Low-Depth Gradient Measurements Can Improve Convergence in Variational Hybrid Quantum-Classical Algorithms

Aram W. Harrow and John C. Napp

Phys. Rev. Lett. **126**, 140502 — Published 7 April 2021

DOI: 10.1103/PhysRevLett.126.140502

Low-depth gradient measurements can improve convergence in variational hybrid quantum-classical algorithms

Aram W. Harrow* and John Napp†
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
(Dated: December 23, 2020)

Within a natural black-box setting, we exhibit a simple optimization problem for which a quantum variational algorithm that measures analytic gradients of the objective function with a low-depth circuit and performs stochastic gradient descent provably converges to an optimum faster than any algorithm that only measures the objective function itself, settling the question of whether measuring analytic gradients in such algorithms can ever be beneficial. We also derive upper bounds on the cost of gradient-based variational optimization near a local minimum.

Introduction.—In recent years, an array of *variational hybrid quantum-classical algorithms* have been widely studied as leading candidates for near-term quantum computers, due to their relatively modest quantum resource requirements and potential of scalability. Variational algorithms have been proposed in the context of quantum simulation (e.g. variational quantum eigensolvers [1, 2]), combinatorial optimization (e.g. QAOA [3]), and machine learning (e.g. quantum classifiers [4–8]).

In the variational setting, one can prepare states belonging to some parameterized family $\{|\boldsymbol{\theta}\rangle\}_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \in \mathcal{X} \subset \mathbb{R}^p$, where p is the number of variational parameters. The set of parameterized states which may be prepared will depend on the specifications of the quantum device. We consider parameterizations consisting of p "pulses" applied to some easy-to-prepare starting state $|\Psi\rangle$,

$$|\boldsymbol{\theta}\rangle := |\theta_1, \ldots, \theta_p\rangle = e^{-iA_p\theta_p/2} \cdots e^{-iA_1\theta_1/2} |\Psi\rangle ,$$

where A_j is the Hermitian operator which generates pulse j. This form of parameterization is well-motivated theoretically [9–11] and is widely considered in the literature.

A classical "outer loop" controls the quantum device, which is used only for preparing variational states and making simple measurements. The classical outer loop uses this measurement information to perform a *classical* optimization of some objective function $f(\boldsymbol{\theta})$ over the feasible set $\mathcal{X}$, where $f(\boldsymbol{\theta})$ is induced by some Hermitian *objective observable* H, via the definition $f(\boldsymbol{\theta}) := \langle\boldsymbol{\theta}|H|\boldsymbol{\theta}\rangle$.

Given the ability to prepare variational states $|\boldsymbol{\theta}\rangle$, there remain the questions of what observables should be measured, and how the measurement outcomes should be used by the classical outer loop to find an approximate minimizer for $f(\boldsymbol{\theta})$. Typically, the objective observable H is decomposed as a linear combination of observables which each can be efficiently measured in low depth. For instance, we may always write a Pauli decomposition $H = \sum_i \alpha_i P_i$, where $\alpha_i > 0$ and P_i are tensor products of Pauli operators. By linearity, it is possible to construct an estimator for $\langle\boldsymbol{\theta}|H|\boldsymbol{\theta}\rangle$ via measurements of the Pauli strings $\{P_i\}_i$.

In this work, we will find it convenient to take a novel but natural approach for estimating the objective function or its derivatives via sampling terms of the Pauli decomposition to measure according to an appropriate distribution; a similar sampling strategy was previously employed in the context of random compiling for Hamiltonian simulation [12]. To this end, we express H as an expectation value, $H = E\,\mathbb{E}_X\, P_X$, where $E := \sum_i \alpha_i$ and the random variable X is distributed as $p_X(x) := \alpha_x/E$. For a given point $\boldsymbol{\theta}$ in parameter space, by linearity we have $f(\boldsymbol{\theta}) = \langle\boldsymbol{\theta}|H|\boldsymbol{\theta}\rangle = E\,\mathbb{E}_X\, \langle\boldsymbol{\theta}|P_X|\boldsymbol{\theta}\rangle$. Hence, an unbiased $\pm E$-valued estimator for $f(\boldsymbol{\theta})$ may be obtained by sampling x from the distribution p_X, measuring P_x w.r.t. $|\boldsymbol{\theta}\rangle$, and then scaling the output by E. With estimates of f obtained in this way, the classical outer loop performs a stochastic zeroth-order (i.e. derivative-free) optimization of the function $f(\boldsymbol{\theta})$; 'stochastic' because of the randomness of the measurement outcomes when estimating $f(\boldsymbol{\theta})$, and 'derivative-free' because the outer loop receives estimates of $f(\boldsymbol{\theta})$ rather than estimates of its gradient $\nabla f(\boldsymbol{\theta})$ or of higher-order derivatives.

