Quantum Lower Bounds for Finding Stationary Points of Nonconvex Functions

Chenyi Zhang [1 2] **Tongyang Li** [3 4]

Abstract

Quantum computing is an emerging technology that has been rapidly advancing in the past decades. In this paper, we conduct a systematic study of quantum lower bounds on finding ϵ-approximate stationary points of nonconvex functions, and we consider the following two important settings: 1) having access to p-th order derivatives; or 2) having access to stochastic gradients. The classical query lower bounds are $\Omega\big(\epsilon^{-\frac{1+p}{p}}\big)$ regarding the first setting and $\Omega(\epsilon^{-4})$ regarding the second setting (or $\Omega(\epsilon^{-3})$ if the stochastic gradient function is mean-squared smooth). In this paper, we extend all these classical lower bounds to the quantum setting. They match the classical algorithmic results respectively, demonstrating that there is no quantum speedup for finding ϵ-stationary points of nonconvex functions with p-th order derivative inputs or stochastic gradient inputs, whether with or without the mean-squared smoothness assumption. Technically, we prove our quantum lower bounds by showing that the sequential nature of classical hard instances in all these settings also applies to quantum queries, preventing any quantum speedup other than revealing information of the stationary points sequentially.

1. Introduction

Quantum computing is an emerging technology with wide applications. Among those, quantum algorithms for optimization are of general interest. On the one hand, optimization algorithms have wide applications in machine learning, statistics, operations research, and many other areas. On the other hand, it is crucial in quantum computing to figure out the extent of quantum speedups in specific problems, and previous literature had established quantum speedups for solving linear systems (Harrow et al., 2009; Childs et al., 2017), semidefinite programs (Brandão & Svore, 2017; Brandão et al., 2019; van Apeldoorn et al., 2020b; van Apeldoorn & Gilyén, 2019; Kerenidis & Prakash, 2020), general convex optimization (Chakrabarti et al., 2020; van Apeldoorn et al., 2020a), etc.

More recently, nonconvex optimization has been a primary research direction in machine learning, since the landscapes of many models, including neural networks, are typically nonconvex. Finding the global optimum of a nonconvex function, even approximately, is NP-hard in general (Murty & Kabadi, 1987; Nemirovski & Yudin, 1983). To give efficient optimization algorithms for nonconvex functions, a first step is to find stationary points (Agarwal et al., 2017; Birgin et al., 2017; Carmon et al., 2018; 2017; Nesterov, 2003; Nesterov & Polyak, 2006). However, quantum algorithms for nonconvex optimization are less understood. Based on gradient descents, Zhang et al. (2021) proposed a quantum algorithm that can find an ϵ-approximate second-order stationary point of a d-dimensional nonconvex function and improve the logarithmic dimension dependence from $\log^6 d$ in the classical result (Jin et al., 2018) to $\log^2 d$, but the ϵ dependence remains the same. The dependence in $\log d$ is further improved to linear in Childs et al. (2022). Moreover, Liu et al. (2022) showed that quantum tunneling can provide quantum speedups in the task of finding an unknown local minimum starting from a known one.

Meanwhile, various results have been developed concerning the classical lower bounds for finding ϵ-approximate first-order stationary points (i.e., points with gradient smaller than ϵ) of nonconvex functions under different assumptions. In particular, Carmon et al. (2020a) discussed the setting where the objective function f has Lipschitz p-th order derivative and proved that any randomized classical algorithm has to make at least $\Omega\big(\epsilon^{-\frac{1+p}{p}}\big)$ derivative queries to guarantee an $\Omega(1)$ success probability in the worst case. Using a similar approach, Carmon et al. (2021) proved a deterministic classical lower bound for first-order method of order $\Omega(\epsilon^{-12/7})$ for functions with Lipschitz first and second derivatives.

As for stochastic settings, Arjevani et al. (2020; 2022) thoroughly investigated classical lower bounds under different

[1]Computer Science Department, Stanford University [2]Institute for Interdisciplinary Information Sciences, Tsinghua University [3]Center on Frontiers of Computing Studies, Peking University [4]School of Computer Science, Peking University. Correspondence to: Tongyang Li <tongyangli@pku.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

assumptions with or without the mean-squared smoothness property, and with access to different orders of stochastic derivatives. In particular, the query lower bound for stochastic first-order methods is $\Omega(\epsilon^{-4})$ (Arjevani et al., 2022). If the stochastic gradient additionally satisfies the mean-squared smoothness property, the query lower bound would be of order $\Omega(\epsilon^{-3})$ (Arjevani et al., 2022), which is also the query lower bound in the case where we have access to second- and higher-order stochastic derivatives (Arjevani et al., 2020). Nevertheless, for objective functions with Lipschitz p-th oder derivative, the query lower bound remains $\Omega(\epsilon^{-3})$ for stochastic p-th order methods if we have the mean-squared smoothness property (Arjevani et al., 2020).

However, despite recent progress on quantum lower bounds for convex optimization (Garg et al., 2020; 2021), quantum lower bounds for nonconvex optimization are widely open.

Contributions We conduct a systematic study of quantum lower bounds for finding an ϵ-stationary point of a nonconvex objective function f, i.e., finding an $\mathbf{x} \in \mathbb{R}^d$ such that

$$\|\nabla f(\mathbf{x})\| \leq \epsilon.$$

For optimization problems with deterministic queries, high-order methods are of general interest (Bubeck et al., 2019b; Gasnikov et al., 2019), which compared to first-order methods can achieve better convergence rate by exploiting higher-order smoothness (Birgin et al., 2017; Cartis et al., 2010; Nesterov & Polyak, 2006). Beyond that, another widely considered setting is having access to stochastic gradients, which is widely applied in modern machine learning tasks (Bottou & Bousquet, 2007; Bottou et al., 2018) as it only requires access to an unbiased gradient estimator. Various classical algorithms have been developed under this setting from variants of stochastic gradient descent (SGD) (Jin et al., 2021; Fang et al., 2018; Zhang & Li, 2021; Zhou et al., 2018) to more advanced methods (Kingma & Ba, 2015; Liu et al., 2018; Duchi et al., 2011). Hence, we study the quantum query lower bounds for finding and ϵ-stationary point under the following two important settings:

1. having access to derivatives of a p-th order Lipschitz function;
2. having access to stochastic gradients of a Lipschitz function without the mean-squared smoothness assumption; or
3. having access to stochastic gradients of a Lipschitz function that additionally satisfy the mean-squared smoothness assumption.

For the first setting , we consider a C^∞ function $f\colon \mathbb{R}^d \to \mathbb{R}$ with L_p-Lipschitz p-th derivative, i.e., $\|\nabla^p f(\mathbf{x})\| \leq L_p$. We define the p-th order response to a query at point $\mathbf{x}$ to be

$$\nabla^{(0,\dots,p)} f(\mathbf{x}) := \{f(\mathbf{x}), \nabla f(\mathbf{x}), \dots, \nabla^p f(\mathbf{x})\}, \tag{1}$$

which we assume can be accessed via the *quantum evaluation oracle* defined as a unitary map on $\mathbb{R}^d \to \mathbb{R}^{d^0+\cdots+d^p}$ such that for any $\mathbf{x} \in \mathbb{R}^d$,

$$O_f^{(p)} |\mathbf{x}\rangle |y\rangle \to |\mathbf{x}\rangle \left| y \oplus \nabla^{(0,\dots,p)} f(\mathbf{x}) \right\rangle. \tag{2}$$

Here, the Dirac notation $|\cdot\rangle$ denotes the register storing quantum states. Specifically, for any $m \in \mathbb{N}$, $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^d$, and $\mathbf{c} \in \mathbb{C}^m$ such that $\sum_{i=1}^m |c_i|^2 = 1$,

$$O_f^{(p)} \Big(\sum_{i=1}^m c_i |\mathbf{x}_i\rangle \otimes |\mathbf{0}\rangle \Big) = \sum_{i=1}^m c_i |\mathbf{x}_i\rangle \otimes \left| \nabla^{(0,\dots,p)} f(\mathbf{x}_i) \right\rangle.$$

If we measure this quantum state, we get $\nabla^{(0,\dots,p)} f(\mathbf{x}_i)$ with probability $|c_i|^2$. Compared to the classical evaluation oracle (i.e., $m = 1$), the quantum evaluation oracle can query different locations in *superposition*, which is the essence of speedups from quantum algorithms. In addition, if we can implement the classical evaluation oracle by arithmetic circuits, the quantum evaluation oracle can be implemented by quantum arithmetic circuits of about the same size (see Footnote 2 of Chakrabarti et al. 2023). Hence, it has been the standard assumption in previous quantum computing literature for convex optimization (van Apeldoorn et al., 2020a; Chakrabarti et al., 2020) and nonconvex optimization (Liu et al., 2022; Zhang et al., 2021), with different magnitudes of quantum speedups obtained.

Despite that the quantum evaluation oracle is powerful by taking queries in superposition, however, we show that it cannot provide quantum speedup for finding stationary points of nonconvex functions. In particular, we prove the following quantum query lower bound.

Theorem 1.1 (Informal version of Theorem 2.4)**.** *For any $\epsilon > 0$ and $p \in \mathbb{N}$, there exists a family $\mathcal{F}$ of functions $f\colon \mathbb{R}^d \to \mathbb{R}$ with L_p-Lipschitz p-th derivative such that any quantum algorithm that finds an ϵ-stationary point of any $f \in \mathcal{F}$ must make*

$$\Omega\left(L_p^{1/p} \epsilon^{-(p+1)/p}\right)$$

queries to the quantum p-th order oracle.

For the second setting where we have access to stochastic gradients, we also consider a C^∞ function $f\colon \mathbb{R}^d \to \mathbb{R}$ that is L-Lipschitz, i.e., $\|\nabla f(\mathbf{x})\| \leq L$. Assume the stochastic gradient $\mathbf{g}(\mathbf{x}, \xi)$ of f indexed by some random seed ξ satisfies

$$\begin{cases} \mathbb{E}_\xi[\mathbf{g}(\mathbf{x}, \xi)] = \nabla f(\mathbf{x}), \\ \mathbb{E}_\xi \|\mathbf{g}(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|^2 \leq \sigma^2, \end{cases}$$

for some constant σ, which can be accessed via the following quantum oracle

$$O_{\mathbf{g}} |\mathbf{x}\rangle |\xi\rangle |\mathbf{v}\rangle = |\mathbf{x}\rangle |\xi\rangle |\mathbf{g}(\mathbf{x}, \xi) + \mathbf{v}\rangle . \tag{3}$$

Then, we can prove the following result.

Theorem 1.2 (Informal version of Theorem 3.6)**.** *For any $\epsilon, \sigma > 0$, there exists a family $\mathcal{F}$ of L-gradient Lipschitz functions $f\colon \mathbb{R}^d \to \mathbb{R}$ such that any quantum algorithm that finds an ϵ-stationary point of any $f \in \mathcal{F}$ must make*

$$\Omega\left(\frac{\min\{L^2\Delta^2, \sigma^4\}}{\epsilon^4}\right)$$

queries to the quantum stochastic gradient oracle.

Moreover, various literature on nonconvex stochastic optimization (Fang et al., 2018; Lei et al., 2017; Zhou et al., 2018) has considered the following additional mean-squared smoothness assumption on the stochastic gradient $\mathbf{g}(\mathbf{x}, \xi)$.

Assumption 1.3. The stochastic gradient $\mathbf{g}$ satisfies that for some constant $\bar{L}$, we have that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\mathbb{E}_\xi \|\mathbf{g}(\mathbf{x}, \xi) - \mathbf{g}(\mathbf{y}, \xi)\|^2 \le \bar{L}^2 \cdot \|\mathbf{x} - \mathbf{y}\|^2.$$

Under the stochastic optimization setting where Assumption 1.3 is satisfied, we prove the following result.

Theorem 1.4 (Informal version of Theorem 3.8)**.** *For any $\epsilon, \sigma > 0$, there exists a family $\mathcal{F}$ of $\bar{L}$-gradient Lipschitz functions $f\colon \mathbb{R}^d \to \mathbb{R}$ such that for any quantum algorithm that finds an ϵ-stationary point of any $f \in \mathcal{F}$ must make $\Omega(\bar{L}\sigma/\epsilon^3)$ queries to the quantum stochastic gradient oracle satisfying the mean-squared smoothness assumption.*

Observe that our quantum lower bounds match the classical algorithmic results (Birgin et al., 2017; Fang et al., 2018; Jin et al., 2021) concerning corresponding settings. Therefore, we essentially prove that **there is no quantum speedup for finding stationary points of nonconvex functions with p-th order derivative inputs or stochastic gradient inputs, whether with or without Assumption 1.3.**

Techniques Inspired by both the classical lower bound results for finding stationary points (Arjevani et al., 2022; Carmon et al., 2020a) as well as the techniques introduced in Garg et al. (2020; 2021) on quantum lower bounds for convex optimization, our work utilizes the underlying similarities and connects these two settings. In particular, the proof of our quantum lower bounds has the following key technical components:

1. Adopt a hard function that is a robust zero-chain and has a "non-informative region" around $\mathbf{0}$, which contains a sequential underlying structure that can only be discovered via adaptive queries.
2. Represent quantum algorithms by sequences of unitaries.
3. Demonstrate that the sequential nature of the robust zero-chain can nullify the advantage of quantum algorithms to make queries in *superpositions*.

First, classical hard instances for finding stationary points of nonconvex functions under different settings share the same intuition originated from the following example proposed in Chapter 2.1.2 of Nesterov (2003),

$$f(\mathbf{x}) := \frac{1}{2}(x_1 - 1)^2 + \frac{1}{2}\sum_{i=1}^{T-1}(x_i - x_{i+1})^2. \tag{4}$$

For every component $i \in [T]$, $\nabla_i f(\mathbf{x}) = 0$ if and only if $x_{i-1} = x_i = x_{i+1}$. Then, if we query a point $\mathbf{x}$ with the first t entries being nonzero, the derivatives $\nabla^{(0,\dots,p)} f(\mathbf{x})$ can only reveal the $(t+1)$-th direction. Such f is called a *zero-chain* which is formally defined as follows.

Definition 1.5 (Carmon et al. 2020a, Definition 3)**.** For $p \in \mathbb{N}$, a function $f\colon \mathbb{R}^T \to \mathbb{R}$ is called a p-th order zero-chain if for every $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned}&\operatorname{supp}\{\mathbf{x}\} \subseteq \{1, \dots, i-1\}\\ &\quad\Rightarrow \bigcup_{q\in[p]} \operatorname{supp}\{\nabla^q f(\mathbf{x})\} \subseteq \{1, \dots, i\},\end{aligned}$$

where the support of a tensor $M \in \mathbb{R}^{\otimes^k T}$ is defined as

$$\operatorname{supp}\{M\} := \{i \in [d] \mid M_i \neq 0\}.$$

We say f is a zero-chain if it is a p-th-order zero-chain for every $p \in \mathbb{N}$.

Intuitively, if the objective function with an unknown set of coordinates is a zero-chain, and if we query its derivatives at point $\mathbf{x}$ with only its first i entries being nonzero, such query can only reveal information of the $(i+1)$-th coordinate, exhibiting a sequential nature. Hence, for any classical algorithm that never explores directions with zero derivatives components that seem not to affect the function, which is referred to as "zero-respecting algorithm" (Carmon et al., 2020a), it takes at least T queries to learn all the T coordinates. Moreover, we can observe that finding a stationary point of the function f defined in Eq. (4) requires complete knowledge of all the T directions, indicating that it takes at least T queries for a zero-respecting algorithm to find the stationary point of f.

Carmon et al. (2020a) extended this lower bound to randomized classical algorithms by constructing a hard instance following the intuition of the quadratic hard instance (4) and additionally **creating a "non-informative" region near $\mathbf{0}$ where small components have no impact on the function value, which can "trap" random perturbations** since

with overwhelming probability they can only create small magnitudes among unknown coordinates and thus has no influence on the function value as well as the algorithm.

Second, Garg et al. (2020; 2021) developed quantum lower bounds for convex optimization with non-smooth and smooth objective functions respectively, and demonstrate that there is no quantum speedup in both settings. The hard instance in Garg et al. (2020) is a variant of the shielded Nemirovski function introduced in Bubeck et al. (2019a) which takes a maximization over several component functions, each related to one of the coordinates and the component function related to the T-th coordinate is the least significant. Then with high probability, each query can reveal only one unknown coordinate with the smallest index. This property also applies to the smoothed hard instance in Garg et al. (2021). To obtain quantum lower bounds, Garg et al. (2020; 2021) **represented quantum algorithms in the form of sequences of unitaries**

$$\cdots V_3 O_f V_2 O_f V_1 O_f V_0$$

applied to the initial state $|0\rangle$, where O_f are the evaluation oracle of f and V_is are unitaries that are independent from f. The key step in their proof is demonstrating that, for any quantum algorithm A_{quan} making $k < T$ queries, if we replace all the k queries to O_f by new evaluation oracles that only partly agree with f but contains no information regarding the T-th coordinate, the output state of the algorithm will barely change. Since finding an ϵ-stationary point requires knowing all the coordinates, the new sequence of unitaries is hence not be able to find an ϵ-stationary point with high probability, so does the original quantum algorithm A_{quan}.

Third, we observe that although the hard instance in Garg et al. (2020; 2021) is a variant of the shielded Nemirovski function and has a different construction compared to the zero-chain hard instance introduced in Carmon et al. (2020a), it also satisfies the properties of robust zero-chains. We note that this connection has been utilized in previous works (Carmon et al., 2020b; Woodworth & Srebro, 2017) but has yet been pointed out explicitly. Similarly, quantum queries to the derivatives of the hard instance of Carmon et al. (2020a) also have only rather limited power, similar to the case in Garg et al. (2020; 2021). Conceptually, this is due to the fact that the hard instance in Carmon et al. (2020a) possesses **a sequential nature that the i-th coordinate direction only emerges when we reach a position that has a large overlap with the $(i-1)$-th direction, which nullifies the unique advantage of quantum algorithms to make queries in superpositions**.

As for the *stochastic setting*, similar to the classical stochastic lower bound result (Arjevani et al., 2022), we still use the classical hard instance defined in Carmon et al. (2020a) but with different scaling parameters. Nevertheless, the stochastic gradient function of the hard instance in Arjevani et al. (2022) could lead to a quantum speedup on this particular hard instance via Grover's search algorithm (Grover, 1996). To address this issue, inspired by the quantum lower bound on multivariate mean estimation (Cornelissen et al., 2022), we construct a new stochastic gradient function with details given in Section 3.3 such that it is also hard for quantum algorithms to obtain an accurate estimation of the exact gradient. Then, a quantum lower bound can be obtained matching the existing classical algorithmic upper bound result following the same procedure as Section 2.

