfied [38] and therefore

$$\lim_{p \to \infty} \frac{1}{p} \ln \left( \mathbb{E}_{C(E)} \left[ e^{p \int [\ln(\lambda - 2rE)]d\mu^*_E} 1 \left\{ \lambda^*_C \geq 2E \right\} \right] \right) = \sup_{\lambda \in R} \left( \int \ln \left( 1 \left\{ \lambda \geq 2E \right\} |\lambda - 2E| \right) d\mu^*_E - I(\lambda) \right).$$

Here, $I$ is as in Eq. (88). The supremum over $\lambda$ is obviously achieved when $\lambda = \lambda^*_{E,1}$ by the properties of $I$ discussed in Lemma 5, giving the leading order term in Eq. (89). The result being exact in the $p \to \infty$ limit gives the subleading $o(1)$.

Using Lemma 6, we can therefore finally calculate the logarithmic asymptotic distribution of local minima of a WHRF.

Theorem 6 (Logarithmic asymptotics of the local minima distribution). Let $d\mu^*_E$ be the spectral measure given in Eq. (82), with $\lambda^*_{E,1}$ the infimum of its support. Let $p, m \gg 1$ with $\frac{p}{2m} = \gamma = O(1)$. Then, the expected distribution of local minima of a WHRF at a fixed energy $E > 0$ is given by

$$\frac{1}{p} \ln \left( \mathbb{E}[\text{Crt}_0(E)] \right) = \frac{1}{2} \ln \left( \frac{\pi q}{2\gamma} \right) + \frac{1}{2\gamma} (1 - E) + \frac{1}{2} (\gamma^{-1} - 1) \ln (E) + \int \ln \left( \left| \frac{\lambda}{r} - 2E \right| 1 \left\{ \frac{\lambda^*_{E,1}}{r} \geq 2E \right\} \right) d\mu^*_E + o(1).$$

Proof. The result follows directly from applying Lemma 6 to Theorem 5.

Note that, though we only prove the asymptotic distribution of local minima in Theorem 6, we expect similar theorems to also hold for critical points of constant index $k$ (taking $\lambda^*_{E,1} \to \lambda^*_{E,k}$ in the integrand); the only difference in the derivation is the exact form of the large deviations of the $k$th smallest eigenvalue of $C(x)$. This is similar to the case in Gaussian hyperspherical random fields [16].

E. Discussion of the Critical Point Distribution

Let us now discuss the implications of Theorem 6. Note first that the rescaled logarithmic number of critical points diverges when $q \to \infty$: following the derivation closely, one sees that this is due to the exponentially suppressed (when $m$ is exponential in $n$) gradient in Eq. (56). We believe that this is a manifestation of the “barren plateau” phenomenon [3][5], where for many models of ansatizes it can be shown that deep circuits have an exponentially vanishing variance of the gradient. This interpretation suggests that these barren plateau regions are filled with many small “bumps” that are exponentially shallow. Furthermore, note that this class of random fields exhibits banded behavior in the eigenvalues. That is, there exist local minima in the band $0 \leq E \leq E_0$, where $E_0$ is the solution to

$$\lambda^*_{E_0,1} = 2rE_0.$$ 

This banded behavior is similar to that in the Gaussian spherical case, used as a model for classical neural networks [13][15]. We will see, however, that unlike the classical case, this does not give necessarily good guarantees on the distribution of local minima. This is due to $E_0$ being generally far from 0 when $\gamma < 1$ as $p, m \to \infty$. To illustrate this, we focus now on two cases: $p \geq 2m$ (the overparameterized regime) and $p \ll m$ (the underparameterized regime).
1. $p \geq 2m$

First, let us consider the case where $p \geq 2m$, i.e. $\gamma \geq 1$. In this limit, the Wishart term of $C$ is low-rank, and $\mu_E^*$ has support on eigenvalues $\leq 0$ for all $E \geq 0$. Therefore, the condition $1 \{ \lambda_{E,1}^* \geq 2E_r \}$ is never satisfied, and to leading order in $p$ there are no critical points at an energy $E > 0$. That is, all local minima are global minima in the $p \to \infty$ limit when $\gamma \geq 1$. Though the choice of ansatz is different, we suspect that a related phenomenon may be what gives rise to the phase transition in training numerically observed in [9][12]. In the case of the HVA for certain Hamiltonians, we believe that the ansatz’s structured form implies that, even for $p = O \left( \text{poly} \left( n \right) \right)$, $p \geq 2m$ such that training almost always find a global minimum. See our discussion in Sec. II C for more details.