However, it is not apparent that such a zeroth-order strategy is best. Indeed, as observed in a number of works (listed in the subsequent section), by performing a slightly more complicated measurement it is possible to directly estimate $\nabla f(\boldsymbol{\theta})$; this estimate can then be used with a first-order (i.e. gradient-based) optimization algorithm. To this end, we may express the j^{th} component of the gradient as an expectation value, $\nabla_j f(\boldsymbol{\theta}) = \langle\boldsymbol{\theta}|G_j|\boldsymbol{\theta}\rangle$, where

$$G_j = \frac{i}{2}[U_{(j+1):p} A_j U_{(j+1):p}^\dagger, H]$$

(see Section II of the Supplemental Material [13] for a derivation). Here, $U_{j:k}$ is shorthand for $e^{-iA_k\theta_k/2} \cdots e^{-iA_j\theta_j/2}$. To measure G_j with a low-depth circuit, we may expand A_j and H as linear combinations of products of Pauli operators with positive coefficients, obtaining

$$\nabla_j f(\boldsymbol{\theta}) = \Gamma_j \mathop{\mathbb{E}}_{K,L} \langle\boldsymbol{\theta}|\frac{i}{2}[U_{(j+1):p} Q_K^{(j)} U_{(j+1):p}^\dagger, P_L]|\boldsymbol{\theta}\rangle ,$$

where $Q_k^{(j)}$ are Pauli operators appearing in the expansion of A_j, Γ_j is the sum of coefficients appearing

in the resulting expansion, and the joint probability of $(K = k, L = l)$, denoted $q_{KL}(k, l)$, is proportional to the coefficient associated with the term in the expansion including $Q_k^{(j)}$ and P_l. A ± 1-valued unbiased estimator for $\langle\boldsymbol{\theta}|\frac{i}{2}[U_{(j+1):p}Q_k^{(j)}U_{(j+1):p}^\dagger, P_l]|\boldsymbol{\theta}\rangle$ can be obtained with a single measurement via a simple Hadamard-test circuit (as described in [14–16]; see also Section II of the Supplemental Material [13]). Now, we may construct a $\pm\Gamma_j$-valued unbiased estimator for $\nabla_j f(\boldsymbol{\theta})$ with a single measurement by sampling (k, l) from q_{KL}, measuring the corresponding observable as described above, and scaling the output by Γ_j. Generalizations of this strategy permit the measurement of higher-order derivatives as well.

Finally, an unbiased estimator $\hat{\mathbf{g}}(\boldsymbol{\theta})$ for the full gradient may be constructed with one measurement by choosing component j with probability $\Gamma_j/\|\vec{\Gamma}\|_1$, estimating $\nabla_j f(\boldsymbol{\theta})$ using the method described above, and then scaling the output by $(\|\vec{\Gamma}\|_1/\Gamma_j)\hat{e}_j$ where $\hat{e}_j$ denotes the unit vector in direction j. Here we have defined the vector $\vec{\Gamma} := (\Gamma_1, \ldots, \Gamma_p)^\top$. It may be verified [13] that $\hat{\mathbf{g}}$ is $\pm\|\vec{\Gamma}\|_1$-valued, and that $\mathbb{E}\,\hat{\mathbf{g}} = \nabla f$. Note that the choice to sample j with probability proportional to Γ_j is optimal for minimizing $\mathbb{E}\,\|\hat{\mathbf{g}}\|^2$ among all choices of sampling weights (as may be verified via a Lagrange multiplier), and furthermore results in this quantity having no explicit dependence on p.

Our method for constructing unbiased estimators for f and its gradient is effectively a form of importance sampling which assigns higher weight to larger terms in the sum; this is reflected in the fact that the magnitude of an estimator depends on an appropriate *sum* of coefficients, but carries no explicit dependence on the number of terms in the decomposition (or on the number of variational parameters for the gradient estimator). This is especially relevant for applications (such as quantum chemistry) for which many terms of the sum may have small weight. After a preprint of this paper was made public, subsequent works [17, 18] have numerically studied similar estimators and have furthermore proposed methods of adaptively setting the sampling weights associated with each observable in the expansion [18, 19].