Furthermore, if we assume that the stochastic gradient function satisfies the *mean-squared smoothness condition described in Assumption 1.3*, we can apply a similar version of the function smoothing technique introduced in Arjevani et al. (2022) to our stochastic gradient function (23) to obtain a "smoothed" stochastic gradient function, whose detailed formula is given in Section 3.4, upon which we can obtain a quantum query lower bound matching the existing classical algorithmic upper bound result given that the stochastic gradient function satisfies Assumption 1.3.

Recently, a simultaneous work by Gong et al. (2022) proved that finding an ϵ-stationary point of a nonconvex function with a noisy zeroth- and first-order inputs requires $\Omega(\epsilon^{-12/7})$ queries. Technically, they used the hard instance introduced in Carmon et al. (2021) that also has a sequential underlying structure such that coordinates can only be revealed sequentially due to the non-informative region near $\mathbf{0}$ created by noise. In contrast, in our setting we do not need external noise to create the non-informative region, which has fundamentally different intuitions.

Open Questions Our paper leaves several natural open questions for future investigation:

- Can we extend our stochastic quantum lower bounds for finding stationary points to higher-order methods?
- Can we extend our quantum lower bounds to the setting of finding approximate second-order stationary points (i.e., approximate local minima) with no additional overhead, or overhead being at most poly-logarithmic in ϵ and d?
- In this work, we show that classically hard optimization instances where information can only be revealed sequentially are also hard for quantum algorithms. Can we develop quantum lower bounds for other computational problems with sequential nature via similar techniques?

2. Quantum Lower Bound with Lipschitz p-th Order Derivatives

2.1. Function Classes and Classical Lower Bound

We consider the following set of objective functions.

Definition 2.1 (Carmon et al. 2020a, Definition 1)**.** Let $p \geq 1$, $\Delta > 0$ and $L_p > 0$. Then the set $\mathcal{F}_p(\Delta, L_p)$ denotes the union, over $d \in \mathbb{N}$, of the collection of C^∞ functions $f\colon \mathbb{R}^d \to \mathbb{R}$ with L_p-Lipschitz p-th derivative and $f(\mathbf{0}) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq \Delta$.

For any $f \in \mathcal{F}_p(\Delta, L_p)$, the response of a p-th order oracle to a query at point $\mathbf{x}$ is

$$\nabla^{(0,\dots,p)} f(\mathbf{x}) = \{f(\mathbf{x}), \nabla f(\mathbf{x}), \dots, \nabla^p f(\mathbf{x})\}$$

as defined in Eq. (1). Then for any dimension $d \in \mathbb{N}$, a classical algorithm A is defined as a map from objective functions $f \in \mathcal{F}_p(\Delta, L_p)$ to a sequence of iterates in $\mathbb{R}^d$, if for any $i \in \mathbb{N}$ it produces iterates of the form

$$\mathbf{x}^{(i)} = A^{(i)}\big(\nabla^{(0,\dots,p)} f(\mathbf{x}^{(1)}), \dots, \nabla^{(0,\dots,p)} f(\mathbf{x}^{(i-1)})\big)$$

where $A^{(i)}$ is a measurable mapping to $\mathbb{R}^d$. We refer to A as a classical p-th-order deterministic algorithm.

Similarly, a classical p-th-order randomized algorithm $A^{(p)}_{\text{rand}}$ as a distribution on p-th order deterministic algorithms. Quantitatively, $A^{(p)}_{\text{rand}}$ would produce iterates of the form

$$\mathbf{x}^{(i)} = A^{(i)}\big(\xi, \nabla^{(0,\dots,p)} f(\mathbf{x}^{(1)}), \dots, \nabla^{(0,\dots,p)} f(\mathbf{x}^{(i-1)})\big)$$

for any $i \in \mathbb{N}$, where ξ is a random uniform variable on $[0, 1]$, for some measurable mappings $A^{(i)}$ into $\mathbb{R}^d$.

The strategy of constructing a classical hard instance is to construct a high-dimensional zero-chain (Definition 1.5) such that the position of its stationary point is related to all the coordinates. In particular, Carmon et al. (2020a) considered the following hard instance $\bar{f}_T(\mathbf{x})\colon \mathbb{R}^d \to \mathbb{R}$,

$$\bar{f}_T(\mathbf{x}) = -\Psi(1)\Phi(x_1) + \sum_{i=2}^{T} [\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)], \quad (5)$$

where

$$\Psi(x) := \begin{cases} 0 & x \leq 1/2, \\ \exp\big(1 - \frac{1}{(2x-1)^2}\big) & x > 1/2, \end{cases}$$
$$\Phi(x) := \sqrt{e} \int_{-\infty}^{x} e^{-t^2/2} \mathrm{d}t,$$

for some $T > 0$. Note that $\bar{f}_T$ is a zero-chain whose derivative $\nabla \bar{f}_T(\mathbf{x})$ has a large norm unless $|x_i| \geq 1$ for all $i \in [T]$ (see Lemma A.2 for details). Thus, it takes at least $\Omega(T)$ queries for a classical algorithm to find a stationary point of $\bar{f}_T$ if it never explores directions with zero derivatives, which is referred to as a *zero-respecting algorithm* in Carmon et al. (2020a), where the authors also showed that the query complexity lower bound for zero-respecting algorithms also holds for all deterministic classical algorithms, and T can at most be of order $O\left(\Delta L_p^p \epsilon^{-(1+p)/p}\right)$ to satisfy L_p-Lipschitzness and other conditions, establishing query lower bound for all deterministic classical algorithms.

This deterministic lower bound is further extended to obtain a distributional complexity lower bounds of all randomized classical algorithms by a simple random orthogonal transformation (or intuitively, a high-dimensional random rotation) on the hard instance $\bar{f}_T$ defined in Eq. (5):

$$\tilde{f}_{T;U}(\mathbf{x}) \coloneqq \alpha \bar{f}_T(U^T \mathbf{x}/\beta) \quad (6)$$

where α and β are scaling constants, $U \in \mathbb{R}^{d \times T}$ with columns $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(T)}$ is an orthogonal matrix with $T \leq d$, and we assume throughout that U is chosen uniformly at random from the space of orthogonal matrices $O(d, T) = \{U \in \mathbb{R}^{d \times T} \,|\, U^T U = I_T\}$.

It is shown in Carmon et al. (2020a) that any random algorithm can "discover" at most one coordinate $\mathbf{u}^{(i)}$ per query with high probability. Quantitatively, for a random orthogonal matrix U, any sequence of bounded iterates $\{\mathbf{x}^{(t)}\}_{t \in \mathcal{N}}$ based on derivatives of $\tilde{f}_{T;U}$ must satisfy $|\langle \mathbf{x}^{(t)}, \mathbf{u}^{(j)} \rangle| \leq 0.5$ with high probability for all t and $j > t$. Then by Lemma A.2 in Appendix A, with high probability $\|\nabla \tilde{f}_{T;U}(\mathbf{x}^{(t)})\|$ is large for any $t \leq T$, and thus establishes the lower bound on randomized algorithms with access to bounded iterates. Moreover, Carmon et al. (2020a) showed that the boundedness of the iterates can be removed by composing $\tilde{f}_{T;U}$ with a soft projection to reach a lower bound for general unbounded iterates, and the query lower bound $\Omega\left(\Delta L_p^p \epsilon^{-(1+p)/p}\right)$ can be obtained for all randomized classical algorithms with access to p-th-order derivatives.

2.2. Quantum Query Model and Complexity Measures

We adopt the quantum query model introduced in Garg et al. (2020). For a function $f\colon \mathbb{R}^d \to \mathbb{R}$ with L_p-Lipschitz p-th derivative, we assume access to the following quantum oracle $O_f^{(p)}$ defined in Eq.(2):

$$O_f^{(p)} |\mathbf{x}\rangle |y\rangle \to |\mathbf{x}\rangle \left| y \oplus \nabla^{(0,\dots,p)} f(\mathbf{x}) \right\rangle,$$

for $\nabla^{(0,\dots,p)} f(\mathbf{x})$ defined in Eq. (1). Then for any p-th order quantum query algorithm A_{quan}, it can be described by the following sequence of unitaries

$$\cdots V_3 O_f^{(p)} V_2 O_f^{(p)} V_1 O_f^{(p)} V_0 \quad (7)$$

applied to an initial state, which can be set to $|0\rangle$ without loss of generality. Moreover, we define $A^{(t)}_{\text{quan}}$ to be the sequence (7) truncated before the $(t+1)$-th query to $O_f^{(p)}$,

$$A^{(t)}_{\text{quan}} := V_t O_f^{(p)} \cdots O_f^{(p)} V_1 O_f^{(p)} V_0,$$

for any $t \in \mathbb{N}$. Next, we extend the classical complexity measure introduced in Carmon et al. (2020a) to the quantum regime. Quantitatively,

Definition 2.2 (Quantum complexity measures)**.** For a function $f\colon \mathbb{R}^d \to \mathbb{R}$ and a sequence of quantum states $\{|\psi\rangle^{(t)}\}_{t\in\mathbb{N}}$, let p_t be the probability distribution over $\mathbf{x}\in\mathbb{R}^d$ obtained by measuring the state $|\psi\rangle^{(t)}$ in the computational basis $\{|\mathbf{x}\rangle \mid \mathbf{x}\in\mathbb{R}^d\}$. Then we can define

$$T_\epsilon\big(\{|\psi\rangle^{(t)}\}_{t\in\mathbb{N}}, f\big) := \inf\Big\{t\in\mathbb{N} \mid \Pr_{\mathbf{x}\sim p_t}\big(\|\nabla f(\mathbf{x})\|\le\epsilon\big)\ge\frac{1}{3}\Big\}.$$

To measure the performance of a quantum algorithm A_{quan} on function f, we define

$$T_\epsilon(A_{\mathrm{quan}}, f) := T_\epsilon\big(\{A^{(t)}_{\mathrm{quan}}|0\rangle\}, f\big)$$

as the complexity of A_{quan} on f. With this setup, we define the complexity of algorithm class $\mathcal{A}_{\mathrm{quan}}$ of all quantum algorithms in the form (7) on a function class $\mathcal{F}$ to be

$$\mathcal{T}_\epsilon(\mathcal{A}_{\mathrm{quan}}, \mathcal{F}) := \inf_{A\in\mathcal{A}_{\mathrm{quan}}}\sup_{f\in\mathcal{F}} T_\epsilon(A, f). \tag{8}$$

2.3. Quantum Lower Bound with Bounded Input Domain

We first prove a query complexity lower bound for any quantum algorithm A_{quan} defined in Section 2.2 on a function class with bounded input domain using the hard instance $\tilde{f}_{T;U}$ defined in Eq. (6). For the convenience of notations, we use $\widetilde{O}^p_{T;U}$ to denote the quantum evaluation oracle encoding the p-th-order derivatives of function $\tilde{f}_{T;U}$, or equivalently

$$\widetilde{O}^{(p)}_{T;U}|\mathbf{x}\rangle|y\rangle \to |\mathbf{x}\rangle\Big|y\oplus\nabla^{(0,\ldots,p)}\tilde{f}_{T;U}(\mathbf{x})\Big\rangle.$$

Consider the truncated sequence $A^{(k)}_{\mathrm{quan}}$ of any possible quantum algorithm A_{quan} with $k<T$, we define a sequence of unitaries starting with $A_0 = A^{(k)}_{\mathrm{quan}}$ as follows:

$$\begin{aligned}
A_0 &:= V_k\widetilde{O}^{(p)}_{T;U}V_{k-1}\widetilde{O}^{(p)}_{T;U}\cdots\widetilde{O}^{(p)}_{T;U}V_1\widetilde{O}^{(p)}_{T;U}V_0\\
A_1 &:= V_k\widetilde{O}^{(p)}_{T;U}V_{k-1}\widetilde{O}^{(p)}_{T;U}\cdots\widetilde{O}^{(p)}_{T;U}V_1\widetilde{O}^{(p)}_{1;U_1}V_0\\
A_2 &:= V_k\widetilde{O}^{(p)}_{T;U}V_{k-1}\widetilde{O}^{(p)}_{T;U}\cdots\widetilde{O}^{(p)}_{2;U_2}V_1\widetilde{O}^{(p)}_{1;U_1}V_0\\
&\vdots\\
A_k &:= V_k\widetilde{O}^{(p)}_{k;U_k}V_{k-1}\widetilde{O}^{(p)}_{k-1;U_{k-1}}\cdots\widetilde{O}^{(p)}_{2;U_2}V_1\widetilde{O}^{(p)}_{1;U_1}V_0,
\end{aligned} \tag{9}$$

where $\widetilde{O}^{(p)}_{t,U_t}$ stands for the evaluation oracle of function $\tilde{f}_{t;U_t}$ and its p-th-order derivatives defined as

$$\tilde{f}_{t;U_t}(\mathbf{x}) := \alpha\bar{f}_t(U_t^T\mathbf{x}/\beta) \tag{10}$$

Our goal to show that the algorithm A_0 does not solve our problem. To achieve that, we follow a similar approach shown in Garg et al. (2020) and develop a hybrid argument in which we first show that the outputs of the algorithm A_i and A_{i+1} are close, so does the outputs of A_0 and A_k. Then, we argue that the algorithm A_k cannot find an ϵ-stationary point with high probability since oracles in the algorithm are independent from $\mathbf{u}_T$. Hence, A_k cannot do better than random guessing a vector $\mathbf{u}_T$, which by Lemma B.2 in Appendix B fails with overwhelming probability.

Lemma 2.3 (A_t and A_{t-1} have similar outputs)**.** *For a hard instance $\tilde{f}_{T;U}(\mathbf{x})\colon\mathbb{R}^d\to\mathbb{R}$ defined on $\mathbb{B}(\mathbf{0}, 2\beta\sqrt{T})$ with $d\ge 200T\log T$, let A_t for $t\in[k-1]$ be the unitaries defined in (9). Then*

$$\mathbb{E}_U\big(\|A_t|\mathbf{0}\rangle - A_{t-1}|\mathbf{0}\rangle\|^2\big)\le 1/36T^4.$$

Proof. Since the series of unitaries in Eq. (9) was constructed by gradually changing the quantum evaluation oracle, the difference between consecutive terms can be expressed as

$$\begin{aligned}
&\|A_t|\mathbf{0}\rangle - A_{t-1}|\mathbf{0}\rangle\|\\
&\quad= \big\|(\widetilde{O}^{(p)}_{t;U_t} - \widetilde{O}^{(p)}_{T;U_T})V_{t-1}\widetilde{O}^{(p)}_{t-1;U_{t-1}}\cdots\widetilde{O}^{(p)}_{1;U_1}V_0|\mathbf{0}\rangle\big\|.
\end{aligned}$$

We will prove the claim for any fixed choice of vectors $\{\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(t-1)}\}$, which will imply the claim for any distribution over those vectors. After fixing these vectors, we can see that the quantum state

$$V_{t-1}\widetilde{O}^{(p)}_{t-1;U_{t-1}}\cdots\widetilde{O}^{(p)}_{1;U_1}V_0|\mathbf{0}\rangle$$

is fixed and we refer to it as $|\psi\rangle$. Thus our problem reduces to showing for all quantum states $|\psi\rangle$,

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)}\}}\big(\|(\widetilde{O}^{(p)}_{t;U_t} - \widetilde{O}^{(p)}_{T;U_T})|\psi\rangle\|^2\big)\le\frac{1}{36T^4}.$$

For any $|\psi\rangle$, it can be expressed as $|\psi\rangle = \sum_{\mathbf{x}}\alpha_{\mathbf{x}}|\mathbf{x}\rangle|\phi_{\mathbf{x}}\rangle$, where $\mathbf{x}$ is the query made to the oracle, and $\sum_{\mathbf{x}}|\alpha_{\mathbf{x}}|^2 = 1$, which leads to

$$\begin{aligned}
&\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)}\}}\Big(\Big\|\sum_{\mathbf{x}}\alpha_{\mathbf{x}}(\widetilde{O}^{(p)}_{t;U_t} - \widetilde{O}^{(p)}_{T;U_T})|\mathbf{x}\rangle|\phi_{\mathbf{x}}\rangle\Big\|^2\Big)\\
&\qquad\le\sum_{\mathbf{x}}|\alpha_{\mathbf{x}}|^2\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)}\}}\big(\|(\widetilde{O}^{(p)}_{t;U_t} - \widetilde{O}^{(p)}_{T;U_T})|\mathbf{x}\rangle\|^2\big).
\end{aligned}$$

Since $|\alpha_{\mathbf{x}}|^2$ defines a probability distribution over $\mathbf{x}$, we can again upper bound the right hand side for any $\mathbf{x}\in\mathbb{B}(\mathbf{0}, 2\beta\sqrt{T})$ instead. Since $\widetilde{O}^{(p)}_{t;U_t}$ and $\widetilde{O}^{(p)}_{T;U_T}$ behave identically for some inputs x, the only nonzero terms are those where the oracles respond differently, which can only happen if

$$\nabla^{(0,\ldots,p)}\tilde{f}_{T;U}(\mathbf{x})\ne\nabla^{(0,\ldots,p)}\tilde{f}_{t;U_t}(\mathbf{x}).$$

When the response is different, we can upper bound $\left\|\left(\widetilde{O}^{(p)}_{t;U_t} - \widetilde{O}^{(p)}_{T;U}\right)|\mathbf{x}\rangle\right\|^2$ by 4 using the triangle inequality. Thus for any $\mathbf{x} \in \mathbb{B}(\mathbf{0}, \sqrt{T})$, we have

$$\begin{aligned}
&\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)}\}}\left(\left\|\left(\widetilde{O}^{(p)}_{t;U_t} - \widetilde{O}^{(p)}_{T;U_T}\right)|\mathbf{x}\rangle\right\|^2\right) \\
&\quad\le 4 \Pr_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)}\}}\left(\nabla^{(0,\ldots,p)}\tilde{f}_{T;U}(\mathbf{x}) \neq \nabla^{(0,\ldots,p)}\tilde{f}_{t;U_t}(\mathbf{x})\right) \\
&\quad\le \frac{1}{36T^4},
\end{aligned}$$

where the last inequality follows from Lemma B.1. □

Based on Lemma 2.3, we can obtain the following result with its proof deferred to Appendix C.2. By combining Lemma 2.3 and the soft projection technique in Carmon et al. (2020a) that can remove the boundedness of the iterates, we obtain the following quantum lower bound.