2. $p \ll m$

When the number of distinct parameters $p$ is poly ($n$) and considering a physically relevant Hamiltonian such that the number of degrees of freedom $m$ is exp ($n$), we have that $p \ll m$ (i.e. $\gamma \ll 1$) for large $n$. In this limit, the spectral distribution $\mu_E^*$ is dominated by the Wishart term of $C$, as its eigenvalues are $O \left( 1 \right)$ while the eigenvalues of the GOE term are $O \left( \sqrt{\gamma} \right)$. Furthermore, the Marchenko–Pastur distribution in this limit only has support at $\lambda = 1 + O \left( \sqrt{\gamma} \right)$. Therefore, the expected number of local minima at an energy $E$ will be proportional to

$$E \left[ \text{Crt}_0 \left( E \right) \right] \propto e^{-mE_0 + O \left( 1 \right)} E^{-\frac{p}{2}} \left( 1 - 2E + O \left( \frac{1}{2} \right) \right)^p \left\{ 0 \leq E + O \left( \sqrt{\gamma} \right) \leq \frac{1}{2} \right\}. \quad (93)$$

In particular, up to shifts on the order of $\sqrt{\gamma}$, the distribution of local minima is roughly that of a compound confluent hypergeometric (CCH) distribution [19]. The CCH distribution can be considered a generalization of the beta distribution, and for our parameters has mean on the order of $\frac{1}{2} - \gamma$ and standard deviation on the order of $m^{-1}$. Restoring the units of Eq. (18), this implies that in this limit the local minima of the variational loss function exponentially concentrate (in expectation) near the mean eigenvalue of $H - \lambda_1$. Even worse, the CCH distributed form of the distribution of local minima implies that, even when beginning at an initial energy well below half of the mean eigenvalue of $H - \lambda_1$, the found energy will only improve by a fraction of the initial energy before the optimization algorithm finds a local minimum. Empirically, we find that this occurs not just in expectation but also for individual instances in Sec. IV A.

### IV. NUMERICAL RESULTS

We now test our analytic predictions using numerical simulations. First, we investigate the numerical performance of the class of randomized ansatz discussed in Sec. II and give numerical evidence of things we were unable to prove. Then, we give numerical evidence of our heuristic reasoning that, for ansatzes dependent on the objective Hamiltonian, the effective $m$ can be much smaller (see Sec. II C). In all cases, we numerically test the predictions of our results by using the variational quantum eigensolver (VQE) [2] to attempt to find the ground state of the 1D $n$ site spinless Fermi–Hubbard Hamiltonian [46]

$$H_{T,U} = -T \sum_{i=1}^{n-1} \left( c^\dagger_i c_{i+1} + c^\dagger_{i+1} c_i \right) + U \sum_{i=1}^{n-1} c^\dagger_i c_i c^\dagger_{i+1} c_{i+1}, \quad (94)$$

where $c$ is the fermionic annihilation operator. Here, we take units such that the mean eigenvalue of the considered Hamiltonian (minus its ground state energy) is $E = 1$. We give further details of our numerical simulations in Appendix B.