A fundamental question is now whether, within the vicinity of a local optimum, "first-order" variational algorithms which perform measurements to construct gradient estimators can converge faster than algorithms which use the simpler, "zeroth-order" strategy of estimating only the objective function itself. This question may be especially important in the context of quantum simulation, in which a precise solution is often desired. Within a natural black-box setting, we answer this question affirmatively by exhibiting an optimization problem for which performing gradient measurements, and using these gradient estimates in conjunction with stochastic gradient descent (SGD) [20], converges to an optimum asymptotically faster than any strategy based on measuring the objective function.

The optimization problem we analyze to demonstrate this separation is quite simple: it is essentially the problem of learning the ground state of a 1-local (non-interacting) spin Hamiltonian. While an analytic solution to this problem may be readily derived, the black-box model ensures the variational algorithm behaves in a generic way, rather than merely solving the problem analytically (as this would be computationally infeasible for more complicated problems). This simple problem provides a counterexample to the proposition that, within the natural black box setting defined below, the convergence rate of gradient-based variational algorithms can be generically matched by that of zeroth-order algorithms. In particular, this rules out the possibility that gradient measurements can always be replaced by gradient estimates obtained by finite-differencing energy measurements without a loss of performance. This observation may be of interest in the design of practical NISQ algorithms, in which gradient measurements could be more difficult to implement than energy measurements. Our results demonstrate that one cannot hope to generically simulate gradient measurements while maintaining equivalent performance; hence, incurring extra overhead for measuring gradients could be worthwhile.

The speedup we obtain for gradient-based algorithms crucially relies on using an appropriate choice of variational ansatz for the problem at hand, making our toy model setting more similar to that of variational algorithms with theoretically motivated ansätze rather than those which use a "hardware-efficient ansatz" [21]. Indeed, the setting of "barren plateaus" [22] in which the ansatz looks random and gradient-based optimization fails may be viewed as the opposite situation to that studied in this work.

While our analysis of a non-interacting system is sufficient to rule out the existence of zeroth-order algorithms which *generically* match the performance of first-order algorithms, we cannot rule out the possibility that certain classes of problems contain additional structure which allows zeroth-order algorithms to match the convergence rate of first-order algorithms. However, we might expect the non-interacting model to exhibit qualitatively similar behavior to that of general models in a disordered phase which flow under RG to non-interacting systems. Furthermore, in the toy model settings we study, the algorithms are constrained to remain within the vicinity of the optimum. Hence, our zeroth-order bounds do not apply to algorithms which may operate far from the vicinity of the optimum to which they are trying to converge. Indeed, in some cases analytic gradient measurements may be performed by performing multiple "non-local" energy measurements and combining the results [5, 23, 24].

Prior work.—Prior works had considered gradient measurements in variational algorithms, but it remained unclear whether they could confer an advantage. It was

first observed that gradients could be directly measured in the context of hybrid quantum-classical algorithms in [2], but the authors pointed out that "it is not clear whether or not access to the derivative would improve the optimization". Many subsequent works [5, 7, 16–19, 23–34] have proposed using gradient measurements in variational algorithms for specific applications, but lacked concrete theoretical evidence for an advantage over zeroth-order algorithms; our work complements these proposals by providing such evidence. In [15], algorithms based on gradient measurements were numerically compared against zeroth-order algorithms for the combinatorial optimization problem MaxCut. Interestingly, the authors found no advantage for gradient-measurement-based algorithms for this problem. The discrepancy between our results and theirs could be explained by the possibility that their simulations were dominated by the cost of *finding* good local optima rather than *converging* to a specific local optima (as is our focus in this Letter). Similar questions about the benefit of noisy gradients for optimization had previously been studied in the purely classical context [35, 36], but fundamental differences between the classical and quantum variational settings prevented these results from being directly applicable in the present setting. Nonetheless, our strategy for proving an advantage for gradient-based variational algorithms adapts some techniques developed in these works, which in turn are inspired by methods from statistical minimax and learning theory.