Theorem 2.4 (Formal version of Theorem 1.1)**.** *There exist numerical constants $0 < c_0, c_1 < \infty$ such that the following lower bound holds. Let $p \ge 1$, $p \in \mathbb{N}$, and let Δ, L_p, and ϵ be positive. Then,*

$$\mathcal{T}_\epsilon\left(\mathcal{A}_{\text{quan}}, \mathcal{F}_p(\Delta, L_p)\right) \ge c_0 \Delta \left(\frac{L_p}{\ell_p}\right)^{1/p} \epsilon^{-\frac{1+p}{p}},$$

where $\ell_p \le e^{\frac{5}{2}p\log p + c_1 p}$, the complexity measure $\mathcal{T}_\epsilon$ is defined in Eq. (8), and the function class $\mathcal{F}_p(\Delta, L_p)$ is defined in Definition 2.1. The lower bound holds even if we restrict $\mathcal{F}_p(\Delta, L_p)$ to functions whose domain has dimension

$$\Omega\left(\frac{200 c_0 \Delta L_p^{1/p}}{\ell_p^{1/p}} \epsilon^{-\frac{1+p}{p}} \cdot \log\left(\frac{c_0 \Delta L_p^{1/p}}{\ell_p^{1/p}} \epsilon^{-\frac{1+p}{p}}\right)\right).$$

3. Quantum Lower Bounds with Access to Stochastic Gradients

Based on Carmon et al. (2020a), Arjevani et al. (2020; 2022) further investigated the classical query lower bound finding ϵ-stationary points given access to stochastic first-order or additionally higher-order derivatives.

3.1. Classical Hard Instance and Lower Bound

Adopt the notation in Arjevani et al. (2022), in this subsection we discuss lower bounds for quantum algorithms finding stationary points of functions in the set $\mathcal{F}(\Delta, L)$ such that for any $F\colon \mathbb{R}^d \to \mathbb{R}$ with $F \in \mathcal{F}$ we have

$$F(\mathbf{0}) - \inf_{\mathbf{x}} F(\mathbf{x}) \le \Delta,$$

and

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Arjevani et al. (2022) proved the classical lower bound by extending the zero-chain property introduced in Carmon et al. (2020a) to the stochastic setting.

Definition 3.1 (Probability-p zero chain)**.** A stochastic gradient function $\mathbf{g}(\mathbf{x}, \xi)$ is a probability-p zero-chain if

$$\begin{aligned}
&\Pr\left(\exists \mathbf{x} : \text{prog}_0(\mathbf{g}(\mathbf{x}, \xi)) = \text{prog}_{\frac{1}{4}}(\mathbf{x}) + 1\right) \le p \\
&\Pr\left(\exists \mathbf{x} : \text{prog}_0(\mathbf{g}(\mathbf{x}, \xi)) > \text{prog}_{\frac{1}{4}}(\mathbf{x}) + 1\right) = 0
\end{aligned}$$

are both satisfied, where

$$\text{prog}_\zeta(\mathbf{x}) := \max\{i \ge 0 \,|\, |x_i| \ge \zeta\}, \tag{11}$$

with $x_0 \equiv 0$ for notation consistency

Intuitively, for a probability-p zero-chain with dimension T, any zero-respecting stochastic algorithm takes $\Omega(1/p)$ queries to discover one new coordinate in expectation. Hence, the expected number of queries to discover all the coordinates is $\Omega(T/p)$.

Arjevani et al. (2022) also used the same underlying function $\bar{f}_T(\mathbf{x})$ defined in Eq. (5) with the following stochastic gradient function

$$[\mathbf{g}_T(\mathbf{x}, \xi)]_i := \nabla_i \bar{f}_T(\mathbf{x}) \left(1 + \mathbb{1}\{i > \text{prog}_{\frac{1}{4}}(\mathbf{x})\}\left(\frac{\xi}{p} - 1\right)\right), \tag{12}$$

where $\xi \sim \text{Bernoulli}(p)$ with $p = O(\epsilon^2)$ in Arjevani et al. (2022). Then, $\bar{f}_T$ together with $\mathbf{g}_T$ forms a probabilistic-$\frac{1}{4}$ zero chain. With a similar approach in Section 2, a lower bound of order $\Omega(1/\epsilon^4)$ for all stochastic first-order algorithms can thus be obtained.

In this work we show that, although the classical hard instance in Arjevani et al. (2022) can be solved via $\tilde{O}(1/\epsilon^3)$ quantum stochastic gradient queries,[1] there exist harder instances such that any quantum algorithm also needs to make at least $\Omega(1/\epsilon^4)$ to guarantee a high success probability in the worst case. If the stochastic gradients additionally satisfy the mean-squared smoothness condition in Assumption 1.3, we can show that any quantum algorithm needs to make at least $\Omega(1/\epsilon^3)$ to find an ϵ-approximate stationary point with high probability in the worst case. These two quantum lower bounds concerning stochastic gradients satisfying Assumption 1.3 or not match the corresponding classical algorithmic upper bounds and are thus tight.

3.2. Stochastic Quantum Query Model and Quantum Speedup on the Classical Hard Instance

We can extend our quantum query model introduced in Section 2.2 to the stochastic settings. For a d-dimensional,

[1] Throughout this paper, the $\tilde{O}$ and $\tilde{\Omega}$ notations omit polylogarithmic terms, i.e., $\tilde{O}(g) = O(g\,\text{poly}(\log g))$ and $\tilde{\Omega}(g) = \Omega(g\,\text{poly}(\log g))$.

L-smooth objective function f, we assume access to the quantum stochastic gradient oracle $O_f^{(p)}$ defined as follows:

Definition 3.2 (Quantum stochastic gradient oracle)**.** For any L-lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$, its quantum stochastic first-order oracle $O_{\mathbf{g}}$ consists of a distribution P_ξ and an unbiased mapping $O_{\mathbf{g}}$ satisfying

$$O_{\mathbf{g}} |\mathbf{x}\rangle |\xi\rangle |\mathbf{v}\rangle = |\mathbf{x}\rangle |\xi\rangle |\mathbf{g}(\mathbf{x},\xi) + \mathbf{v}\rangle ,$$

where the stochastic gradient $\mathbf{g}(\mathbf{x},\xi)$ satisfies both

$$\begin{aligned} &\mathbb{E}_{\xi\sim P_\xi}[\mathbf{g}(\mathbf{x},\xi)] = \nabla f(\mathbf{x}) \\ &\mathbb{E}_{\xi\sim P_\xi}[\|\mathbf{g}(\mathbf{x},\xi) - \nabla f(\mathbf{x})\|^2] \le \sigma^2 \end{aligned} \tag{13}$$

for some constant σ.

To measure the performance of a quantum algorithm A_{quan} on function f with queries to its stochastic gradient oracle $O_{\mathbf{g}}$, based on Definition 2.2 we define

$$T_\epsilon(A_{\mathrm{quan}}, f) := T_\epsilon\big(\{A^{(t)}_{\mathrm{quan}} |0\rangle\}, f\big)$$

as the complexity of A_{quan} on f. With this setup, we define the complexity of algorithm class $\mathcal{A}_{\mathrm{quan}}$ of all quantum algorithms in the form (7) on a function class $\mathcal{F}$ to be

$$\mathcal{T}^{\mathrm{stoc}}_\epsilon(\mathcal{A}_{\mathrm{quan}}, \mathcal{F}, \sigma) := \inf_{A\in\mathcal{A}_{\mathrm{quan}}} \sup_{f\in\mathcal{F}} \sup_{\mathbf{g}} T_\epsilon(A, f), \tag{14}$$

where the last supremum is over all possible stochastic gradient functions $\mathbf{g}$ of f satisfying the bounded-variance requirement (13) in Definition 3.2.

Based on this complexity measure, we show that quantum algorithm can find an ϵ-approximate stationary point of the classical hard instance based on $\bar{f}_T$ with fewer queries than the classical lower bound of order $\Omega(1/\epsilon^4)$ by approximating the exact gradient via Grover's algorithm. In particular, we prove the following result.

Lemma 3.3. *Consider the classical hard instance in Arjevani et al. (2022) obtained by projecting $\bar{f}_T$ defined in Eq. (5) into a d-dimensional space while adopting the stochastic gradient function $\mathbf{g}_T(\mathbf{x},\xi)$ defined in Eq. (12), there exists a quantum algorithm using $\tilde{O}(1/\epsilon^3)$ queries to the stochastic quantum gradient oracle $O_{\mathbf{g}}$ defined in Definition 3.2 that can find find a point $\mathbf{x}$ satisfying $\bar{f}_T(\mathbf{x}) = 0$ with probability at least $1/2$.*

The proof of Lemma 3.3 is deferred to Appendix D.2.

3.3. Quantum Lower Bound

In this subsection, we introduce a new hard instance where any quantum algorithm also have to make $\Omega\big(\epsilon^{-4}\big)$ queries to the quantum stochastic oracle defined in Definition 3.2 to find an ϵ-stationary point with high probability.

Before presenting the construction of the stochastic gradient function, we first review some existing results on quantum multivariate mean estimation.

Problem 3.4. Consider a matrix $M \in \mathbb{R}^{d\times 2T}$ for some $T \le d/4$, denote $\mathbf{m}^{(j)}$ as the i-th column of M. Suppose T columns of M forms a set of orthonormal vectors, while the other T columns are all zero. The goal is to get a good estimation of the direction of the vector

$$\mathbf{g} := \frac{1}{2T}\sum_{j=1}^{2T} \mathbf{m}^{(j)}.$$

Formally, we define

$$W_{\mathbf{g}}(\eta) = \left\{ |\tilde{\mathbf{g}}\rangle \;\middle|\; \frac{|\langle\tilde{\mathbf{g}},\mathbf{g}\rangle|}{\|\tilde{\mathbf{g}}\|\cdot\|\mathbf{g}\|} \ge 1-\eta, \quad \tilde{\mathbf{g}}\in\mathbb{R}^d \right\},$$

and define $\Pi_{\mathbf{g}}(\eta)$ to be the projection operator onto $W_{\mathbf{g}}(\eta)$. For some small η, the goal is to produce a quantum state $|\psi\rangle$ with large value of $\|\Pi_{\mathbf{g}}(\eta)\cdot|\psi\rangle\|$, given the following quantum oracle

$$O_M |\mathbf{x}\rangle |j\rangle |\mathbf{v}\rangle = |\mathbf{x}\rangle |j\rangle \left|\mathbf{m}^{(j)} + \mathbf{v}\right\rangle. \tag{15}$$

Lemma 3.5 (Cornelissen et al. 2022, Theorem 3.7)**.** *For any $n < T/2$ and any quantum algorithm that uses at most n queries to the quantum oracle O_M defined in (15) with output state $|\psi\rangle$, we have*

$$\mathbb{E}_M[\|\Pi_{\mathbf{g}}(\eta)\cdot|\psi\rangle\|^2] \le \exp(-\zeta T),$$

for some small constant ζ, and

$$\eta \ge \frac{3\sqrt{2}}{8\|\mathbf{g}\|}\sqrt{\frac{\mathbb{E}_j\big[\|\mathbf{m}^{(j)} - \mathbf{g}\|^2\big]}{T}} \ge \frac{3}{4}. \tag{16}$$

and the expectation is over all possible matrices M.

Drawing inspiration from the hard instance for quantum multivariate mean estimation introduced by Cornelissen et al. (2022), we design our stochastic gradient function accordingly. In a manner similar to the approach employed by Arjevani et al. (2022), we incorporate stochasticity to amplify the difficulty of achieving significant progress in individual coordinates through stochastic gradient information. Specifically, for the d-dimensional function $\tilde{f}T;U$, where $d \ge 4T$, we ensure that for any point $\mathbf{x}$ with gradient $\mathbf{g}(\mathbf{x})$, there exists a matrix $M_{\mathbf{x}} \in \mathbb{R}^{d\times 2T}$. This matrix possesses T columns consisting of zeros ($\mathbf{0}$), while the remaining T columns form a set of orthonormal vectors that satisfy

$$\nabla_{\mathrm{prog}_{\frac{\beta}{4}}(\mathbf{x})+1}\tilde{f}_{T;U}(\mathbf{x}) = \frac{1}{2T}\sum_j 2\gamma\sqrt{T}\cdot\mathbf{m}^{(j)}_{\mathbf{x}}, \tag{17}$$

where $\mathbf{m}_{\mathbf{x}}^{(j)}$ stands for the j-th column of $M_{\mathbf{x}}$ and

$$\gamma = \left\|\nabla_{\mathrm{prog}_{\frac{\beta}{4}}(\mathbf{x})+1}\tilde{f}_{T;U}(\mathbf{x})\right\| \leq 23$$

is the norm of the $(\mathrm{prog}_{\beta/4}(\mathbf{x})+1)$-th gradient component at certain points whose exact value is specified later.

Moreover, to guarantee that all the stochastic gradients at $\mathbf{x}$ can only reveal the $(\mathrm{prog}_{\beta/4}(\mathbf{x})+1)$-th coordinate $\mathbf{u}_{\mathrm{prog}_{\beta/4}(\mathbf{x})+1}$ even with infinite number of queries and will not "accidentally" make further progress, we additionally require that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ with $\mathrm{prog}_{\beta/4}(\mathbf{x}) \neq \mathrm{prog}_{\beta/4}(\mathbf{y})$, all the columns of $M_{\mathbf{x}}$ are orthogonal to all the columns of $M_{\mathbf{y}}$. This can be achieved by creating T orthogonal subspaces

$$\{\mathcal{V}_1, \ldots, \mathcal{V}_T\},$$

where each subspace is of dimension $2T$ and has no overlap with $\{\mathbf{u}_1, \ldots, \mathbf{u}_T\}$, such that the columns of $M_{\mathbf{x}}$ are within

$$\mathrm{span}\left\{\mathbf{u}_{\mathrm{prog}_{\frac{\beta}{4}}(\mathbf{x})+1}, \mathcal{V}_{\mathrm{prog}_{\frac{\beta}{4}}(\mathbf{x})+1}\right\},$$

as long as the dimension d is larger than $2T^2 + T = O(T^2)$.

Then, we can define the following stochastic gradient function for $\nabla\tilde{f}_{T;U}(\mathbf{x})$:

$$\mathbf{g}(\mathbf{x}, j) = \mathbf{g}(\mathbf{x}) - \mathbf{g}_{\mathrm{prog}_{\beta/4}(\mathbf{x})+1}(\mathbf{x}) + 2\gamma\sqrt{T} \cdot \mathbf{m}_{\mathbf{x}}^{(j)} \quad (18)$$

where j is uniformly distributed in the set $[2T]$. Then based on Lemma 3.5, we know that it is hard for quantum algorithms to get a accurate estimation of the direction of $\mathbf{g}_{\mathrm{prog}_{\beta/4}(\mathbf{x})+1}(\mathbf{x})$ given only access to stochastic gradients at point $\mathbf{x}$ defined in Eq. (18) using less than $T/2$ queries.

Further, we can show that if one only knows about the first t components $\{\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(t)}\}$, even if we permit the quantum algorithm to query the stochastic gradient oracle at different positions of $\mathbf{x}$, it is still hard to learn $\mathbf{u}^{(t+1)}$ as well as other components with larger indices. Quantitatively, for any $1 \leq t \leq T$ we define

$$W_{t;\perp} := \Big\{\mathbf{x} \in \mathbb{B}(\mathbf{0}, \beta\sqrt{T}) \,\Big|\, \exists i, \text{ s.t. } |\langle \mathbf{x}, \mathbf{u}^{(q)}\rangle| \geq \frac{\beta}{4} \text{ and } t < i \leq T\Big\}, \quad (19)$$

and

$$W_{i;\|} := \mathbb{B}(\mathbf{0}, \beta\sqrt{T}) - W_{i;\perp}.$$

Intuitively, $W_{t;\perp}$ is the subspace of $\mathbb{B}(\mathbf{0}, \beta\sqrt{T})$ such that any vector in $W_{t;\perp}$ has a relatively large overlap with at least one of $\mathbf{u}^{(t+1)}, \ldots, \mathbf{u}^{(T)}$. Moreover, we use $\Pi_{t;\perp}$ and $\Pi_{t;\|}$ to denote the quantum projection operators onto $W_{t;\perp}$ and $W_{t;\|}$, respectively. As shown in Lemma D.1, if starting in the subspace $W_{t;\|}$, any quantum algorithm using at most $T/2$ queries at arbitrary locations cannot output a quantum state that has a large overlap with $W_{t;\perp}$ in expectation. Then, we can obtain the following result.

Theorem 3.6 (Formal version of Theorem 1.2)**.** *For any Δ, L, σ, and ϵ that are all positive, we have*

$$\mathcal{T}_{\epsilon}^{\mathrm{stoc}}\big(\mathcal{A}_{\mathrm{quan}}, \mathcal{F}_1(\Delta, L), \sigma\big) \geq \Omega\Big(\frac{\min\{L^2\Delta^2, \sigma^4\}}{\epsilon^4}\Big),$$

where the complexity measure $\mathcal{T}_{\epsilon}^{\mathrm{stoc}}(\cdot)$ is defined in Eq. (14), and the function class $\mathcal{F}_1(\Delta, L)$ is defined in Definition 2.1. The lower bound still holds even if we restrict $\mathcal{F}_1(\Delta, L)$ to functions whose domain has dimension

$$O(\min\{L^2\Delta^2, \sigma^4\}/\epsilon^4).$$

Remark 3.7. Compared to the classical result (Arjevani et al., 2022), the lowest possible dimension of the hard instance is improved from $\Theta(\epsilon^{-6})$ to $\Theta(\epsilon^{-4})$, which is due to a sharper analysis and may be of independent interest.

The proof of Theorem 3.6 is deferred to Appendix D.3.2.

3.4. Quantum Lower Bound with the Mean-Squared Smoothness Assumption

We also prove a quantum query lower bound for finding an ϵ-stationary point with access to the quantum stochastic gradient oracle defined in Definition 3.2 and additionally satisfies the *mean-squared smoothness* assumption defined in Assumption 1.3 for some constant $\bar{L}$.

Theorem 3.8 (Formal version of Theorem 1.4)**.** *For any Δ, $\bar{L}$, σ, and ϵ that are all positive, we have*

$$\mathcal{T}_{\epsilon}^{\mathrm{stoc}}\big(\mathcal{A}_{\mathrm{quan}}, \mathcal{F}_1(\Delta, \bar{L}), \sigma\big) \geq \Omega(\Delta\bar{L}\sigma/\epsilon^3).$$

if we further assume the stochastic gradient function $\mathbf{g}(\mathbf{x})$ satisfies Assumption 1.3 with mean-squared smoothness parameter $\bar{L}$, where the complexity measure $\mathcal{T}_{\epsilon}^{\mathrm{stoc}}(\cdot)$ is defined in Eq. (14), the function class $\mathcal{F}_1(\Delta, \bar{L})$ is defined in Definition 2.1. The lower bound still holds even if we restrict $\mathcal{F}_1(\Delta, \bar{L})$ to functions whose domain has dimension

$$\tilde{O}(\Delta\bar{L}\sigma/\epsilon^3).$$

Remark 3.9. Compared to the classical result (Arjevani et al., 2022), the lowest possible dimension of the hard instance is improved from $\Theta(\epsilon^{-4})$ to $\Theta(\epsilon^{-3})$, which is due to a sharper analysis and may be of independent interest.