#### A. Empirical Performance of Random Ansatzes

First, we analyzed the performance of VQE on this loss function via the random ansatz construction procedure defined in Theorem 3. Previous numerical results on a related problem-agnostic ansatz have already shown concentration of local minima far away from the global minimum below some degrees-of-freedom transition, and concentration at the global minimum above this transition [9]. Here, we tracked where our analysis predicts the local minima to lie as a function of $\gamma$ for $\gamma \ll 1$, up to deviations on the order of $\sqrt{\gamma}$ (as discussed in Sec. II E).
FIG. 1. Here we plot the distribution of found local minima found after 52 separate VQE training instances on the spinless Fermi–Hubbard Hamiltonian given in Eq. (94) using the randomized ansatz on (a) $n = 6$ and (b) $n = 8$ qubits. Dashed lines denote the predicted region local minima will lie. Note the clustering of local minima at a finite energy when $\gamma \ll 1$.

FIG. 2. Here we plot the distribution of found local minima found after 52 separate VQE training instances on the spinless Fermi–Hubbard Hamiltonian given in Eq. (94) using the randomized ansatz, with $p = 48$ on $n = 6$ qubits. Note that even for small system sizes, qualitatively the expected distribution of critical points and the distribution of critical points for a fixed random ansatz are in agreement.

Concretely, for a given training instance and depth $q = p$, we generated an ansatz $|\theta\rangle$ composed of $p$ layers of Pauli rotations, where each Pauli rotation was chosen uniformly from all nonidentity Pauli matrices on $n$ qubits. A summary of the normalized distribution of found local minima for the randomized ansatz with $n = 6, 8$ is given in Fig. 1, along with the predicted region in which all local minima should lie in the $p \to \infty$ limit as discussed in Sec. III D 3. See Appendix B for details on how this distribution was generated.

We see that almost all found local minima lie within the predicted region, even at small $p,n$. In particular, for small $\gamma$, the distribution of local minima is almost entirely localized within $\sqrt{\gamma}$ of the predicted $\frac{1}{2} - \gamma$ (in units of the mean eigenvalue of $H - \lambda_1$). Finally, we numerically observe that the distribution of local minima are qualitatively similar in expectation and for a single random ansatz in Fig. 2.
B. Empirical Performance of the Hamiltonian Variational Ansatz

Previous numerical results \cite{10} on models related to Eq. (94) have shown that only a number of parameters polynomial in the volume of the model suffices for efficient training when using the Hamiltonian variational ansatz (HVA) \cite{20}. As discussed in Sec. II C, we believe this is due to this class of ansatzes effectively limiting the degrees of freedom \( m \) of the associated WHRF model; to test this, we once again tracked where our analysis predicts the local minima to lie as a function of \( \gamma \) for \( \gamma \ll 1 \), up to deviations on the order of \( \sqrt{\gamma} \).

Unlike the random class of ansatzes we discuss in our mapping in Sec. II, this choice of ansatz depends on the Hamiltonian objective. For the Fermi–Hubbard Hamiltonian of Eq. (94), each HVA layer is of the form

\[
U_{T,U}(\theta_i) = e^{-i\theta_i 2 H_{T,U,\text{Coulomb}}} e^{-i\theta_i 2 H_{T,U,\text{even}}} e^{-i\theta_i 1 H_{T,U,\text{odd}}},
\]

Here, \( H_{T,U,\text{Coulomb}} \) is composed of the terms proportional to \( U \) in \( H_{T,U,\text{even}} \) the hopping terms on even links, and \( H_{T,U,\text{odd}} \) the hopping terms on odd links. We took the starting state \( |\psi_0\rangle \) to be the computational basis state \( |1\rangle \) on the first \( \frac{n}{2} \) qubits and \( |0\rangle \) on the other \( \frac{n}{2} \) qubits. To observe the effects of scaling the number of independent parameters \( p \), we overparameterize our ansatz at a fixed overall depth by fixing the total number of ansatz layers \( U_{T,U} \) to be 6, but introduce extra parameters to govern each evolution. For instance, for a multiplicative factor \( f = 2 \), we double the number of parameters by splitting into a sum of two terms

\[
H_{T,U,\text{Coulomb}} = H_{T,U,\text{Coulomb}}^{(1)} + H_{T,U,\text{Coulomb}}^{(2)},
\]