Black-box model.—We now discuss our rigorous separation between the performance of zeroth-order and first-order variational algorithms. As a prerequisite, we first introduce a black-box model for variational algorithms. To see why such a setting is useful, note that a classical computer with unbounded computational resources is capable of simulating any hybrid quantum-classical algorithm; in this sense, no quantum measurements are required. Of course, this simulation will generally require exponential space and runtime, and therefore be intractable. Since one of our goals is to prove lower bounds on the number of quantum measurements required by a variational algorithm, we must impose extra constraints on the classical component of the algorithm to rule out such brute-forcing behavior. A natural way to do this which is also amenable to theoretical analysis is via a model in which the 'quantum' component of the algorithm is only accessible via queries to a black box. Note that our motivation for a black-box model is analogous to the motivation for a black-box model in the context of purely classical optimization, where it has been found to be a highly useful and insightful framework [20].

We now describe the black box setting. We assume the classical outer loop is not given an explicit description of the objective observable H, but rather has access to an oracle $\mathcal{O}_H$ encoding H. Suppose $H = E\,\mathbb{E}_L\,P_L$ as above. The classical outer loop may query $\mathcal{O}_H$ with a variational state ansatz description Θ, a parameter $\boldsymbol{\theta} \in \mathbb{R}^p$, and optionally an index $j \in [p]$. Upon querying $\mathcal{O}_H$ without the optional index, which we call a *zeroth-order query*, the oracle prepares the variational state $|\boldsymbol{\theta}\rangle$ according to the ansatz described by Θ and outputs an unbiased $\pm E$-valued estimate of $f(\boldsymbol{\theta})$ following the sampling approach described above. Similarly, upon querying $\mathcal{O}_H$ with index j, which we call a *first-order query*, the oracle outputs an unbiased $\pm\Gamma_j$-valued estimate of $\nabla_j f(\boldsymbol{\theta})$ following the sampling approach. Higher-order queries to the oracle may be defined analogously, as described explicitly in the Supplemental Material. In the black-box setting, we say an algorithm is k^{th}*-order* if it only makes queries of order k or lower.

In this oracle model, following the classical optimization literature [20], the classical outer loop is given black-box access to $\mathcal{O}_H$ and may be promised that H belongs to some family $\mathcal{H}$, but is not given explicit knowledge of H. The relevant performance metric of an optimization algorithm in this setting is the query complexity, that is, the number of oracle calls made by the classical algorithm. If the oracles are implemented physically via the observable sampling procedures described above, the query complexity exactly corresponds to the number of quantum state preparations and measurements performed.

To formally state our separation between zeroth- and first-order variational algorithms, it will be necessary to make some additional definitions. Let $\mathcal{H}$ be some fixed set of objective observables, and suppose $\mathcal{A}$ is a (possibly randomized) classical algorithm which has oracle access to $H \in \mathcal{H}$ and outputs a (generally random) description of a quantum state $|\psi\rangle$ from some distribution $\mathcal{D}_H$ which may depend on H. Then the optimization error of $\mathcal{A}$ with respect to $\mathcal{H}$, $\mathrm{Err}(\mathcal{A}, \mathcal{H})$, is defined as

$$\mathrm{Err}(\mathcal{A}, \mathcal{H}) := \sup_{H \in \mathcal{H}} \mathop{\mathbb{E}}_{\phi \sim \mathcal{D}_H} [\, \langle \phi | H | \phi \rangle - \lambda_{\min}(H)],$$

where $\lambda_{\min}(H)$ is the smallest eigenvalue of H, and the expectation is over the possible randomness of the output state $|\phi\rangle$. That is, $\mathrm{Err}(\mathcal{A}, \mathcal{H})$ quantifies the worst-case (over $H \in \mathcal{H}$) expected optimization error of $\mathcal{A}$. In some cases, we will be particularly interested in the setting in which the variational algorithm is close to an optimum and is trying to converge. To this end, it is helpful to define $\mathcal{A}$ to be a δ*-vicinity algorithm* with respect to $\mathcal{H}$ if $\mathcal{A}$ only queries the oracle with descriptions of variational states in the δ*-optimum* of $\mathcal{H}$; this defined to be the set of states $|\boldsymbol{\theta}\rangle$ such that $\langle \boldsymbol{\theta} | H | \boldsymbol{\theta} \rangle - \lambda_{\min}(H) \leq \delta$ for some $H \in \mathcal{H}$.