The proof of Theorem 3.8 is deferred to Appendix D.4.3.

Acknowledgements

We thank Hao Wang for helpful discussions regarding Section 3, and also thank anonymous reviewers for helpful suggestions on an initial version of this paper. TL was supported by a startup fund from Peking University.

References

Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., and Ma, T. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1195–1199, 2017. arXiv:1611.01146

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Sekhari, A., and Sridharan, K. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pp. 242–299. PMLR, 2020. arXiv:2006.13476

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pp. 1–50, 2022. arXiv:1912.02365

Birgin, E. G., Gardenghi, J. L., Martínez, J. M., Santos, S. A., and Toint, P. L. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1):359–368, 2017.

Bottou, L. and Bousquet, O. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pp. 161–168, 2007.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018. arXiv:1606.04838

Brandão, F. G. and Svore, K. Quantum speed-ups for semidefinite programming. In *Proceedings of the 58th Annual Symposium on Foundations of Computer Science*, pp. 415–426, 2017. arXiv:1609.05537

Brandão, F. G., Kalev, A., Li, T., Lin, C. Y.-Y., Svore, K. M., and Wu, X. Quantum SDP solvers: Large speed-ups, optimality, and applications to quantum learning. In *Proceedings of the 46th International Colloquium on Automata, Languages, and Programming*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 27:1–27:14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019. arXiv:1710.02581

Bubeck, S., Jiang, Q., Lee, Y.-T., Li, Y., and Sidford, A. Complexity of highly parallel non-smooth convex optimization. *Advances in neural information processing systems*, 32, 2019a. arXiv:1906.10655

Bubeck, S., Jiang, Q., Lee, Y. T., Li, Y., and Sidford, A. Near-optimal method for highly smooth convex optimization. In *Conference on Learning Theory*, pp. 492–507. PMLR, 2019b. arXiv:1812.08026

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. "Convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International conference on machine learning*, pp. 654–663. PMLR, 2017. arXiv:1705.02766

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018. arXiv:1611.00756

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1):71–120, 2020a. arXiv:1710.11606

Carmon, Y., Jambulapati, A., Jiang, Q., Jin, Y., Lee, Y. T., Sidford, A., and Tian, K. Acceleration with a ball optimization oracle. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19052–19063, 2020b. arXiv:2003.08078

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points II: First-order methods. *Mathematical Programming*, 185(1):315–355, 2021. arXiv:1711.00841

Cartis, C., Gould, N. I. M., and Toint, P. L. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.

Chakrabarti, S., Childs, A. M., Li, T., and Wu, X. Quantum algorithms and lower bounds for convex optimization. *Quantum*, 4:221, 2020. arXiv:1809.01731

Chakrabarti, S., Childs, A. M., Hung, S.-H., Li, T., Wang, C., and Wu, X. Quantum algorithm for estimating volumes of convex bodies. *ACM Transactions on Quantum Computing*, 4(3):1–60, 2023. arXiv:1908.03903

Childs, A. M., Kothari, R., and Somma, R. D. Quantum algorithm for systems of linear equations with exponentially improved dependence on precision. *SIAM Journal on Computing*, 46(6):1920–1950, 2017. arXiv:1511.02306

Childs, A. M., Leng, J., Li, T., Liu, J.-P., and Zhang, C. Quantum simulation of real-space dynamics. *Quantum*, 6:680, 2022. arXiv:2203.17006

Cornelissen, A., Hamoudi, Y., and Jerbi, S. Near-optimal quantum algorithms for multivariate mean estimation. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 33–43, 2022. arXiv:2111.09787

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pp. 689–699, 2018. arXiv:1807.01695

Garg, A., Kothari, R., Netrapalli, P., and Sherif, S. No quantum speedup over gradient descent for non-smooth convex optimization, 2020. arXiv:2010.01801

Garg, A., Kothari, R., Netrapalli, P., and Sherif, S. Near-optimal lower bounds for convex optimization for all orders of smoothness. *Advances in Neural Information Processing Systems*, 34:29874–29884, 2021. arXiv:2112.01118

Gasnikov, A., Dvurechensky, P., Gorbunov, E., Vorontsova, E., Selikhanovych, D., and Uribe, C. A. Optimal tensor methods in smooth convex and uniformly convex optimization. In *Conference on Learning Theory*, pp. 1374–1391. PMLR, 2019.

Gong, W., Zhang, C., and Li, T. Robustness of quantum algorithms for nonconvex optimization, 2022. arXiv:2212.02548

Grover, L. K. A fast quantum mechanical algorithm for database search. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, pp. 212–219. ACM, 1996. arXiv:quant-ph/9605043

Harrow, A. W., Hassidim, A., and Lloyd, S. Quantum algorithm for linear systems of equations. *Physical review letters*, 103(15):150502, 2009. arXiv:0811.3171

Jin, C., Netrapalli, P., and Jordan, M. I. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference on Learning Theory*, pp. 1042–1085, 2018. arXiv:1711.10456

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021. arXiv:1902.04811

Kerenidis, I. and Prakash, A. A quantum interior point method for LPs and SDPs. *ACM Transactions on Quantum Computing*, 1(1), 2020. ISSN 2643-6809. doi: 10.1145/3406306. URL https://doi.org/10.1145/3406306. arXiv:1808.09266

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. arXiv:1412.6980

Lei, L., Ju, C., Chen, J., and Jordan, M. I. Non-convex finite-sum optimization via SCSG methods. In *Advances in Neural Information Processing Systems*, volume 30, 2017. arXiv:1706.09156

Liu, M., Li, Z., Wang, X., Yi, J., and Yang, T. Adaptive negative curvature descent with applications in non-convex optimization. *Advances in Neural Information Processing Systems*, 31:4858–4867, 2018.

Liu, Y., Su, W. J., and Li, T. On quantum speedups for nonconvex optimization via quantum tunneling walks, 2022. arXiv:2209.14501

Murty, K. G. and Kabadi, S. N. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming: Series A and B*, 39(2):117–129, 1987.

Nemirovski, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization, 1983.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Nesterov, Y. and Polyak, B. T. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

van Apeldoorn, J. and Gilyén, A. Improvements in quantum SDP-solving with applications. In *Proceedings of the 46th International Colloquium on Automata, Languages, and Programming*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 99:1–99:15. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019. arXiv:1804.05058

van Apeldoorn, J., Gilyén, A., Gribling, S., and de Wolf, R. Convex optimization using quantum oracles. *Quantum*, 4:220, 2020a. arXiv:1809.00643

van Apeldoorn, J., Gilyén, A., Gribling, S., and de Wolf, R. Quantum SDP-solvers: Better upper and lower bounds. *Quantum*, 4:230, 2020b. arXiv:1705.01843

Woodworth, B. and Srebro, N. Lower bound for randomized first order convex optimization, 2017. arXiv:1709.03594

Zhang, C. and Li, T. Escape saddle points by a simple gradient-descent based algorithm. *Advances in Neural Information Processing Systems*, 34:8545–8556, 2021. arXiv:2111.14069

Zhang, C., Leng, J., and Li, T. Quantum algorithms for escaping from saddle points. *Quantum*, 5:529, 2021. arXiv:2007.10253

Zhou, D., Xu, P., and Gu, Q. Finding local minima via stochastic nested variance reduction, 2018. arXiv:1806.08782

A. Auxiliary Lemmas

We list the auxiliary lemmas used in our proofs here.

Lemma A.1 (Carmon et al. 2020a, Lemma 1)**.** *The functions $\bar{f}_T$, Ψ and Φ satisfy the following.*

1. *For all $x \le \frac{1}{2}$ and all $k \in \mathbb{N}$, $\Psi^{(k)}(x) = 0$.*
2. *For all $x \ge 1$ and $|y| < 1$, $\Psi(x)\Phi'(y) > 1$.*
3. *$\forall \mathbf{x} \in \mathbb{R}^T$,*
$$\sqrt{\sum_{i=t}^{T} x_i^2} \le \frac{1}{2} \Rightarrow \nabla^{(0,\dots,p)} \bar{f}_T(x_1, x_2, \dots, x_T) = \nabla^{(0,\dots,p)} \bar{f}_T(x_1, \dots, x_t, 0, \dots, 0).$$
4. *Both Ψ and Φ are infinitely differentiable, and for all $k \in \mathbb{N}$ we have*
$$\sup_{\mathbf{x}} |\Psi^{(k)}(\mathbf{x})| \le \exp\Big(\frac{5k}{2}\log(4k)\Big), \qquad \sup_{\mathbf{x}} |\Phi^{(k)}(\mathbf{x})| \le \exp\Big(\frac{3k}{2}\log\frac{3k}{2}\Big).$$
5. *The functions and derivatives Φ, Ψ, Φ' and Φ' are non-negative and bounded, with*
$$0 \le \Psi < e, \quad 0 \le \Psi' \le \sqrt{54/e}, \quad 0 < \Phi < \sqrt{2\pi e}, \quad 0 < \Phi' \le \sqrt{e}.$$

Lemma A.2 (Carmon et al. 2020a, Lemma 2)**.** *If $|x_i| < 1$ for any $i \le T$, then there exists $j \le i$ such that $|x_j| < 1$ and*
$$\|\nabla \bar{f}_T(\mathbf{x})\| \ge \Big|\frac{\partial}{\partial x_j} \bar{f}_T(\mathbf{x})\Big| > 1.$$

Lemma A.3 (Carmon et al. 2020a, Lemma 3)**.** *The function $\bar{f}_T \colon \mathbb{R}^d \to \mathbb{R}$ defined in Eq. (5) satisfies the following.*

1. *$\bar{f}_T(\mathbf{0}) - \inf_{\mathbf{x}} \bar{f}_T(\mathbf{x}) \le 12T$.*
2. *For all $\mathbf{x} \in \mathbb{R}^d$, we have $\|\nabla \bar{f}_T(\mathbf{x})\|_\infty \le 23$ and $\|\nabla \bar{f}_T(\mathbf{x})\| \le 23\sqrt{T}$.*
3. *For all $p \ge 1$, the p-th order derivatives of $\bar{f}_T$ are ℓ_p-Lipschitz continuous, where*
$$\ell_p \le \exp\Big(\frac{5}{2}p\log p + cp\Big)$$
for some numerical constant $c < \infty$.

Lemma A.4 (Garg et al. 2020, Proposition 14)**.** *Let $\mathbf{x} \in \mathbb{B}(\mathbf{0}, 1)$. Then for a d-dimensional random unit vector $\mathbf{u}$ and all $c > 0$,*
$$\Pr_{\mathbf{u}}(|\langle \mathbf{x}, \mathbf{u}\rangle| \ge c) \le 2e^{-dc^2/2}.$$

Lemma A.5 (Grover 1996, Grover's algorithm)**.** *For any function $\omega(\xi) \colon \Xi \to \mathbb{R}$ satisfying*
$$\Pr_{\xi \in \Xi}\{\omega(\xi) \ne 0\} = 1 - p$$
for some $p < 1$, with probability at least $1/2$ we can find a ξ satisfying $\omega(\xi) \ne 0$ using $O(1/\sqrt{p})$ queries to the following quantum oracle
$$U_\omega |\xi\rangle |y\rangle = |\xi\rangle |y + \omega(\xi)\rangle .$$

Lemma A.6 (Arjevani et al. 2022, Observation 1)**.** *The function Γ defined in Eq. (29) and Θ_i defined in Eq. (28) satisfies*[2]

1. *$\Gamma(t) = 0$ for all $t \in (-\infty, 1/(4\beta)]$.*
2. *$\Gamma(t) = 1$ for all $t \in [1/(2\beta), \infty)$.*
3. *$\Gamma \in C^\infty$, with $0 \le \Gamma'(t) \le 6/\beta$ and $|\Gamma''(t)| \le 128/\beta^2$ for all $t \in \mathbb{R}$.*
4. *Θ_i is $(36/\beta)$-Lipschitz.*

[2]The last entry can be found in the proof of Lemma 4 of Arjevani et al. (2022).

B. Probabilistic Facts about $\tilde{f}_{T;U}$

In this subsection, we discuss some probabilistic facts of the hard instance $\tilde{f}_{T;U}$ defined in Eq. (6) that are useful for the proof of quantum lower bounds.

Lemma B.1. *Let $1 \le t \le T$ be integers and $\{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(t-1)}\}$ be a set of orthonormal vectors. Let $\{\mathbf{u}^{(t)}, \dots, \mathbf{u}^{(T)}\}$ be chosen uniformly at random so that the set $\{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(T)}\}$ is orthonormal. Then*

$$\forall \mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T}), \quad \Pr_{\mathbf{u}^{(t)},\dots,\mathbf{u}^{(T)}} \left(\nabla^{(0,\dots,p)} \tilde{f}_{T;U}(\mathbf{x}) \neq \nabla^{(0,\dots,p)} \tilde{f}_{t;U_t}(\mathbf{x})\right) \le \frac{1}{144T^4},$$

where $U_t \in \mathbb{R}^{d\times t}$ is defined as the orthogonal matrix with columns $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(t)}$ and $\tilde{f}_{t;U_t}(\mathbf{x})$ is defined in (10), given that the dimension d of $\tilde{f}_{T;U}$ satisfies $d \ge 200T\log T$.

If d further satisfies $d \ge 400T\log T$, we have

$$\forall \mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T}), \quad \Pr_{\mathbf{u}^{(t)},\dots,\mathbf{u}^{(T)}} \left(\nabla^{(0,\dots,p)} \tilde{f}_{T;U}(\mathbf{x}) \neq \nabla^{(0,\dots,p)} \tilde{f}_{t;U_t}(\mathbf{x})\right) \le \frac{1}{144T^6}.$$

If d satisfies $d \ge 400\mathscr{T}T\log\mathscr{T}$ for some $\mathscr{T}$ satisfying $\mathscr{T} \ge T$, we have

$$\forall \mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T}), \quad \Pr_{\mathbf{u}^{(t)},\dots,\mathbf{u}^{(T)}} \left(\nabla^{(0,\dots,p)} \tilde{f}_{T;U}(\mathbf{x}) \neq \nabla^{(0,\dots,p)} \tilde{f}_{t;U_t}(\mathbf{x})\right) \le \frac{1}{144\mathscr{T}^2T^4}.$$

Proof. Without loss of generality, in this proof we set $\alpha = \beta = 1$. Use $\mathbf{x}_\perp$ to denote the projection of $\mathbf{x}$ to the span $\{\mathbf{u}^{(t)}, \dots, \mathbf{u}^{(T)}\}$. Intuitively, as long as each component of $\mathbf{x}_\perp$ has a small absolute value, the components $\{\mathbf{u}^{(t)}, \dots, \mathbf{u}^{(T)}\}$ will have no impact on the function value and any order of derivative. Quantitatively, by Lemma A.1 we can derive that

$$\Pr\left(\nabla^{(0,\dots,p)} \tilde{f}_{T;U}(\mathbf{x}) \neq \nabla^{(0,\dots,p)} \tilde{f}_{t;U_t}(\mathbf{x})\right) \le 1 - \Pr\left(|\langle \mathbf{u}^{(t)}, \mathbf{x}\rangle|, \dots, |\langle \mathbf{u}^{(T)}, \mathbf{x}\rangle| \le \frac{\beta}{2}\right).$$

Since $\{\mathbf{u}^{(t)}, \dots, \mathbf{u}^{(T)}\}$ are chosen uniformly at random in the $(d-t+1)$-dimensional orthogonal complement of span $\{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(t-1)}\}$, by Lemma A.4 we can further derive that

$$\begin{aligned}
1 - \Pr\left(|\langle \mathbf{u}^{(t)}, \mathbf{x}\rangle|, \dots, |\langle \mathbf{u}^{(T)}, \mathbf{x}\rangle| \le \frac{\beta}{2}\right) &\le 1 - \prod_{i=t}^{T} \left[1 - \Pr\left(|\langle \mathbf{u}^{(i)}, \mathbf{x}\rangle| \ge \beta/2\right)\right] \\
&\le \sum_{i=t}^{T} \Pr\left[|\langle \mathbf{u}^{(i)}, \mathbf{x}\rangle| \ge \beta/2\right] \\
&= 2Te^{-(d-T)\cdot(4\sqrt{T})^{-2}/2}.
\end{aligned}$$

We can then reach our desired conclusions considering different values of d. □

Lemma B.2 (Cannot guess stationary point)**.** *Let $k < T$ be a positive integer and $\{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(k)}\}$ be a set of orthonormal vectors. Let $\{\mathbf{u}^{(k+1)}, \dots, \mathbf{u}^{(T)}\}$ be chosen uniformly at random from $\operatorname{span}(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(k)})^\perp$ such that all columns of the matrix $U = \left[\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(T)}\right]$ forms a set of orthonormal vectors. Then,*

$$\forall \mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T}), \quad \Pr_{\{\mathbf{u}^{(k+1)},\dots,\mathbf{u}^{(T)}\}} \left[\|\nabla \tilde{f}_{T;U}(\mathbf{x})\| \le \alpha/\beta\right] \le \frac{1}{144T^4},$$

for the function $\tilde{f}_{T;U}(\mathbf{x})\colon \mathbb{R}^d \to \mathbb{R}$ defined in Eq. (6), given that the dimension d satisfies $d \ge 200T\log T$. If d further satisfies $d \ge 400T\log T$, we have

$$\forall \mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T}), \quad \Pr_{\mathbf{u}^{(t)},\dots,\mathbf{u}^{(T)}} \left(\nabla^{(0,\dots,p)} \tilde{f}_{T;U}(\mathbf{x}) \neq \nabla^{(0,\dots,p)} \tilde{f}_{t;U_t}(\mathbf{x})\right) \le \frac{1}{144T^6}.$$

Proof. For any $\mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T})$, by Lemma A.4 we can claim that the quantity

$$\Pr_{\mathbf{u}^{(T)}}[\left|\langle \mathbf{u}^{(T)}, \mathbf{x}\rangle\right| \geq \beta] \leq 2\exp\Big(-d(2\sqrt{T})^{-2}/2\Big)$$

is less than $\frac{1}{144T^4}$ if $d \geq 200T\log T$ and is further less than $\frac{1}{144T^6}$ if $d \geq 400T\log T$. Further by Lemma A.2, given the condition that $\left|\langle \mathbf{u}^{(T)}, \mathbf{x}\rangle\right| < \beta$, we have

$$\|\nabla \tilde{f}_{T;U}(\mathbf{x})\| \geq \frac{\alpha}{\beta}\Big|\frac{\partial}{\partial x_T}\bar{f}_T(U^T\mathbf{x}/\beta)\Big| > \frac{\alpha}{\beta}.$$

Hence, we can conclude that with probability at most $\frac{1}{144T^4}$ or $\frac{1}{144T^6}$ separately under the conditions $d \geq 200T\log T$ or $d \geq 400T\log T$, the following inequality is true:

$$\|\nabla \tilde{f}_{T;U}(\mathbf{x})\| \leq \alpha/\beta. \tag{20}$$

□

C. Proof of Quantum Lower Bound with Lipschitz p-th Order Derivatives

C.1. Lower Bound with Bounded Input Domain

In this subsection, we prove a query complexity lower bound for any quantum algorithm A_{quan} defined in Section 2.2 on a function class with bounded input domain using the hard instance $\tilde{f}_{T;U}$ defined in Eq. (6). Quantitatively, we define the function class $\tilde{\mathcal{F}}_p(\Delta, L_p, \mathcal{R})$ with bounded input domain as follows.