\[
H_{T,U,\text{even}} = H_{T,U,\text{even}}^{(1)} + H_{T,U,\text{even}}^{(2)},
\]

\[
H_{T,U,\text{odd}} = H_{T,U,\text{odd}}^{(1)} + H_{T,U,\text{odd}}^{(2)},
\]

and parameterize the evolution under each term separately. For \( f = 1 \), this ansatz preserves the fermion number of the initial state; thus, for these simulations we calculate \( m \) in this \( \frac{n}{2} \)-fermion subspace. For large \( f \), this parameterization breaks the fermion number conservation of the ansatz, but still preserves the parity of the fermion number. In practice, then, the \( \gamma \) we compute should be considered an upper bound on the true \( \gamma \).

We show the empirical distribution of local minima in Fig. 3 for \( n = 8 \), along with the predicted region local minima should lie as discussed in Sec. III E 2 and Appendix B. Even when using an upper bound on \( \gamma \), the predicted local minima distribution is overly pessimistic (particularly at larger \( p \)). We suspect this is due to the fact that the ansatz is constructed in a way that minimizes the effective degrees of freedom of the WHRF \( m \) such that \( \gamma \) is close to 1 for smaller \( p \) than is predicted. See Sec. II C for a heuristic explanation of this phenomenon.

V. CONCLUSION

Though variational quantum algorithms are perhaps the most promising way to use noisy intermediate-scale quantum devices for practical computational tasks, there are many caveats with regard to their performance. In particular,
previous work has shown that utilizing variational ansatzes that are independent of the problem Hamiltonian can introduce a barren plateau phenomenon where, though the ansatz is expressive enough to capture the ground state of interest, in practice optimizing the loss function is infeasible [6–8]. We extended these results by showing a particular class of random ansatzes independent of the problem instance not only exhibits barren plateaus, but also has a concentration of local minima near the mean eigenvalue of the objective Hamiltonian. This is in contrast to the case in classical neural networks, where even generic model structure leads to a concentration of local minima in a band near the global minimum of the loss function.

Though our results may not seem encouraging for variational quantum algorithms, we emphasize that we expect our analytic results to hold only when the ansatz is independent of the problem Hamiltonian. In particular, we believe that choosing from a restricted class of ansatzes that depend on the problem Hamiltonian may restrict the effective degrees of freedom in the random field the variational quantum algorithm maps to, effectively shifting the concentration of local minima to be closer to the global minimum of the loss function. Indeed, we see numerically that this is the case for the Hamiltonian variational ansatz, where our analytic results seem much too pessimistic. In principle, this new way of thinking about variational quantum algorithms may inform future ansatz design; we leave for future work the study of how various ansatz choices may impact the distribution of critical points of the loss function positively.

ACKNOWLEDGMENTS

We are grateful to Tongyang Li, John Napp, and Aram Harrow for insightful discussion and helpful points regarding a draft of this work. E.R.A. is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 400063445, and a Lester Wolfe Fellowship and the Henry W. Kendall Fellowship Fund from M.I.T.
Appendix A: The Kac–Rice Formula and its Assumptions

For completeness, we state the formal version of Lemma 7—with all assumptions—here. We borrow heavily from [15]. By $\nabla f$, we mean the covariant gradient of $f$.

**Lemma 7** (Kac–Rice formula). Let $M$ be a compact, oriented, $N$-dimensional $C^1$ manifold with $C^1$ Riemannian metric $g$. Let $B \subset \mathbb{R}^K$ be an open set such that $\partial B$ has dimension $K - 1$. Let $f : M \to \mathbb{R}^K$ be a random field on $M$, and let $\iota(\cdot)$ denote the index of $\cdot$. Furthermore, assume that:

1. All components of $f$, $\nabla f$, and $\nabla^2 f$ are almost surely continuous and have finite variances over $M$.
2. The marginal density $p_t(\nabla f(t))$ of $\nabla f$ at $t \in M$ is continuous at $\nabla f = 0$.
3. The conditional densities $p_t(\nabla f(t) | f(t), \nabla^2 f(t))$ are bounded above and continuous at $\nabla f = 0$, uniformly in $t \in M$.
4. The conditional densities $p_t(\det(\nabla^2 f(t)) | \nabla f(t) = 0)$ are continuous in the neighborhood of $\det(\nabla^2 f) = 0$ and $\nabla f(t) = 0$, uniformly in $t \in M$.
5. The conditional densities $p_t(f(t) | \nabla f(t) = 0)$ are continuous for all $f$ and for all $\nabla f$ in a neighborhood of 0, uniformly in $t \in M$.
6. The Hessian moments are bounded, i.e.

$$\sup_{t \in M} \max_{i,j} \mathbb{E} \left[ \left| (\nabla^2 f(t))_{i,j} \right|^N \right] < \infty. \quad (A1)$$

7. The moduli of continuity with respect to (the canonical metric induced by) $g$ of each component of $f$, $\nabla f$, and $\nabla^2 f$ all satisfy

$$\mathbb{P} [\omega(\eta) > \epsilon] = o(\eta^n) \quad (A2)$$

for all $\epsilon > 0$ as $\eta \to 0^+$. 


Then,

\[
E \left[ C r t_k^f (B) \right] = \int_M p_\sigma (\nabla f (\sigma) = 0) \\
\times E \left[ \left| \left| \det (\nabla^2 f (\sigma)) \right| \right| 1 \{ f (\sigma) \in B \} 1 \{ \iota (\nabla^2 f (\sigma)) \leq k \} \mid \nabla f (\sigma) = 0 \right] d\sigma ,
\]

where \( d\sigma \) is the volume element induced by \( g \) on \( M \).

It is obvious by Lemma 2 that conditions 2-6 are satisfied given \( B = (0, u) \). Furthermore, as \( F \) is a polynomial in \( \{ \cos (\theta_i), \sin (\theta_i) \} \), \( F \) and its derivatives are continuous for any value of the components of \( m^{-1} J \), and all have finite variance. Similarly, it is easy to see that the modulus of continuity of \( f \) and its gradients go as \( J \eta^r \) as \( \eta \to 0^+ \), where \( J \) is the largest component of \( m^{-1} J \). As the distributions of the components of a Wishart matrix have exponential tails, the probability that \( J = \Omega (\eta^{-r}) \) is indeed \( o (\eta N) \) and therefore all conditions are satisfied.

**Appendix B: Details of Numerical Simulations**

Here we give further details on the numerical simulations performed in Sec. IV. We performed all simulations via Qiskit [47], and used standard gradient descent (via the method of finite differences) to optimize the loss function

\[
F (\theta) = \langle \theta | H_{1,2} | \theta \rangle
\]

until convergence. Our implementation of gradient descent used a learning rate of 0.05 and a momentum of 0.9, and halted when either the function value improved by no more than \( 10^{-5} \) or after \( 10^6 \) iterations, whichever came first.

To estimate the empirical distribution of local minima for the studied instances of VQE, we repeated this procedure 52 times, using a new ansatz and uniformly random starting point for each training instance. We also verified numerically that \( m \) as defined in Eq. (14) is on the order of \( 2^n \) for \( H_{1,2} \), though we directly used Eq. (14) when computing \( \gamma \). In all plotted instances, we normalize the energy scale by a factor of \( c_{VQA} \), where

\[
c_{VQA} = \lambda - \lambda_1 ;
\]

this is just the overall factor of Eq. (18). These units are such that the mean eigenvalue of \( H - \lambda_1 \) in the subspace of interest is at \( E = 1 \).

Finally, though we write \( H_{1,2} \) for brevity when discussing our objective Hamiltonian, in practice we took i.i.d normal \( T_i \) and \( U_i \) at each link to break translation invariance. In our simulations, these random variables were centered at \( T = 1 \) and \( U = 2 \), respectively, and each had a variance of \( 10^{-2} \).