We now introduce the parameterized family of objective observables which we use to prove our sample complexity separation. First, for any $\delta \in \mathbb{R}$ and $v \in \{-1, 1\}^n$, define the n-qubit observable

$$H_v^\delta := -\sum_{i=1}^{n} \Big[\sin\Big(\frac{\pi}{4} + v_i \delta\Big) X_i + \cos\Big(\frac{\pi}{4} + v_i \delta\Big) Z_i \Big],$$

where X_i (Z_i) denotes the Pauli X (Z) operator acting on qubit i. Now, for a fixed parameter $\epsilon > 0$ we define $\delta(\epsilon) := \sqrt{\frac{45\epsilon}{n}}$ and

$$\mathcal{H}_n^\epsilon := \{H_v^{\delta(\epsilon)} \ : \ \forall v \in \{-1,1\}^n\}.$$

We prove lower and upper bounds on the query cost of finding a low-energy state w.r.t. observables in of the family $\mathcal{H}_n^\epsilon$.

Lower bounds.—We now state our lower bound for zeroth-order variational algorithms. (The numerical constants are chosen for ease of proof and have not been carefully optimized.)

Theorem 1 (Lower bound for zeroth-order methods)**.** *For any $n > 15$ and $\epsilon < 0.01n$, let $\mathcal{A}$ be any zeroth-order 100ϵ-vicinity algorithm for the family $\mathcal{H}_n^\epsilon$ that makes T queries to the oracle. Then, if $\mathrm{Err}(\mathcal{A}, \mathcal{H}_n^\epsilon) \leq \epsilon$, it must hold that $T \geq \Omega(\frac{n^3}{\epsilon^2})$ where the implicit factor is some fixed constant.*

The proof of Theorem 1 is information-theoretic, and may be found in Section IV of the Supplemental Material [13]. We choose a set $\mathcal{M} \subset \mathcal{H}_n^\epsilon$ that is both large and has well-separated points, then run $\mathcal{A}$ on a randomly chosen $H \in \mathcal{M}$. Since the points in $\mathcal{M}$ are sufficiently well separated, if $\mathrm{Err}(\mathcal{A}, \mathcal{H}_n^\epsilon) \leq \epsilon$, we can unambiguously distinguish which $H \in \mathcal{M}$ we are given. On the other hand, if $\mathcal{M}$ is large then learning this information means that the oracle outputs must have large mutual information with the identity of H (via Fano's inequality [37]; indeed our strategy is also known as Fano's method). Finally, in the vicinity of the ground state the output distributions produced by zeroth-order queries to $\mathcal{O}_H$ and $\mathcal{O}_{H'}$ for any $H, H' \in \mathcal{M}$ have small relative entropy, which implies an upper bound on the amount of mutual information obtained by each oracle query. Putting this together yields a lower bound on the number of queries needed to optimize $\mathcal{H}_n^\epsilon$ with error ϵ.

Theorem 1 gives a lower bound for *zeroth-order* variational algorithms *restricted to the vicinity* of the optimum. Upon lifting these two restrictions, we obtain a more general lower bound following a similar proof strategy. The primary difference is that now, for this unrestricted case, the oracle output distributions associated with two different $H, H' \in \mathcal{M}$ may be more distinguishable, yielding a weaker lower bound.

Theorem 2 (General lower bound)**.** *For any $n > 15$, $\epsilon < 0.01n$, and $k \in \mathbb{Z}_+$, suppose $\mathcal{A}$ is a k^{th}-order algorithm that makes T queries and satisfies $\mathrm{Err}(\mathcal{A}, \mathcal{H}_n^\epsilon) \leq \epsilon$. Then $T \geq \Omega(\frac{n^2}{\epsilon})$.*

Upper bounds.—The arguments above indicate that zeroth-order measurements taken in the vicinity of the optimum may be less informative in some sense than more general measurements. *A priori*, it is unclear if this observation translates into an algorithmic advantage for variational algorithms making gradient measurements. To this end, we show that a first-order algorithm based on SGD can attain an upper bound which matches the lower bound of Theorem 2, even when restricted to the vicinity of an optimum. Hence, not only does this show that a first-order algorithm can converge faster than any zeroth-order algorithm in the vicinity of the optimum, but it also shows that for the specific problem under consideration, the first-order SGD-based algorithm is in fact essentially optimal among all k^{th}-order algorithms for any k. This result is stated as the following theorem.