Definition C.1. Let $p \geq 1$, $\Delta > 0$ and $L_p > 0$. Then the set $\tilde{\mathcal{F}}_p(\Delta, L_p, \mathcal{R})$ denotes the union, over $d \in \mathbb{N}$, of the collection of C^∞ functions $f\colon \mathbb{B}(\mathbf{0}, \mathcal{R}) \to \mathbb{R}$ with L_p-Lipschitz p-th derivative and $f(\mathbf{0}) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq \Delta$.

For the convenience of notations, we use $\widetilde{O}^p_{T;U}$ to denote the quantum evaluation oracle encoding the p-th-order derivatives of function $\tilde{f}_{T;U}$, or equivalently

$$\widetilde{O}^{(p)}_{T;U}\left|\mathbf{x}\right\rangle\left|y\right\rangle \to \left|\mathbf{x}\right\rangle\left|y \oplus \nabla^{(0,\ldots,p)}\tilde{f}_{T;U}(\mathbf{x})\right\rangle. \tag{21}$$

Consider the truncated sequence $A^{(k)}_{\text{quan}}$ of any possible quantum algorithm A_{quan} with $k < T$, we define a sequence of unitaries starting with $A_0 = A^{(k)}_{\text{quan}}$ as follows:

$$\begin{aligned}
A_0 &:= V_k\widetilde{O}^{(p)}_{T;U}V_{k-1}\widetilde{O}^{(p)}_{T;U}\cdots\widetilde{O}^{(p)}_{T;U}V_1\widetilde{O}^{(p)}_{T;U}V_0\\
A_1 &:= V_k\widetilde{O}^{(p)}_{T;U}V_{k-1}\widetilde{O}^{(p)}_{T;U}\cdots\widetilde{O}^{(p)}_{T;U}V_1\widetilde{O}^{(p)}_{1;U_1}V_0\\
A_2 &:= V_k\widetilde{O}^{(p)}_{T;U}V_{k-1}\widetilde{O}^{(p)}_{T;U}\cdots\widetilde{O}^{(p)}_{2;U_2}V_1\widetilde{O}^{(p)}_{1;U_1}V_0\\
&\vdots\\
A_k &:= V_k\widetilde{O}^{(p)}_{k;U_k}V_{k-1}\widetilde{O}^{(p)}_{k-1;U_{k-1}}\cdots\widetilde{O}^{(p)}_{2;U_2}V_1\widetilde{O}^{(p)}_{1;U_1}V_0,
\end{aligned} \tag{22}$$

where $\widetilde{O}^{(p)}_{t,U_t}$ stands for the evaluation oracle of function $\tilde{f}_{t;U_t}$ and its p-th-order derivatives as defined in Eq. (10). Our goal to show that the algorithm A_0 does not solve our problem. To achieve that, we follow a similar approach shown in Garg et al. (2020) and develop a hybrid argument in which we first show that the outputs of the algorithm A_i and A_{i+1} are close, so does the outputs of A_0 and A_k. Then, we argue that the algorithm A_k cannot find an ϵ-stationary point with high probability since oracles in the algorithm are independent from $\mathbf{u}_T$. Hence, A_k cannot do better than random guessing a vector $\mathbf{u}_T$, which by Lemma B.2 in Appendix B fails with overwhelming probability.

Proposition C.2. *There exist numerical constants $0 < c_0, c_1 < \infty$ such that the following lower bound holds. Let $p \geq 1$, $p \in \mathbb{N}$, and let Δ, L_p, and ϵ be positive. Then,*

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\text{quan}}, \tilde{\mathcal{F}}_p(\Delta, L_p, \mathcal{R})\big) \geq c_0\Delta(L_p/\ell_p)^{1/p}\epsilon^{-\frac{1+p}{p}},$$

where $\ell_p \le e^{\frac{5}{2}p\log p + c_1 p}$, $\mathcal{R} = \sqrt{c_0\Delta}\big(\frac{\ell_p}{L_p}\big)^{\frac{1}{2p}}\epsilon^{-\frac{p-1}{2p}}$, *the complexity measure is defined in Eq. (8), and the function class* $\tilde{F}_p(\Delta, L_p, \mathcal{R})$ *is defined in Definition C.1. The lower bound holds even if we restrict* $\tilde{\mathcal{F}}_p(\Delta, L_p, \mathcal{R})$ *to functions whose domain has dimension*

$$\frac{200c_0\Delta L_p^{1/p}}{\ell_p^{1/p}}\epsilon^{-\frac{1+p}{p}} \cdot \log\Big(\frac{c_0\Delta L_p^{1/p}}{\ell_p^{1/p}} \cdot \epsilon^{-\frac{1+p}{p}}\Big).$$

The following lemma is helpful for proving Proposition C.2.

Lemma C.3. *Consider the* d*-dimensional function* $\tilde{f}_{T;U}(\mathbf{x})\colon \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T}) \to \mathbb{R}$ *defined in (6) with the rotation matrix* U *being chosen arbitrarily and the dimension* $d \ge 200T\log T$*. Consider the truncated sequence* $A^{(k)}_{\mathrm{quan}}$ *of any possible quantum algorithm* A_{quan} *containing* $k < T$ *queries to the oracle* $O_f^{(p)}$ *defined in Eq. (2), let* p_U *be the probability distribution over* $\mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T})$ *obtained by measuring the state* $A^{(t)}_{\mathrm{quan}}|0\rangle$*, which is related to the rotation matrix* U*. Then,*

$$\Pr_{U, \mathbf{x}_{out}\sim p_U}\big[\|\nabla\tilde{f}_{T;U}(\mathbf{x}_{out})\| \le \alpha/\beta\big] \le \frac{1}{3}.$$

Proof. We first demonstrate that A_k defined in Eq. (9) cannot find an α/β-approximate stationary point with high probability. In particular, let p_{U_k} be the probability distribution over $\mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T})$ obtained by measuring the output state $A_k|0\rangle$. Then,

$$\Pr_{U_k, \mathbf{x}_{\text{out}}\in p_{U_k}}\big[\|\nabla\tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \le \alpha/\beta\big] \le \Pr_{\{\mathbf{u}^{(k+1)},\ldots,\mathbf{u}^{(T)}\}}\big[\|\nabla\tilde{f}_{T;U}(\mathbf{x})\| \le \alpha/\beta\big].$$

for any fixed $\mathbf{x}$. Hence by Lemma B.2,

$$\Pr_{U_k, \mathbf{x}_{\text{out}}\in p_{U_k}}\big[\|\nabla\tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \le \alpha/\beta\big] \le \frac{1}{6}.$$

Moreover, by Lemma 2.3 and Cauchy-Schwarz inequality, we have

$$\mathbb{E}_U\big[\|A_k|0\rangle - A_0|0\rangle\|^2\big] \le k\cdot\mathbb{E}_U\Big[\sum_{t=1}^{k-1}\|A_{t+1}|0\rangle - A_t|0\rangle\|^2\Big] \le \frac{1}{36T^2}.$$

Then by Markov's inequality,

$$\Pr_U\Big[\|A_k|0\rangle - A_0|0\rangle\|^2 \ge \frac{1}{6T}\Big] \le \frac{1}{6T},$$

since both norms are at most 1. Thus, the total variance distance between the probability distribution p_U obtained by measuring $A_0|0\rangle$ and the probability distribution $p_U^{(k)}$ obtained by measuring $A_k|0\rangle$ is at most

$$\frac{1}{6T} + \frac{1}{6T} = \frac{1}{3T} \le \frac{1}{6}.$$

Hence, we can conclude that

$$\Pr_{U, \mathbf{x}_{\text{out}}\sim p_U}\big[\|\nabla\tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \le \alpha/\beta\big] \le \Pr_{U_k, \mathbf{x}_{\text{out}}\sim p_{U_k}}\big[\|\nabla\tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \le \alpha/\beta\big] + \frac{1}{6} = \frac{1}{3}.$$

□

Equipped with Lemma C.3, we are now ready to prove Proposition C.2.

Proof of Proposition C.2. We set up the scaling parameters α, β in hard instance $\tilde{f}_{T;U}\colon \mathbb{R}^d \to \mathbb{R}$ defined in Eq. (6) for some T as

$$\alpha = \frac{L_p \beta^{p+1}}{\ell_p}, \qquad \beta = \Big(\frac{\ell_p \epsilon}{L_p}\Big)^{1/p},$$

where $\ell_p \le e^{2.5p\log p + c_1}$ as in the third entry of Lemma A.3. Then by Lemma A.3, we know that the p-th order derivatives of $\tilde{f}_{T;U}$ are L_p-Lipschitz continuous. Moreover, note that

$$\tilde{f}_{T;U}(\mathbf{0}) - \inf_{\mathbf{x}} \tilde{f}_{T;U}(\mathbf{x}) = \alpha\big(\bar{f}_T(\mathbf{0}) - \inf_{\mathbf{x}} \bar{f}_T(\mathbf{x})\big) \le \frac{\ell_p^{1/p} \epsilon^{\frac{1+p}{p}}}{12 L_p^{1/p}} T.$$

Then by choosing

$$T = \frac{c_0 \Delta L_p^{1/p}}{\ell_p^{1/p}} \epsilon^{-\frac{1+p}{p}},$$

for some positive constant c_0, we can have $\tilde{f}_{T;U} \in \tilde{\mathcal{F}}(\Delta, L_p, \mathcal{R})$ for arbitrary dimension d and rotation matrix U. Moreover, by Lemma C.3, for any truncated sequence $A^{(t)}_{\mathrm{quan}}$ of any possible quantum algorithm A_{quan} containing $t < T$ queries to the oracle $O_f^{(p)}$ on input domain $\mathbb{B}(0, \mathcal{R})$, we have

$$\Pr_{U, \mathbf{x}_{\mathrm{out}} \sim p_U}\big[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\mathrm{out}})\| \le \alpha/\beta\big] = \Pr_{U, \mathbf{x}_{\mathrm{out}} \sim p_U}\big[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\mathrm{out}})\| \le \epsilon\big] \le \frac{1}{3},$$

where p_U is the probability distribution over $\mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T}) = \mathbb{B}(\mathbf{0}, \mathcal{R})$ obtained by measuring the state $A^{(t)}_{\mathrm{quan}}|0\rangle$, given that the dimension d satisfies

$$d \ge 200 T \log T = \frac{200 c_0 \Delta L_p^{1/p}}{\ell_p^{1/p}} \epsilon^{-\frac{1+p}{p}} \cdot \log\Big(\frac{c_0 \Delta L_p^{1/p}}{\ell_p^{1/p}} \epsilon^{-\frac{1+p}{p}}\Big).$$

Finally, due to Definition 2.2 we can conclude that

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathrm{quan}}, \tilde{\mathcal{F}}_p(\Delta, L_p, \mathcal{R})\big) \ge T = c_0 \Delta \Big(\frac{L_p}{\ell_p}\Big)^{1/p} \epsilon^{-\frac{1+p}{p}}.$$

□

C.2. Lower Bound with Unbounded Input Domain

In this subsection, we extend the lower bound in Proposition C.2 for bounded input domain to the setting of unbounded input domain and then present the proof of Theorem 2.4. In particular, we consider the hard instance $\hat{f}_{T;U}$ introduced by Carmon et al. (2020a),

$$\hat{f}_{T;U}(\mathbf{x}) := \tilde{f}_{T;U}(\chi(\mathbf{x})) + \frac{\alpha}{10} \cdot \frac{\|\mathbf{x}\|^2}{\beta^2},$$

where

$$\chi(\mathbf{x}) := \frac{\mathbf{x}}{\sqrt{1 + \|\mathbf{x}\|^2/\hat{\mathcal{R}}^2}},$$

with $\hat{\mathcal{R}} = 230\sqrt{T}$ and α, β, T defined in Eq. (6). The quadratic term $\|\mathbf{x}\|^2/10$ guarantees that with overwhelming probability one cannot obtain an ϵ-stationary point by randomly choosing an $\mathbf{x}$ with large norm. In all, the constants in $\hat{f}_{T;U}$ are chosen carefully such that stationary points of $\hat{f}_{T;U}$ are in one-to-one correspondence to stationary points of the hard instance $\tilde{f}_{T;U}$ concerning the setting with bounded input domain. Quantitatively,

Lemma C.4 (Carmon et al. 2020a, Section 5.2)**.** *Let Δ, L_p, and ϵ be positive constants. There exist numerical constants $0 < c_0, c_1 < \infty$ such that, under the following choice of parameters*

$$T = \frac{c_0 \Delta L_p^{1/p}}{\ell^{1/p}} \epsilon^{-\frac{1+p}{p}}, \quad \alpha = \frac{L_p \beta^{p+1}}{\ell_p},$$
$$\beta = \Big(\frac{\ell_p \epsilon}{L_p}\Big)^{1/p}, \quad \mathcal{R} = \sqrt{c_0 \Delta}\Big(\frac{\ell_p}{L_p}\Big)^{\frac{1}{2p}} \epsilon^{-\frac{p-1}{2p}},$$

where $\ell_p \le e^{2.5p \log p + c_1}$ as in the third entry of Lemma A.3, such that for any function pairs $(\tilde{f}_{T;U}, \hat{f}_{T;U}) \in \tilde{\mathcal{F}}_p(\Delta, L_p, \mathcal{R}) \times \mathcal{F}_p(\Delta, L_p)$ with dimension $d \ge 200T \log T$ and the same rotation matrix U, where the function classes are defined in Definition 2.1 and Definition C.1 separately, there exists a bijection between the ϵ-stationary points of $\tilde{f}_{T;U}$ and the ϵ-stationary points of $\hat{f}_{T;U}$ that is independent from U.

Equipped with Lemma C.4, we are now ready to prove Theorem 2.4.

Proof of Theorem 2.4. Note that one quantum query to the p-th order derivatives of $\hat{f}_{T;U}$ can be implemented by one quantum query to the p-th order derivatives of $\tilde{f}_{T;U}$ with the same rotation U. Combined with Lemma C.4, we can note that the problem of finding ϵ-stationary points of $\tilde{f}_{T;U}$ with unknown U can be reduced to the problem of finding ϵ-stationary points of $\hat{f}_{T;U}$ with no additional overhead in terms of query complexity. Then by Proposition C.2, we can conclude that

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\text{quan}}, \mathcal{F}_p(\Delta, L_p)\big) \ge \mathcal{T}_\epsilon\big(\mathcal{A}_{\text{quan}}, \tilde{\mathcal{F}}_p(\Delta, L_p, \mathcal{R})\big) = c_0 \Delta \Big(\frac{L_p}{\ell_p}\Big)^{1/p} \epsilon^{-\frac{1+p}{p}},$$

and the dimension dependence is the same as Proposition C.2. □

D. Proof of Quantum Lower Bounds with Access to Stochastic Gradients

D.1. Overview of the Proof Techniques

Similar to the classical stochastic lower bound result (Arjevani et al., 2022), we still utilize the classical hard instance defined in Carmon et al. (2020a) but with different scaling parameters. Nevertheless, the stochastic gradient function of the hard instance in Arjevani et al. (2022) exhibits a relatively simple form. In particular, after learning the first $(t-1)$ coordinate directions, the next query would be to reveal the t-th coordinate direction via the component $\nabla_t f(\mathbf{x})$, upon which direction Arjevani et al. (2022) applied all the stochasticity to obtain the following stochastic gradient function

$$[\mathbf{g}(\mathbf{x}, \xi)]_i := \nabla_i f(\mathbf{x}) \cdot \Big(1 + \mathbb{1}\{i = t\}\Big(\frac{\xi}{p} - 1\Big)\Big) \tag{23}$$

for some probability parameter $p = O(\epsilon^2)$. Intuitively, it takes $1/p$ classical queries in expectation to reveal the gradient component $\nabla_t f$ and obtain an accurate estimation of ∇f. For a quantum algorithm however, it takes only $O(1/\sqrt{p})$ queries by Grover's search algorithm (Grover, 1996), which leads to a quadratic quantum speedup, as shown in Appendix D.2.

To address this issue, inspired by the quantum lower bound on multivariate mean estimation (Cornelissen et al., 2022), we construct a new stochastic gradient function where each stochastic gradient $\mathbf{g}(\mathbf{x}, \xi)$ all has a very small overlap with $\nabla_t f$ and one has to take at least $\Omega(1/p)$ quantum queries to obtain enough knowledge of the stochastic gradients to estimate $\nabla_t f$ and ∇f accurately. Full details regarding the construction of $\mathbf{g}(\mathbf{x}, \xi)$ are presented in Section 3.3. Then, a quantum lower bound can be obtained matching the existing classical algorithmic upper bound result following the same procedure as Section 2.

Furthermore, if we assume that the stochastic gradient function satisfies the mean-squared smoothness condition described in Assumption 1.3, the stochastic gradient function defined in Eq. (23) is no longer applicable as it is not continuous on certain inputs. To address this issue, we apply a similar version of the function smoothing technique introduced in Arjevani et al. (2022) to our stochastic gradient function (23) to obtain a "smoothed" stochastic gradient function, as shown in Appendix D.4.1, upon which we can obtain a quantum query lower bound matching the existing classical algorithmic upper bound result given that the stochastic gradient function satisfies Assumption 1.3.