Convexity of $f(\boldsymbol{\theta})$	Zeroth-order	SGD	SMD
Convex	$\min\left(\frac{p^{32}E^2}{\epsilon^2}, \frac{p^2E^4(R_2/r_2)^2}{\epsilon^4}\right)$	$\frac{R_2^2\|\vec{\Gamma}\|_1^2}{\epsilon^2}$	$\frac{R_2^2\|\vec{\Gamma}\|_2^2}{\epsilon^2}$
λ_2-strongly convex w.r.t. $\|\cdot\|_2$	$\min\left(\frac{p^{32}E^2}{\epsilon^2}, \frac{p^2E^4(R_2/r_2)^2}{\epsilon^4}\right)$	$\frac{\|\vec{\Gamma}\|_1^2}{\lambda_2\epsilon}$	$\frac{p\|\vec{\Gamma}\|_2^2}{\lambda_2\epsilon}$
λ_1-strongly convex w.r.t. $\|\cdot\|_1$	$\min\left(\frac{p^{32}E^2}{\epsilon^2}, \frac{p^2E^4(R_2/r_2)^2}{\epsilon^4}\right)$	$\frac{\|\vec{\Gamma}\|_1^2}{\lambda_1\epsilon}$	$\frac{\|\vec{\Gamma}\|_2^2}{\lambda_1\epsilon}$

TABLE I. Rigorous upper bounds for the query complexity of optimizing $f(\boldsymbol{\theta})$ to precision ϵ in a convex region $\mathcal{X} \subset \mathbb{R}^p$ contained in a 2-ball of radius R_2, contained in an 1-ball of radius R_1, and containing a 2-ball of radius r_2, using zeroth-order strategies or gradient measurements in conjunction with SGD or SMD with an l_1 setup. Constants, logarithmic factors, and some Lipschitz constants are hidden for clarity (see [13] for full details and caveats). For the family $\mathcal{H}_n^\epsilon$, and with respect to the variational ansatz used in our proof of Theorem 3, we have $p = n$, $E = \Theta(n)$, $\Gamma_i = \Theta(1)$, $R_2 = \Theta(\sqrt{\epsilon})$, $R_1 = \Theta(\sqrt{\epsilon n})$, $r_2 = O(\sqrt{\epsilon/n})$, $\lambda_2 = \Theta(1)$, and $\lambda_1 = \Theta(1/n)$.

Theorem 3 (Upper bound for first-order methods)**.** *For any $\epsilon < 0.01n$, there exists a first-order, 100ϵ-vicinity algorithm $\mathcal{A}$ based on SGD that makes $O(\frac{n^2}{\epsilon})$ queries and achieves an error $\mathrm{Err}(\mathcal{A}, \mathcal{H}_n^\epsilon) \leq \epsilon$.*

En route to showing this theorem, we first obtain general upper bounds on the query cost of variational algorithms in the vicinity of a local minimum, reported in Table I. More precisely, the bounds are applicable when the induced objective function f is known to be convex within some fixed convex feasible set. They are obtained by combining objective function or gradient estimators with known convergence results [20] from the theory of stochastic optimization. In particular, the SGD bounds utilize the estimator $\hat{\mathbf{g}}(\boldsymbol{\theta})$ defined previously (note that $\hat{\mathbf{g}}(\boldsymbol{\theta})$ can be constructed from a single first-order oracle query). While Theorem 3 will only require an SGD bound, we also report bounds based on stochastic mirror descent (SMD), as well as (for comparison) zeroth-order algorithms. The zeroth-order bounds, based on [38, 39], are the best rigorous bounds we are aware of, but would likely be outperformed in practice. SMD [20] is a non-Euclidean generalization of SGD; the SMD bounds we report are based on taking the norm in parameter space to be the 1-norm rather than the Euclidean 2-norm, as

is the case for SGD. Further background on these algorithms, motivation for considering SMD, and full derivation of the bounds in Table I may be found in Section III of the Supplemental Material [13].

We now describe an algorithm $\mathcal{A}$ attaining the upper bound in Theorem 3. We refer the reader to Section IV of the Supplemental Material [13] for full technical details of the argument. Start by fixing the following n-parameter variational ansatz Θ:

$$|\theta_1, \ldots, \theta_n\rangle := \exp\left(-i\sum_{j=1}^{n}(\theta_j + \pi/4)Y_j/2\right)|0\rangle^{\otimes n}.$$

This parameterization has a simple geometric interpretation: $|\theta\rangle$ is the product state on n qubits for which the polarization of qubit j is $\sin(\pi/4+\theta_j)\hat{x} + \cos(\pi/4+\theta_j)\hat{z}$.