D.2. Quantum Speedup on the Classical Hard Instance

In this subsection, we present the proof of Lemma 3.3, which show that for minimizing the objective function in the hard instance for proving the classical lower bound in Arjevani et al. (2022), there exists a quantum algorithm based on Grover's search algorithm that can find its stationary point using only $\tilde{O}(1/\epsilon^3)$ queries to the stochastic quantum gradient oracle, even in the absence of the mean-squared smoothness assumption. Consequently, proving the $\Omega(1/\epsilon^4)$ quantum lower bound in Theorem 3.6 necessitates the development of fundamentally new ideas. Further details regarding these ideas will be presented in the next subsection, Appendix D.3.

Proof of Lemma 3.3. Note that at any $\mathbf{x} \in \mathbb{R}^d$ such that $\mathrm{prog}_{1/4}(\mathbf{x}) = T$, one can directly reveal the exact gradient using one query to the stochastic quantum gradient oracle $O_{\mathbf{g}}$. As for points $\mathbf{x}$ with $\mathrm{prog}_{1/4}(\mathbf{x}) = t < T$, by Lemma A.5 we know that, using

$$\log(1/\delta) \cdot O(1/\sqrt{p}) = O(\epsilon^{-1}\log(1/\delta))$$

queries to the oracle $O_{\mathbf{g}}$, we can obtain an accurate estimation of $\nabla_t f$ and hence ∇f, with success probability at least $1-\delta$. Then, we can perform gradient descent with step size being $1/L$ to reach a stationary point within $O(\epsilon^{-2})$ iterations, with overall success probability at least $1 - O(\delta/\epsilon^2)$. Hence, we can complete the proof by setting $\delta = O(\epsilon^2)$, and the overall query complexity equals

$$O(\epsilon^{-1}\log(1/\delta)) \cdot O(1/\epsilon^{-2}) = \tilde{O}(\epsilon^{-3}).$$

□

D.3. Proof of Quantum Lower Bound

We prove Theorem 3.6 in this subsection. To achieve this, we first prove that any quantum algorithm must make at least $\Omega(T)$ queries to the quantum stochastic gradient oracle to obtain an accurate estimation of the actual gradient (Lemma D.1), and based on which we establish the quantum lower bound with a bounded input domain in Appendix D.3.1. This can be further extended to the setting of unbounded input domains, consisting the proof of Theorem 3.6 in Appendix D.3.2.

Lemma D.1. *For any $n < T/2$ and $t \le T$, suppose in the form of Definition 3.2 we are given the quantum stochastic gradient oracle $\widetilde{O}_{\mathbf{g};U}$ of $\mathbf{g}(\mathbf{x}, j)$ defined in Eq. (18). Then for any quantum algorithm A_{quan} in the form of Eq. (7), consider the sequence of unitaries $A^{(n)}_{\mathrm{quan}}$ truncated after the n stochastic gradient oracle query*

$$A^{(n)}_{\mathrm{quan}} := \widetilde{O}_{\mathbf{g};U} V_n \widetilde{O}_{\mathbf{g};U} \cdots \widetilde{O}_{\mathbf{g};U} V_2 \widetilde{O}_{\mathbf{g};U} V_1,$$

and any input state $|\phi\rangle$, we have

$$\delta_\perp(n) := \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big[\|\Pi_{t;\perp} \cdot A^{(n)}_{\mathrm{quan}} |\phi\rangle\|^2\big] \le \frac{n}{18T^6}, \tag{24}$$

where the expectation is over all possible sets $\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)}\}$ and all possible set of matrices $\{M_{\mathbf{x}}\}$ at all positions $\mathbf{x} \in \mathbb{B}(\mathbf{0}, \beta\sqrt{T})$ satisfying Eq. (17), given that the dimension d of the objective function $\tilde{f}_{T;U}$ satisfies $d \ge 2T^2 + T$.

Proof. We use induction to prove this claim. Firstly for $n = 1$, we have

$$\begin{aligned}
\delta_\perp(1) &= \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big[\|\Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} V_0 |\phi\rangle\|^2\big] \\
&= \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big[\|\Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} |\phi\rangle\|^2\big] \\
&\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big[\big\|\Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} \big|\phi_{\|}\big\rangle\big\|^2\big] + \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big[\big\||\phi_\perp\rangle\big\|^2\big],
\end{aligned}$$

where $\big|\phi_{\|}\big\rangle := \Pi_{t;\|} |\phi\rangle$ and $|\phi_\perp\rangle := \Pi_{t;\perp} |\phi\rangle$. Since for all components in the (possibly superposition) state $\Pi_{T;\perp} |\psi\rangle$ all the stochastic gradients have no overlap with $\{\mathbf{u}^{t+2},\ldots,\mathbf{u}^T\}$, by Lemma 3.5, we have

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big[\big\|\Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} \big|\phi_{\|}\big\rangle\big\|^2\big] \le \exp(-\zeta T),$$

where ζ is a small enough constant. Moreover, by Lemma B.1 we have

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\left[\left\| |\phi_\perp\rangle \right\|^2\right] \le \frac{1}{36T^6}.$$

Hence,

$$\delta_\perp(1) \le \exp(-\zeta T) + \frac{1}{36T^6} \le \frac{1}{18T^6}.$$

Suppose the inequality (24) holds for all $n \le \tilde{n}$ for some $\tilde{n} < \frac{T}{2}$. Then for $n = \tilde{n}+1$, we denote

$$|\phi_{\tilde{n}}\rangle := \widetilde{O}_{\mathbf{g};U} V_{\tilde{n}-1} \widetilde{O}_{\mathbf{g};U} \cdots \widetilde{O}_{\mathbf{g};U} V_1 \widetilde{O}_{\mathbf{g};U} V_0 |\phi\rangle .$$

Then,

$$\begin{aligned}
\delta_\perp(\tilde{n}+1) &= \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\left[\|\Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} V_{\tilde{n}} |\phi_{\tilde{n}}\rangle \|^2\right] \\
&\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\left[\left\|\Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} V_{\tilde{n}} \left|\phi_{\tilde{n};\|}\right\rangle \right\|^2\right] + \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\left[\left\| |\phi_{\tilde{n};\perp}\rangle \right\|^2\right] \\
&\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\left[\left\|\Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} V_{\tilde{n}} \left|\phi_{\tilde{n};\|}\right\rangle \right\|^2\right] + \delta_\perp(\tilde{n}).
\end{aligned}$$

Consider the following sequence

$$\widetilde{O}_{\mathbf{g};U} V_{\tilde{n}} \widetilde{O}_{\mathbf{g};U} \cdots \widetilde{O}_{\mathbf{g};U} V_0 |\phi'\rangle = \widetilde{O}_{\mathbf{g};U} V_{\tilde{n}} \left|\phi_{\tilde{n};\|}\right\rangle ,$$

note that it contains $\tilde{n}+1 \le \frac{T}{2}$ queries to the stochastic gradient oracle, and at each query except the last one, the input state has no overlap with the desired space $W_{t;\perp}$. Then by Lemma 3.5, we have

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\left[\left\|\Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} V_{\tilde{n}} \left|\phi_{\tilde{n};\|}\right\rangle \right\|^2\right] \le \exp(-\zeta T) + \frac{1}{36T^6} \le \frac{1}{18T^6}.$$

Hence, the inequality (24) also holds for $n = \tilde{n}+1$. □

D.3.1. Lower Bound with Bounded Input Domain

Through this construction of quantum stochastic gradient oracle, we can prove the query complexity lower bound for any quantum algorithm A_{quan} defined in Section 2.2 using the hard instance $\tilde{f}_{T;U}$ defined in Eq. (6). For the convenience of notations, we use $\widetilde{O}_{\mathbf{g};U}$ to denote the stochastic gradient oracle defined in Eq. (18) of function $\tilde{f}_{T;U}$. Similar to Section 2.3, we consider the truncated sequence $A_{\text{quan}}^{(K\cdot T/2)}$ of any possible quantum algorithm A_{quan} with $K < T$, and define a sequence of unitaries starting with $A_0 = A_{\text{quan}}^{(K\cdot T/2)}$ as follows:

$$\begin{aligned}
A_0 &:= V_{K+1} \widetilde{O}_{\mathbf{g};U} V_{K;T/2} \cdots \widetilde{O}_{\mathbf{g};U} V_{K;1} \cdots \widetilde{O}_{\mathbf{g};U} V_{2;T/2} \cdots \widetilde{O}_{\mathbf{g};U} V_{2;1} \widetilde{O}_{\mathbf{g};U} V_{1;T/2} \cdots \widetilde{O}_{\mathbf{g};U} V_{1;1} \\
A_1 &:= V_{K+1} \widetilde{O}_{\mathbf{g};U} V_{K;T/2} \cdots \widetilde{O}_{\mathbf{g};U} V_{K;1} \cdots \widetilde{O}_{\mathbf{g};U} V_{2;T/2} \cdots \widetilde{O}_{\mathbf{g};U} V_{2;1} \widetilde{O}_{\mathbf{g};U_1} V_{1;T/2} \cdots \widetilde{O}_{\mathbf{g};U_1} V_{1;1} \\
A_2 &:= V_{K+1} \widetilde{O}_{\mathbf{g};U} V_{K;T/2} \cdots \widetilde{O}_{\mathbf{g};U} V_{K;1} \cdots \widetilde{O}_{\mathbf{g};U_2} V_{2;T/2} \cdots \widetilde{O}_{\mathbf{g};U_2} V_{2;1} \widetilde{O}_{\mathbf{g};U_1} V_{1;T/2} \cdots \widetilde{O}_{\mathbf{g};U_1} V_{1;1} \\
&\vdots \\
A_K &:= V_{K+1} \widetilde{O}_{\mathbf{g};U_K} V_{K;T/2} \cdots \widetilde{O}_{\mathbf{g};U_K} V_{K;1} \cdots \widetilde{O}_{\mathbf{g};U_2} V_{2;T/2} \cdots \widetilde{O}_{\mathbf{g};U_2} V_{2;1} \widetilde{O}_{\mathbf{g};U_1} V_{1;T/2} \cdots \widetilde{O}_{\mathbf{g};U_1} V_{1;1},
\end{aligned} \tag{25}$$

where $\widetilde{O}_{\mathbf{g};U_t}$ stands for the stochastic gradient oracle of function $\tilde{f}_{t;U_t}$. Note that for the sequence of unitaries A_0, it can be decomposed into the product of V_{K+1} and K sequences of unitaries, each of the form

$$\mathscr{A}_k(n) = \widetilde{O}_{\mathbf{g};U} V_{k;n} \widetilde{O}_{\mathbf{g};U} \cdots \widetilde{O}_{\mathbf{g};U} V_{k;2} \widetilde{O}_{\mathbf{g};U} V_{k;1}$$

for $n = T/2$ and $k \in [K]$ for some unitaries $V_1, \ldots, V_n$. In the following lemma we demonstrate that, for such sequence $\mathscr{A}_k(n)$, if we replace $\widetilde{O}_{\mathbf{g};U}$ by another oracle that only reveals part information of f, the sequence will barely change on random inputs.

Lemma D.2. *For any $t \in [T-1]$ and any $n \le \frac{T}{2}$, consider the following two sequences of unitaries*

$$\mathscr{A}(n) = \widetilde{O}_{\mathbf{g};U} V_n \widetilde{O}_{\mathbf{g};U} \cdots \widetilde{O}_{\mathbf{g};U} V_2 \widetilde{O}_{\mathbf{g};U} V_1,$$

and

$$\hat{\mathscr{A}_t}(n) = \widetilde{O}_{\mathbf{g};U_t} V_n \widetilde{O}_{\mathbf{g};U_t} \cdots \widetilde{O}_{\mathbf{g};U_t} V_2 \widetilde{O}_{\mathbf{g};U_t} V_1,$$

we have

$$\delta(n) := \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\| (\hat{\mathscr{A}_t}(n) - \mathscr{A}(n)) |\psi\rangle \|^2 \right] \le \frac{n}{36T^5}, \tag{26}$$

for any fixed pure state $|\psi\rangle$.

Proof. We use induction to prove this claim. Firstly for $n = 1$, we have

$$\begin{aligned} &\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\left\| (\hat{\mathscr{A}_t}(n) - \mathscr{A}(n)) |\psi\rangle \right\|^2 \right] \\ &\quad = \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\left\| (\widetilde{O}_{\mathbf{g};U} - \widetilde{O}_{\mathbf{g};U_t}) |\psi\rangle \right\|^2 \right] \le \frac{1}{36T^6}, \end{aligned} \tag{27}$$

where the last inequality follows from Lemma B.1. Suppose the inequality (27) holds for all $n \le \tilde{n}$ for some $\tilde{n} < \frac{T}{2}$. Then for $n = \tilde{n} + 1$, we have

$$\begin{aligned} \delta(\tilde{n}+1) &= \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\| (\hat{\mathscr{A}_t}(n) - \mathscr{A}(n)) |\psi\rangle \|^2 \right] \\ &\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\| (\widetilde{O}_{\mathbf{g};U} - \widetilde{O}_{\mathbf{g};U_t}) |\psi_t\rangle \|^2 \right] + \delta(\tilde{n}), \end{aligned}$$

where

$$|\psi_t\rangle = V_{\tilde{n}} \widetilde{O}_{\mathbf{g};U_t} \cdots \widetilde{O}_{\mathbf{g};U_t} V_1 |\psi\rangle$$

is a function of U_t obtained by $\tilde{n}$ queries to $\tilde{O}_{\mathbf{g};U_t}$. By Lemma D.1, we have

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_x\}} [\| \Pi_{t;\perp} |\psi_t\rangle \|^2] \le \frac{n}{18T^6} \le \frac{1}{36T^5},$$

indicating that $|\psi_t\rangle$ only has a very little overlap with the subspace $W_{t;\perp}$ defined in (19), outside of which the columns $\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)}\}$ of U has no impact on the function value and derivatives of $\tilde{f}_{T;U}$. Thus,

$$\begin{aligned} \delta(\tilde{n}+1) &\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\| (\widetilde{O}_{\mathbf{g};U} - \widetilde{O}_{\mathbf{g};U_t}) |\psi_t\rangle \| \right] + \delta(\tilde{n}) \\ &\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_x\}} [\| \Pi_{t;\perp} |\psi_t\rangle \|^2] + \delta(\tilde{n}) \le \frac{\tilde{n}+1}{36T^5}, \end{aligned}$$

indicating that Eq. (26) also holds for $n = \tilde{n} + 1$. □

Lemma D.3 (A_t and A_{t-1} have similar outputs)**.** *For a hard instance $\tilde{f}_{T;U}(\mathbf{x})\colon \mathbb{R}^d \to \mathbb{R}$ defined on $\mathbb{B}(\mathbf{0}, 2\beta\sqrt{T})$ with $d \ge 2T^2 + T$, let A_t for $t \in [K]$ be the sequence unitaries defined in Eq. (25). Then*

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left(\| A_t |\mathbf{0}\rangle - A_{t-1} |\mathbf{0}\rangle \|^2 \right) \le \frac{1}{72T^4}.$$

Proof. From the definition of the unitaries in Eq. (25), we have

$$\| A_t |\mathbf{0}\rangle - A_{t-1} |\mathbf{0}\rangle \| = \| (\mathscr{A}(T/2) - \hat{\mathscr{A}_t}(T/2)) |\psi\rangle \|,$$

for some fixed quantum state $|\psi\rangle$ dependent on the vectors $\{\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(t-1)}\}$, where

$$\mathscr{A}(T/2) = \widetilde{O}_{\mathbf{g};U} V_{T/2} \widetilde{O}_{\mathbf{g};U} \cdots \widetilde{O}_{\mathbf{g};U} V_2 \widetilde{O}_{\mathbf{g};U} V_1,$$

and

$$\hat{\mathscr{A}_t}(T/2) = \widetilde{O}_{\mathbf{g};U_t} V_{T/2} \widetilde{O}_{\mathbf{g};U_t} \cdots \widetilde{O}_{\mathbf{g};U_t} V_2 \widetilde{O}_{\mathbf{g};U_t} V_1.$$

By Lemma D.2, we have

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left(\| (\mathscr{A}(T/2) - \hat{\mathscr{A}_t}(T/2)) \, |\psi\rangle \, \|^2 \right) \leq \frac{1}{36T^5} \cdot \frac{T}{2} = \frac{1}{72T^4},$$

which leads to

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left(\| A_t \, |\mathbf{0}\rangle - A_{t-1} \, |\mathbf{0}\rangle \, \|^2 \right) \leq \frac{1}{72T^4}.$$

□

Proposition D.4. *Consider the d-dimensional function $\tilde{f}_{T;U}(\mathbf{x})\colon \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T}) \to \mathbb{R}$ defined in (6) with the rotation matrix U being chosen arbitrarily and the dimension $d \geq 2T^2 + T$. Consider the truncated sequence $A_{\mathrm{quan}}^{(K\cdot T/2)}$ of any possible quantum algorithm A_{quan} containing $KT/2$ queries to the quantum stochastic gradient oracle $\widetilde{O}_{\mathbf{g};U}$ of $\mathbf{g}(\mathbf{x}, j)$ defined in Eq. (18) with $K < T$, let p_U be the probability distribution over $\mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T})$ obtained by measuring the state $A_{\mathrm{quan}}^{(K\cdot T/2)} \, |0\rangle$, which is related to the rotation matrix U. Then,*

$$\Pr_{U,\{M_{\mathbf{x}}\},\mathbf{x}_{out}\sim p_U} \left[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{out})\| \leq \alpha/\beta \right] \leq \frac{1}{3},$$

where the probability is subject to all possible orthogonal rotation matrices U, and all possible matrices $\{M_{\mathbf{x}}\}$ in the quantum stochastic gradient function $\mathbf{g}(\mathbf{x}, j)$ for any $\mathbf{x}$.