Now, consider some objective observable $H_v^\delta \in \mathcal{H}_n^\epsilon$. The induced objective function $f(\theta)$ is found to be $f(\theta) = \langle\theta|H_v^\delta|\theta\rangle = -\sum_{i=1}^n \cos(\theta_i - \delta v_i)$. Let $\mathcal{B}_\infty(\delta) \subset \mathbb{R}^n$ denote the ∞-ball of radius δ centered at the origin. That is, $\mathcal{B}_\infty(\delta) = \{\theta : \max(\theta_1, \theta_2, \ldots, \theta_n) \leq \delta\}$. Note that the ground state of H_v^δ is the state $|\delta v_1, \delta v_2, \ldots, \delta v_n\rangle$, and hence corresponds to a parameter inside the set $\mathcal{B}_\infty(\delta)$ for any choice of v. Furthermore, the set of states associated with $\mathcal{B}_\infty(\delta)$ is contained in the 100ϵ-optimum of $\mathcal{H}_n^\epsilon$, and the induced objective function $f(\theta)$ is 0.01-strongly convex w.r.t. the 2-norm (strong convexity is reviewed in the Supplemental Material). It is also straightforward to show that, for this problem, $\|\vec{\Gamma}\|_1 = O(n)$. Theorem 3 now follows from the SGD upper bound for strongly convex functions in Table I, taking $\mathcal{B}_\infty(\delta)$ as the feasible set. We note that SMD with a 1-norm setup achieves an identical performance for this toy problem, up to logarithmic factors.

Conclusion.—Our results provide theoretical evidence that taking analytic gradient measurements in variational algorithms can be advantageous, supporting recent gradient-based proposals. We expect the rigorous upper bounds we report in Table I may be helpful in guiding expectations on the performance of gradient-based variational algorithms for particular classes of problems, even if more heuristic algorithms may be used in practice. To this end, an interesting direction for future work is to understand how the parameters appearing in Table I behave for various problems of practical interest. Further discussion, open questions, and comparison with the literature may be found in Section V of the Supplemental Material.

We thank Xiaodi Wu for helpful discussions. JN and AWH were funded by ARO contract W911NF-17-1-0433 and NSF grants CCF-1729369 and PHY-1818914. AWH was also funded by NSF grant CCF-1452616 and the MIT-IBM Watson AI Lab under the project *Machine Learning in Hilbert space*.

* aram@mit.edu
† napp@mit.edu

[1] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. Obrien, Nature Communications **5**, 4213 (2014).
[2] D. Wecker, M. B. Hastings, and M. Troyer, Physical Review A **92**, 042303 (2015).
[3] E. Farhi, J. Goldstone, and S. Gutmann, arXiv:1411.4028 (2014).
[4] E. Farhi and H. Neven, arXiv:1802.06002 (2018).
[5] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Physical Review A **98**, 032309 (2018).
[6] M. Schuld and N. Killoran, Physical Review Letters **122**, 040504 (2019).
[7] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, Physical Review A **101**, 032308 (2020).
[8] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Nature **567**, 209 (2019).
[9] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, New Journal of Physics **18**, 023023 (2016).
[10] Z.-C. Yang, A. Rahmani, A. Shabani, H. Neven, and C. Chamon, Physical Review X **7**, 021027 (2017).
[11] A. Bapat and S. Jordan, Quantum Information & Computation **19**, 424 (2019).
[12] E. Campbell, Phys. Rev. Lett. **123**, 070503 (2019).
[13] See Supplemental Material at [URL] for full details of proofs and derivations, stochastic optimization preliminaries, and further discussion.
[14] Y. Li and S. C. Benjamin, Physical Review X **7**, 021050 (2017).
[15] G. G. Guerreschi and M. Smelyanskiy, arXiv:1701.01450 (2017).
[16] J. Romero, R. Babbush, J. McClean, C. Hempel, P. Love, and A. Aspuru-Guzik, Quantum Science and Technology (2018).
[17] R. Sweke, F. Wilde, J. Meyer, M. Schuld, P. K. Fährmann, B. Meynard-Piganeau, and J. Eisert, arXiv preprint arXiv:1910.01155 (2019).
[18] A. Arrasmith, L. Cincio, R. D. Somma, and P. J. Coles, arXiv preprint arXiv:2004.06252 (2020).
[19] J. M. Kübler, A. Arrasmith, L. Cincio, and P. J. Coles, Quantum **4**, 263 (2020).
[20] S. Bubeck, Foundations and Trends in Machine Learning **8**, 231 (2015).
[21] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Nature **549**, 242 (2017).
[22] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Nature Communications **9**, 1 (2018).
[23] J. Li, X. Yang, X. Peng, and C.-P. Sun, Phys. Rev. Lett. **118**, 150503 (2017).
[24] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Physical Review A **99**, 032331 (2019).
[25] J.-G. Liu and L. Wang, Phys. Rev. A **98**, 062324 (2018).
[26] M. Benedetti, E. Grant, L. Wossnig, and S. Severini, New Journal of Physics **21**, 043023 (2019).
[27] S. Khatri, R. LaRose, A. Poremba, L. Cincio, A. T. Sornborger, and P. J. Coles, Quantum **3**, 140 (2019).
[28] J. Romero and A. Aspuru-Guzik, arXiv preprint arXiv:1901.00848 (2019).