Proof. We first demonstrate that the sequence of unitaries A_K defined in Eq. (25) cannot find an α/β-approximate stationary point with high probability. In particular, let p_{U_K} be the probability distribution over $\mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T})$ obtained by measuring the output state $A_K \, |0\rangle$. Then,

$$\Pr_{U_K,\{M_{\mathbf{x}}\},\mathbf{x}_{\text{out}}\in p_{U_K}} \left[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \leq \alpha/\beta \right] \leq \Pr_{\{\mathbf{u}^{(k+1)},\ldots,\mathbf{u}^{(T)}\}} \left[\|\nabla \tilde{f}_{T;U}(\mathbf{x})\| \leq \alpha/\beta \right].$$

for any fixed $\mathbf{x}$. Then by Lemma B.2 we have

$$\Pr_{U_K,M_{\mathbf{x}},\mathbf{x}_{\text{out}}\in p_{U_K}} \left[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \leq \alpha/\beta \right] \leq \frac{1}{6}.$$

Moreover, by Lemma D.3 and Cauchy-Schwarz inequality, we have

$$\mathbb{E}_U \left[\| A_K \, |0\rangle - A_0 \, |0\rangle \, \|^2 \right] \leq K \cdot \mathbb{E}_U \Big[\sum_{t=1}^{K-1} \| A_{t+1} \, |0\rangle - A_t \, |0\rangle \, \|^2 \Big] \leq \frac{1}{72T^2}.$$

Then by Markov's inequality,

$$\Pr_U \left[\| A_{K-1} \, |0\rangle - A_0 \, |0\rangle \, \|^2 \geq \frac{1}{6T} \right] \leq \frac{1}{6T},$$

since both norms are at most 1. Thus, the total variance distance between the probability distribution p_U obtained by measuring $A_0 \, |0\rangle$ and the probability distribution p_{U_K} obtained by measuring $A_K \, |0\rangle$ is at most

$$\frac{1}{6T} + \frac{1}{6T} = \frac{1}{3T} \leq \frac{1}{6}.$$

Hence, we can conclude that

$$\begin{aligned} &\Pr_{U,\{M_{\mathbf{x}}\},\mathbf{x}_{\text{out}}\sim p_U^{(t)}} \left[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \leq \alpha/\beta \right] \\ &\leq \Pr_{U_K,\{M_{\mathbf{x}}\},\mathbf{x}_{\text{out}}\sim p_{U_K}} \left[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \leq \alpha/\beta \right] + \frac{1}{6} = \frac{1}{3}. \end{aligned}$$

□

Proposition D.5. *Suppose Δ, L, σ, and ϵ are positive. Then,*

$$\mathcal{T}_\epsilon^{\mathrm{stoc}}\big(\mathcal{A}_{\mathrm{quan}}, \tilde{\mathcal{F}}_1(\Delta, L, \mathcal{R}), \sigma\big) = \Omega\Big(\frac{\min\{\Delta^2 L^2, \sigma^4\}}{\epsilon^4}\Big),$$

where $\mathcal{R} = c \cdot \min\{\sqrt{L\Delta}, \sigma\}$ for some constant c, the complexity measure $\mathcal{T}_\epsilon^{\mathrm{stoc}}(\cdot)$ is defined in Eq. (14), and the function class $\tilde{\mathcal{F}}_1(\Delta, L, \mathcal{R})$ is defined in Definition C.1. The lower bound holds even if we restrict $\tilde{\mathcal{F}}_1(\Delta, L, \mathcal{R})$ to functions whose domain has dimension

$$\Theta\Big(\frac{\min\{L^2\Delta^2, \sigma^4\}}{\epsilon^4}\Big).$$

Proof. We set up the scaling parameters α and β in the hard instance $\tilde{f}_{T;U} : \mathbb{R}^d \to \mathbb{R}$ defined in Eq. (6) as

$$\alpha = \frac{L\beta^2}{\ell}, \qquad \beta = \frac{2\ell\epsilon}{L},$$

where ℓ is the gradient Lipschitz constant of $\bar{f}_T$ whose value is given in Lemma A.3. We also set the parameter

$$T = \min\Big\{\frac{\Delta\ell}{12L\beta^2}, \frac{\sigma^2\beta^2}{4\gamma^2\alpha^2}\Big\} = \Theta\Big(\frac{\min\{L\Delta, \sigma^2\}}{\epsilon^2}\Big).$$

Then by Lemma A.3, we know that $\tilde{f}_{T;U}$ is L-smooth, and

$$\tilde{f}_{T;U}(\mathbf{0}) - \inf_{\mathbf{x}} \tilde{f}_{T;U}(\mathbf{x}) = \alpha\big(\bar{f}_T(\mathbf{0}) - \inf_{\mathbf{x}} \bar{f}_T(\mathbf{x})\big) \le \frac{12L\beta^2}{\ell} \cdot T \le \Delta,$$

indicating that $\tilde{f}_{T;U} \in \tilde{\mathcal{F}}(\Delta, L_p, \mathcal{R})$ for arbitrary dimension d and rotation matrix U. Moreover, at every $\mathbf{x}$, we have

$$\mathbb{E}\big[\|\mathbf{g}(\mathbf{x}, j) - \nabla\tilde{f}_{T;U}(\mathbf{x})\|^2\big] \le 4\alpha^2\gamma^2 T/\beta^2 \le \delta^2,$$

indicating that the variance of the stochastic gradient function $\mathbf{g}$ defined in Eq. (18) is bounded by σ^2. Further, we notice that the radius

$$\mathcal{R} = 2\beta\sqrt{T} = c \cdot \min\big\{\sqrt{L\Delta}, \sigma\big\}$$

for some constant c.

By Proposition D.4, for any truncated sequence $A_{\mathrm{quan}}^{(KT/2)}$ of any possible quantum algorithm A_{quan} containing $KT/2 < T^2/2$ queries to the oracle $O_f^{(p)}$ on input domain $\mathbb{B}(0, \mathcal{R})$, we have

$$\Pr_{U, \mathbf{x}_{\mathrm{out}} \sim p_U}\big[\|\nabla\tilde{f}_{T;U}(\mathbf{x}_{\mathrm{out}})\| \le \alpha/\beta\big] = \Pr_{U, \mathbf{x}_{\mathrm{out}} \sim p_U}\big[\|\nabla\tilde{f}_{T;U}(\mathbf{x}_{\mathrm{out}})\| \le \epsilon\big] \le \frac{1}{3},$$

where p_U is the probability distribution over $\mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T}) = \mathbb{B}(\mathbf{0}, \mathcal{R})$ obtained by measuring the state $A_{\mathrm{quan}}^{(KT/2)}|0\rangle$, given that the dimension d satisfies

$$d \ge 2T^2 + T = \Theta\Big(\frac{\min\{L^2\Delta^2, \sigma^4\}}{\epsilon^4}\Big).$$

Then according to Definition 2.2, we can conclude that

$$\mathcal{T}_\epsilon^{\mathrm{stoc}}\big(\mathcal{A}_{\mathrm{quan}}, \tilde{\mathcal{F}}_1(\Delta, L, \mathcal{R}), \sigma\big) \ge \frac{T^2}{2} = \Omega\Big(\frac{\min\{L^2\Delta^2, \sigma^4\}}{\epsilon^4}\Big).$$

□

D.3.2. LOWER BOUND WITH UNBOUNDED INPUT DOMAIN

In this subsection, we extend the quantum lower bound proved in Proposition D.5 to the function class $\mathcal{F}_1(\Delta, L)$ with unbounded input domain via similar scaling techniques adopted in Arjevani et al. (2022) and Appendix C.2. In particular, we consider the scaled hard instance $\hat{f}_{T;U}$ introduced in Carmon et al. (2020a) and also used in Arjevani et al. (2022),

$$\hat{f}_{T;U}(\mathbf{x}) := \tilde{f}_{T;U}(\chi(\mathbf{x})) + \frac{\alpha}{10} \cdot \frac{\|\mathbf{x}\|^2}{\beta^2},$$

where

$$\chi(\mathbf{x}) := \frac{\mathbf{x}}{\sqrt{1 + \|\mathbf{x}\|^2/\hat{\mathcal{R}}^2}},$$

with parameters

$$\alpha = \frac{L\beta^2}{\ell}, \qquad \beta = \frac{2\ell\epsilon}{L}, \qquad \beta = \frac{2\ell\epsilon}{L}, \qquad T = \min\Big\{\frac{\Delta\ell}{12L\beta^2}, \frac{\sigma^2\beta^2}{4\gamma^2\alpha^2}\Big\}, \qquad \hat{\mathcal{R}} = 230\sqrt{T},$$

whose values are also adopted in the proof of Proposition D.5. The constants in $\hat{f}_{T;U}$ are chosen carefully such that stationary points of $\hat{f}_{T;U}$ are in one-to-one correspondence to stationary points of the hard instance $\tilde{f}_{T;U}$ concerning the setting with bounded input domain. Quantitatively,

Lemma D.6 (Arjevani et al. 2022)**.** *Let Δ, L_p, and ϵ be positive constants. There exist numerical constants $0 < c_0, c_1 < \infty$ such that, under the following choice of parameters*

$$T = \min\Big\{\frac{\Delta\ell}{12L\beta^2}, \frac{\sigma^2\beta^2}{4\gamma^2\alpha^2}\Big\}, \qquad \alpha = \frac{L\beta^2}{\ell}, \qquad \beta = \frac{2\ell\epsilon}{L}, \qquad \mathcal{R} = c\sqrt{T} \cdot \min\{\sqrt{L\Delta}, \sigma\},$$

where ℓ is the gradient Lipschitz parameter of $\bar{f}_T$ whose value is given in Lemma A.3, such that for any function pairs $(\tilde{f}_{T;U}, \hat{f}_{T;U}) \in \tilde{\mathcal{F}}_1(\Delta, L, \mathcal{R}) \times \mathcal{F}_1(\Delta, L)$ with dimension $d \geq 400T\log T$ and the same rotation matrix U, where the function classes are defined in Definition 2.1 and Definition C.1 separately, there exists a bijection between the ϵ-stationary points of $\tilde{f}_{T;U}$ and the ϵ-stationary points of $\hat{f}_{T;U}$ that is independent from U.

Equipped with Lemma D.6, we are now ready to prove Theorem 3.6.

Proof of Theorem 3.6. Note that one quantum query to the stochastic gradient of $\hat{f}_{T;U}$ can be implemented by one quantum query to the stochastic gradient of $\tilde{f}_{T;U}$ with the same rotation U, if we directly scale the stochastic gradient function of $\tilde{f}_{T;U}$ to $\hat{f}_{T;U}$, which will not increase the variance of the stochastic gradient function. Combined with Lemma D.6, we can note that the problem of finding ϵ-stationary points of $\tilde{f}_{T;U}$ with unknown U can be reduced to the problem of finding ϵ-stationary points of $\hat{f}_{T;U}$ with no additional overhead in terms of query complexity. Then by Proposition D.5, we can conclude that

$$\mathcal{T}_\epsilon^{\mathrm{stoc}}\big(\mathcal{A}_{\mathrm{quan}}, \mathcal{F}_1(\Delta, L), \sigma\big) \geq \mathcal{T}_\epsilon^{\mathrm{stoc}}\big(\mathcal{A}_{\mathrm{quan}}, \tilde{\mathcal{F}}_1(\Delta, L, \mathcal{R}), \sigma\big) = \Omega\Big(\frac{\max\{L^2\Delta^2, \sigma^4\}}{\epsilon^4}\Big),$$

and the dimension dependence is the same as Proposition D.5. □

D.4. Proof of Quantum Lower Bound with the Mean-Squared Smoothness Assumption

In this subsection, we prove a quantum query lower bound for finding an ϵ-stationary point with access to the quantum stochastic gradient oracle defined in Definition 3.2 and additionally satisfies the *mean-squared smoothness* assumption defined in Assumption 1.3 for some constant $\bar{L}$.

D.4.1. CONSTRUCTION OF THE STOCHASTIC GRADIENT FUNCTION SATISFYING ASSUMPTION 1.3

Note that the stochastic gradient function (18) in Section 3.3 does not satisfy Assumption 1.3 since the function $\mathrm{prog}_\alpha(\cdot)\colon \mathbb{R}^d \to \mathbb{R}$ defined in (11) contains a maximization over all the d components, which makes the stochastic gradient discontinuous.

This issue can be addressed using a smoothing technique similar to which introduced in Arjevani et al. (2022). In particular, Arjevani et al. (2022) defines the following smoothed version of the indicator function $\mathbb{I}\{i > \mathrm{prog}_{\frac{\beta}{4}}(\mathbf{x})\}$ for any i (with rotation U):

$$\Theta_i(\mathbf{x}) \coloneqq \Gamma\bigg(1 - \Big(\sum_{k=i}^{T} \Gamma^2(|x_k/\beta|)\Big)^{1/2}\bigg) = \Gamma(1 - \|\Gamma(\mathbf{x}_{\geq i})\|), \tag{28}$$

where $\Gamma(|\mathbf{x}_{\geq i}|)$ is a shorthand for a vector with entries

$$\Gamma(|x_i|), \Gamma(|x_{i+1}|), \ldots, \Gamma(|x_T|),$$

and the function $\Gamma\colon \mathbb{R} \to \mathbb{R}$ is defined as

$$\Gamma(t) = \frac{\int_{1/4}^{t/\beta} \Lambda(\tau)\mathrm{d}\tau}{\int_{1/4}^{1/2} \Lambda(\tau)\mathrm{d}\tau}, \quad \text{where} \quad \Lambda(t) = \begin{cases} 0, & \frac{t}{\beta} \leq \frac{1}{4} \text{ or } \frac{t}{\beta} \geq \frac{1}{2}, \\ \exp\Big(-\frac{1}{100\left(\frac{t}{\beta} - \frac{1}{4}\right)\left(\frac{1}{2} - \frac{t}{\beta}\right)}\Big), & \frac{1}{4} < \frac{t}{\beta} < \frac{1}{2}. \end{cases} \tag{29}$$

Note that Γ is a smooth non-decreasing Lipschitz function with $\Gamma(t) = 0$ for all $t \leq \beta/4$ and $\Gamma(t) = 1$ for all $t \geq \beta/2$. Then, the function $\Theta_i(\mathbf{x})$ defined in Eq. (28) satisfies

$$\mathbb{I}\left\{i > \mathrm{prog}_{\frac{\beta}{4}}(\mathbf{x})\right\} \leq \Theta_i(\mathbf{x}) \leq \mathbb{I}\left\{i > \mathrm{prog}_{\frac{\beta}{2}}(\mathbf{x})\right\}.$$

Following the same intuition of the gradient function defined in Eq. (18) without the mean-squared smoothness assumption, here we also arrange the stochasticity to harden the attempts on increasing the coordinate progress via stochastic gradient information. In particular, similar to Eq. (17), for the d-dimensional function $\tilde{f}_{T;U}$ with $d \geq 4\mathscr{T}$ for some integer $\mathscr{T}$ whose value is specified later, we note that for any point $\mathbf{x}$ with gradient $\mathbf{g}(\mathbf{x})$ there exists a matrix $M_{\mathbf{x}} \in \mathbb{R}^{d \times 2\mathscr{T}}$ with $\mathscr{T}$ columns being $\mathbf{0}$ and the other $\mathscr{T}$ columns forming a set of orthonormal vectors such that

$$\nabla_{\mathrm{prog}_{\frac{\beta}{2}}(\mathbf{x})+1} \tilde{f}_{T;U}(\mathbf{x}) = \frac{1}{2\mathscr{T}} \sum_{j} 2\gamma\sqrt{\mathscr{T}} \cdot \mathbf{m}_{\mathbf{x}}^{(j)}, \tag{30}$$

where $\mathbf{m}_{\mathbf{x}}^{(j)}$ stands for the j-th column of $M_{\mathbf{x}}$ and

$$\gamma = \left\|\nabla_{\mathrm{prog}_{\frac{\beta}{2}}(\mathbf{x})+1} \tilde{f}_{T;U}(\mathbf{x})\right\| \leq 23$$

is the norm of the $(\mathrm{prog}_{\beta/2}+1)$-th gradient component at certain points whose exact value is specified later.

Moreover, to guarantee that all the stochastic gradients at $\mathbf{x}$ can only reveal the $(\mathrm{prog}_{\beta/4}(\mathbf{x})+1)$-th coordinate direction $\mathbf{u}_{\mathrm{prog}_{\beta/4}(\mathbf{x})+1}$ even with infinite number of queries and will not "accidentally" make further progress, we additionally require that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ with $\mathrm{prog}_{\beta/4}(\mathbf{x}) \neq \mathrm{prog}_{\beta/4}(\mathbf{y})$, all the columns of $M_{\mathbf{x}}$ are orthogonal to all the columns of $M_{\mathbf{y}}$. This can be achieved by creating T orthogonal subspaces

$$\{\mathcal{V}_1, \ldots, \mathcal{V}_T\},$$

where each subspace is of dimension $2T$ and has no overlap with $\{\mathbf{u}_1, \ldots, \mathbf{u}_T\}$, such that for any $\mathbf{x}$ the columns of $M_{\mathbf{x}}$ are within the subspace

$$\mathrm{span}\left\{\mathbf{u}_{\mathrm{prog}_{\frac{\beta}{4}}(\mathbf{x})+1}, \mathcal{V}_{\mathrm{prog}_{\frac{\beta}{4}}(\mathbf{x})+1}\right\},$$

as long as the dimension d is larger than $2\mathscr{T}T + T = O(\mathscr{T}T)$.

Now, we can define the following stochastic gradient function for $\nabla \tilde{f}_{T;U}(\mathbf{x})$:

$$\hat{\mathbf{g}}(\mathbf{x}, j) = \mathbf{g}(\mathbf{x}) + \Theta_{\mathrm{prog}_{\beta/2}(\mathbf{x})+1}(\mathbf{x}) \cdot \left(2\gamma\sqrt{\mathscr{T}} \cdot \mathbf{m}^{(j)} - \mathbf{g}_{\mathrm{prog}_{\beta/2}(\mathbf{x})+1}(\mathbf{x})\right), \tag{31}$$

where j is uniformly distributed in the set $[2\mathscr{T}]$. Then, we can prove that this stochastic gradient function satisfies Assumption 1.3.

Lemma D.7. *The stochastic gradient function $\hat{g}$ defined in (31) is unbiased for $\nabla \tilde{f}_{T;U}(\mathbf{x})$ and satisfies*

$$\mathbb{E}\big\|\hat{\mathbf{g}}(\mathbf{x},j)-\nabla\tilde{f}_{T;U}(\mathbf{x})\big\|^2\le\frac{4\mathscr{T}\alpha^2}{\beta^2},\qquad \mathbb{E}\big\|\hat{\mathbf{g}}(\mathbf{x},j)-\hat{\mathbf{g}}(\mathbf{y},j)\big\|^2\le\frac{\hat{\ell}^2\mathscr{T}\alpha^2\|\mathbf{x}-\mathbf{y}\|^2}{\beta^2},$$

for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, where $\hat{\ell}=328$.