[29] J.-G. Liu, Y.-H. Zhang, Y. Wan, and L. Wang, Phys. Rev. Research **1**, 023025 (2019).

[30] C. Zoufal, A. Lucchi, and S. Woerner, npj Quantum Information **5**, 103 (2019).

[31] X. Xu, J. Sun, S. Endo, Y. Li, S. C. Benjamin, and X. Yuan, arXiv preprint arXiv:1909.03898 (2019).

[32] C. Bravo-Prieto, R. LaRose, M. Cerezo, Y. Subasi, L. Cincio, and P. J. Coles, arXiv preprint arXiv:1909.05820 (2019).

[33] S. Lu, L.-M. Duan, and D.-L. Deng, arXiv preprint arXiv:2001.00030 (2019).

[34] M. Cerezo, K. Sharma, A. Arrasmith, and P. J. Coles, arXiv preprint arXiv:2004.01372 (2020).

[35] A. Agarwal, M. J. Wainwright, P. L. Bartlett, and P. K. Ravikumar, in *Advances in Neural Information Processing Systems* (2009) pp. 1–9.

[36] K. G. Jamieson, R. Nowak, and B. Recht, in *Advances in Neural Information Processing Systems* (2012) pp. 2672–2680.

[37] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, NY, USA, 1991).

[38] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, in *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms* (Society for Industrial and Applied Mathematics, 2005) pp. 385–394.

[39] A. Agarwal, D. P. Foster, D. J. Hsu, S. M. Kakade, and A. Rakhlin, in *Advances in Neural Information Processing Systems* (2011) pp. 1035–1043.

[40] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, SIAM Journal on Optimization **19**, 1574 (2009).

[41] A. Juditsky and A. Nemirovski, Optimization for Machine Learning , 121 (2011).

[42] E. Hazan and S. Kale, The Journal of Machine Learning Research **15**, 2489 (2014).

[43] A. Agarwal and O. Dekel, in *COLT* (Citeseer, 2010) pp. 28–40.

[44] O. Shamir, in *Conference on Learning Theory* (2013) pp. 3–24.

[45] A. Gilyén, S. Arunachalam, and N. Wiebe, Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms , 1425 (2019).

[46] S. P. Jordan, Physical Review Letters **95**, 050501 (2005).

[47] A. Guntuboyina, IEEE Transactions on Information Theory **57**, 2386 (2011).

[48] M. Raginsky and A. Rakhlin, IEEE Transactions on Information Theory **57**, 7036 (2011).

[49] S.-i. Amari, Neural Computation **10**, 251 (1998), https://doi.org/10.1162/089976698300017746.

[50] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum **4**, 269 (2020).

[51] B. Koczor and S. C. Benjamin, arXiv preprint arXiv:1912.08660 (2019).

[52] B. van Straaten and B. Koczor, arXiv preprint arXiv:2005.05172 (2020).

[53] S. McArdle, T. Jones, S. Endo, Y. Li, S. C. Benjamin, and X. Yuan, npj Quantum Information **5**, 75 (2019).