Proof. For any $\mathbf{x}$ and j, we define

$$\delta(\mathbf{x},j)\coloneqq\hat{\mathbf{g}}(\mathbf{x},j)-\nabla\tilde{f}_{T;U}(\mathbf{x})=\Theta_{\mathrm{prog}_{\beta/2}(\mathbf{x})+1}(\mathbf{x})\cdot\big(\mathbf{g}_{\mathrm{prog}_{\beta/2}(\mathbf{x})+1}(\mathbf{x})-2\gamma\sqrt{\mathscr{T}}\cdot\mathbf{m}^{(j)}\big).$$

Then we have

$$\mathbb{E}\big\|\hat{\mathbf{g}}(\mathbf{x},j)-\nabla\tilde{f}_{T;U}(\mathbf{x})\big\|^2=\mathbb{E}\|\delta(\mathbf{x},\mathbf{z})\|^2\le4\mathscr{T}\alpha^2|\Theta_{\mathrm{prog}_{\beta/2}(\mathbf{x})+1}(\mathbf{x})|/\beta^2\le4\mathscr{T}\alpha^2/\beta^2.$$

For any $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$,

$$\hat{\mathbf{g}}(\mathbf{x},j)-\hat{\mathbf{g}}(\mathbf{y},j)=\delta(\mathbf{x},j)-\delta(\mathbf{y},j)+\nabla\tilde{f}_{T;U}(\mathbf{x})-\nabla\tilde{f}_{T;U}(\mathbf{y}).$$

Since $\mathbb{E}[\delta(\mathbf{x},j)-\delta(\mathbf{y},j)]=0$, we can derive that

$$\mathbb{E}\big\|\hat{\mathbf{g}}(\mathbf{x},j)-\hat{\mathbf{g}}(\mathbf{y},j)\big\|^2=\mathbb{E}\big\|\delta(\mathbf{x},j)-\delta(\mathbf{y},j)\big\|^2+\|\nabla\tilde{f}_{T;U}(\mathbf{x})-\nabla\tilde{f}_{T;U}(\mathbf{y})\|^2.$$

Note that

$$\delta(\mathbf{x},j)-\delta(\mathbf{y},j)=\Theta_{i_\mathbf{x}}(\mathbf{x})\cdot\big(\mathbf{g}_{i_\mathbf{x}}(\mathbf{x})-2\gamma_\mathbf{x}\sqrt{\mathscr{T}}\cdot\mathbf{m}_\mathbf{x}^{(j)}\big)-\Theta_{i_\mathbf{y}}(\mathbf{y})\cdot\big(\mathbf{g}_{i_\mathbf{y}}(\mathbf{y})-2\gamma_\mathbf{y}\sqrt{\mathscr{T}}\cdot\mathbf{m}_\mathbf{y}^{(j)}\big),$$

where we denote $i_\mathbf{x}=\mathrm{prog}_{\beta/2}(\mathbf{x})+1$ and $i_\mathbf{y}=\mathrm{prog}_{\beta/2}(\mathbf{y})+1$. Then,

$$\begin{aligned}\mathbb{E}\big\|\delta(\mathbf{x},j)-\delta(\mathbf{y},j)\big\|^2\le&\,2\mathscr{T}(\nabla_{i_\mathbf{x}}\tilde{f}_{T;U}(\mathbf{x}))^2(\Theta_{i_\mathbf{x}}(\mathbf{x})-\Theta_{i_\mathbf{x}}(\mathbf{y}))^2\\&+2\mathscr{T}\big(\nabla_{i_\mathbf{x}}\tilde{f}_{T;U}(\mathbf{x})-\nabla_{i_\mathbf{x}}\tilde{f}_{T;U}(\mathbf{y})\big)^2\Theta_{i_\mathbf{x}}^2(\mathbf{y})\\&+2\mathscr{T}(\nabla_{i_\mathbf{y}}\tilde{f}_{T;U}(\mathbf{y}))^2(\Theta_{i_\mathbf{y}}(\mathbf{y})-\Theta_{i_\mathbf{y}}(\mathbf{x}))^2\\&+2\mathscr{T}\big(\nabla_{i_\mathbf{y}}\tilde{f}_{T;U}(\mathbf{y})-\nabla_{i_\mathbf{y}}\tilde{f}_{T;U}(\mathbf{x})\big)^2\Theta_{i_\mathbf{y}}^2(\mathbf{x}).\end{aligned}$$

As Θ_i is 36-Lipschitz for any $i\in[T]$ according to Lemma A.6, we have

$$\begin{aligned}\mathbb{E}\big\|\delta(\mathbf{x},j)-\delta(\mathbf{y},j)\big\|^2&\le\mathscr{T}\cdot\big(2\alpha^2\cdot(23\cdot6)^2\|\mathbf{x}-\mathbf{y}\|^2/\beta^2+2\|\nabla\tilde{f}_{T;U}(\mathbf{x})-\nabla\tilde{f}_{T;U}(\mathbf{y})\|^2\big)\\&\quad+\|\nabla\tilde{f}_{T;U}(\mathbf{x})-\nabla\tilde{f}_{T;U}(\mathbf{y})\|^2\\&\le\mathscr{T}\hat{\ell}^2\alpha^2\|\mathbf{x}-\mathbf{y}\|^2/\beta^2,\end{aligned}$$

where the last inequality uses the fact that the gradient of $\nabla\tilde{f}_{T;U}$ is $152\alpha/\beta$-Lipschitz continuous, which is demonstrated in Lemma A.3. □

Similar to the case of Section 3.3, we can show that if one only knows about the first t components $\{\mathbf{u}^{(1)},\dots,\mathbf{u}^{(t)}\}$, even if we permit the quantum algorithm to query the stochastic gradient oracle at different positions of $\mathbf{x}$, it is still hard to learn $\mathbf{u}^{(t+1)}$ as well as other components with larger indices. Quantitatively, following the same notation in Section 3.3, for any $1\le t\le T$ we denote

$$W_{t;\perp}\coloneqq\Big\{\mathbf{x}\in\mathbb{B}(\mathbf{0},\beta\sqrt{T})\,\big|\,\exists i,\ \text{s.t. }|\langle\mathbf{x},\mathbf{u}^{(q)}\rangle|\ge\frac{\beta}{4}\text{ and }t<i\le T\Big\},$$

and

$$W_{i;\|}\coloneqq\mathbb{B}(\mathbf{0},\beta\sqrt{T})-W_{i;\perp},$$

where $W_{t;\perp}$ is the subspace of $\mathbb{B}(\mathbf{0},\beta\sqrt{T})$ such that any vector in $W_{t;\perp}$ has a relatively large overlap with at least one of $\mathbf{u}^{(t+1)},\dots,\mathbf{u}^{(T)}$. Moreover, we still use $\Pi_{t;\perp}$ and $\Pi_{t;\|}$ to denote the quantum projection operators onto $W_{t;\perp}$ and $W_{t;\|}$, respectively. The following lemma demonstrates that, if starting in the subspace $W_{t;\|}$, any quantum algorithm using at most $\mathscr{T}/2$ queries at arbitrary locations cannot output a quantum state that has a large overlap with $W_{t;\perp}$ in expectation.

Lemma D.8. *For any $n < \mathscr{T}/2$ and $t \le T$, suppose in the form of Definition 3.2 we are given the quantum stochastic gradient oracle $\widetilde{O}_{\mathbf{g};U}$ of $\mathbf{g}(\mathbf{x}, j)$ defined in Eq. (31). Then for any quantum algorithm A_{quan} in the form of Eq. (7), consider the sequence of unitaries $A^{(n)}_{\mathrm{quan}}$ truncated after the n stochastic gradient oracle query*

$$A^{(n)}_{\mathrm{quan}} := \widetilde{O}_{\mathbf{g};U} V_n \widetilde{O}_{\mathbf{g};U} \cdots \widetilde{O}_{\mathbf{g};U} V_2 \widetilde{O}_{\mathbf{g};U} V_1,$$

and any input state $|\phi\rangle$, we have

$$\delta_\perp(n) := \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\| \Pi_{t;\perp} \cdot A^{(n)}_{\mathrm{quan}} |\phi\rangle \|^2 \right] \le \frac{n}{18\mathscr{T}^2 T^4}, \tag{32}$$

where the expectation is over all possible sets $\{\mathbf{u}^{(t)}, \ldots, \mathbf{u}^{(T)}\}$ and all possible sets of matrices $\{M_{\mathbf{x}}\}$ at all positions $\mathbf{x} \in \mathbb{B}(\mathbf{0}, \beta\sqrt{T})$ satisfy Eq. (30), given that the dimension d of the objective function $\tilde{f}_{T;U}$ satisfies $d \ge 2\mathscr{T} T \log \mathscr{T}$ and $\mathscr{T} \ge T$.

Proof. We use induction to prove this claim. First, for $n = 1$, we have

$$\begin{aligned}
\delta_\perp(1) &= \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\| \Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} V_0 |\phi\rangle \|^2 \right] \\
&= \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\| \Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} |\phi\rangle \|^2 \right] \\
&\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\left\| \Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} \left|\phi_{\|}\right\rangle \right\|^2 \right] + \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\left\| |\phi_\perp\rangle \right\|^2 \right],
\end{aligned}$$

where $\left|\phi_{\|}\right\rangle := \Pi_{t;\|} |\phi\rangle$ and $|\phi_\perp\rangle := \Pi_{t;\perp} |\phi\rangle$. Since for all components in the (possibly superposition) state $\Pi_{T;\perp} |\psi\rangle$ all the stochastic gradients have no overlap with $\{\mathbf{u}^{t+2}, \ldots, \mathbf{u}^T\}$, by Lemma 3.5 we have

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\left\| \Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} \left|\phi_{\|}\right\rangle \right\|^2 \right] \le \exp(-\zeta\mathscr{T}),$$

where ζ is a small enough constant. Moreover, by Lemma B.1 we have

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\left\| |\phi_\perp\rangle \right\|^2 \right] \le \frac{1}{36\mathscr{T}^2 T^4}.$$

Hence,

$$\delta_\perp(1) \le \exp(-\zeta\mathscr{T}) + \frac{1}{36\mathscr{T}^2 T^4} \le \frac{1}{18\mathscr{T}^2 T^4}.$$

Suppose the inequality (32) holds for all $n \le \tilde{n}$ for some $\tilde{n} < \frac{\mathscr{T}}{2}$. Then for $n = \tilde{n} + 1$, we denote

$$|\phi_{\tilde{n}}\rangle := \widetilde{O}_{\mathbf{g};U} V_{\tilde{n}-1} \widetilde{O}_{\mathbf{g};U} \cdots \widetilde{O}_{\mathbf{g};U} V_1 \widetilde{O}_{\mathbf{g};U} V_0 |\phi\rangle .$$

Then,

$$\begin{aligned}
\delta_\perp(\tilde{n}+1) &= \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\| \Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} V_{\tilde{n}} |\phi_{\tilde{n}}\rangle \|^2 \right] \\
&\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\left\| \Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} V_{\tilde{n}} \left|\phi_{\tilde{n};\|}\right\rangle \right\|^2 \right] + \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\left\| |\phi_{\tilde{n};\perp}\rangle \right\|^2 \right] \\
&\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}} \left[\left\| \Pi_{t;\perp} \cdot \widetilde{O}_{\mathbf{g};U} V_{\tilde{n}} \left|\phi_{\tilde{n};\|}\right\rangle \right\|^2 \right] + \delta_\perp(\tilde{n}).
\end{aligned}$$

Consider the following sequence

$$\widetilde{O}_{\mathbf{g};U} V_{\tilde{n}} \widetilde{O}_{\mathbf{g};U} \cdots \widetilde{O}_{\mathbf{g};U} V_0 |\phi'\rangle = \widetilde{O}_{\mathbf{g};U} V_{\tilde{n}} \left|\phi_{\tilde{n};\|}\right\rangle ,$$

note that it contains $\tilde{n} + 1 \le \frac{\mathscr{T}}{2}$ queries to the stochastic gradient oracle, and at each query except the last one, the input state has no overlap with the desired space $W_{t;\perp}$. Observe that within this restricted input subspace where these queries happen, we always have

$$\mathbb{I}\{i > \mathrm{prog}_{\frac{\beta}{4}}(\mathbf{x})\} = \Theta_i(\mathbf{x}) = \mathbb{I}\{i > \mathrm{prog}_{\frac{\beta}{2}}(\mathbf{x})\}.$$

Hence, the oracle behaves as if there is no scaling to the indicator function $\mathbb{I}\{i > \mathrm{prog}_{\frac{\beta}{4}}(\mathbf{x})\}$, and we can apply Lemma 3.5 to obtain the following result:

$$\begin{aligned}\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\left[\left\|\Pi_{t;\perp}\cdot\widetilde{O}_{\mathbf{g};U}V_{\tilde{n}}\left|\phi_{\tilde{n};\|}\right\rangle\right\|^2\right] &\le \exp(-\zeta\mathscr{T})+\frac{1}{36\mathscr{T}^2T^4}\\ &\le \exp(-\zeta T)+\frac{1}{36\mathscr{T}^2T^4}\\ &\le \frac{1}{18\mathscr{T}^2T^4}.\end{aligned}$$

Hence, the inequality (32) also holds for $n=\tilde{n}+1$. □

D.4.2. Lower Bound with Bounded Input Domain

Through this construction of quantum stochastic gradient oracle with mean-squared smoothness, we can prove the query complexity lower bound for any quantum algorithm A_{quan} defined in Section 2.2 using the hard instance $\tilde{f}_{T;U}$ defined in Eq. (6). For the convenience of notations, we use $\widetilde{O}_{\mathbf{g};U}$ to denote the stochastic gradient oracle defined in Eq. (31) of function $\tilde{f}_{T;U}$. Similar to Section 2.3 and Appendix D.3.1, we consider the truncated sequence $A_{\mathrm{quan}}^{(K\cdot\mathscr{T}/2)}$ of any possible quantum algorithm A_{quan} with $K<T$, and define a sequence of unitaries starting with $A_0=A_{\mathrm{quan}}^{(K\cdot\mathscr{T}/2)}$ as follows:

$$\begin{aligned}A_0 &:= V_{K+1}\widetilde{O}_{\mathbf{g};U}V_{K;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U}V_{K;1}\cdots\widetilde{O}_{\mathbf{g};U}V_{2;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U}V_{2;1}\widetilde{O}_{\mathbf{g};U}V_{1;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U}V_{1;1}\\ A_1 &:= V_{K+1}\widetilde{O}_{\mathbf{g};U}V_{K;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U}V_{K;1}\cdots\widetilde{O}_{\mathbf{g};U}V_{2;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U}V_{2;1}\widetilde{O}_{\mathbf{g};U_1}V_{1;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U_1}V_{1;1}\\ A_2 &:= V_{K+1}\widetilde{O}_{\mathbf{g};U}V_{K;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U}V_{K;1}\cdots\widetilde{O}_{\mathbf{g};U_2}V_{2;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U_2}V_{2;1}\widetilde{O}_{\mathbf{g};U_1}V_{1;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U_1}V_{1;1}\\ &\vdots\\ A_K &:= V_{K+1}\widetilde{O}_{\mathbf{g};U_K}V_{K;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U_K}V_{K;1}\cdots\widetilde{O}_{\mathbf{g};U_2}V_{2;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U_2}V_{2;1}\widetilde{O}_{\mathbf{g};U_1}V_{1;\mathscr{T}/2}\cdots\widetilde{O}_{\mathbf{g};U_1}V_{1;1},\end{aligned}\tag{33}$$

where $\widetilde{O}_{\mathbf{g};U_t}$ stands for the stochastic gradient oracle of the function $\tilde{f}_{t;U_t}$. Note that for the sequence of unitaries A_0, it can be decomposed into the product of V_{K+1} and K unitaries, each of the form

$$\mathscr{A}_k(n)=\widetilde{O}_{\mathbf{g};U}V_{k;n}\widetilde{O}_{\mathbf{g};U}\cdots\widetilde{O}_{\mathbf{g};U}V_{k;2}\widetilde{O}_{\mathbf{g};U}V_{k;1}$$

for $n=\mathscr{T}/2$ and $k\in[K]$ for some unitaries $V_1,\ldots,V_n$. In the following lemma, we demonstrate that for such a sequence $\mathscr{A}_k(n)$, if we replace $\widetilde{O}_{\mathbf{g};U}$ by another oracle that only reveals part information of f, the sequence will barely change on random inputs.

Lemma D.9. *For any $t\in[T-1]$ and any $n\le\frac{\mathscr{T}}{2}$, consider the following two sequences of unitaries*

$$\mathscr{A}(n)=\widetilde{O}_{\mathbf{g};U}V_n\widetilde{O}_{\mathbf{g};U}\cdots\widetilde{O}_{\mathbf{g};U}V_2\widetilde{O}_{\mathbf{g};U}V_1,$$

and

$$\hat{\mathscr{A}}_t(n)=\widetilde{O}_{\mathbf{g};U_t}V_n\widetilde{O}_{\mathbf{g};U_t}\cdots\widetilde{O}_{\mathbf{g};U_t}V_2\widetilde{O}_{\mathbf{g};U_t}V_1,$$

we have

$$\delta(n):=\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\left[\left\|\left(\hat{\mathscr{A}}_t(n)-\mathscr{A}(n)\right)|\psi\rangle\right\|^2\right]\le\frac{n}{36\mathscr{T}T^4}\tag{34}$$

for any pure state $|\psi\rangle$.

Proof. We use induction to prove this claim. First, for $n=1$, we have

$$\begin{aligned}&\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\left[\left\|\left(\hat{\mathscr{A}}_t(n)-\mathscr{A}(n)\right)|\psi\rangle\right\|^2\right]\\ &\quad=\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\left[\left\|\left(\widetilde{O}_{\mathbf{g};U}-\widetilde{O}_{\mathbf{g};U_t}\right)|\psi\rangle\right\|^2\right]\le\frac{1}{36\mathscr{T}^2T^4},\end{aligned}$$

where the last inequality follows from Lemma B.1. Suppose the inequality (26) holds for all $n \le \tilde{n}$ for some $\tilde{n} < \frac{T}{2}$. Then for $n = \tilde{n} + 1$, we have

$$\delta(\tilde{n}+1) = \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big[\|(\hat{\mathscr{A}_t}(n) - \mathscr{A}(n))\,|\psi\rangle\,\|^2\big] \tag{35}$$

$$\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big[\|(\widetilde{O}_{\mathbf{g};U} - \widetilde{O}_{\mathbf{g};U_t})\,|\psi_t\rangle\,\|^2\big] + \delta(\tilde{n}), \tag{36}$$

where

$$|\psi_t\rangle = V_{\tilde{n}}\widetilde{O}_{\mathbf{g};U_t}\cdots\widetilde{O}_{\mathbf{g};U_t}V_1\,|\psi\rangle \tag{37}$$

is a function of U_t obtained by $\tilde{n}$ queries to $\tilde{O}_{\mathbf{g};U_t}$. By Lemma D.1, we have

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_x\}}[\|\Pi_{t;\perp}\,|\psi_t\rangle\,\|^2] \le \frac{n}{18\mathscr{T}^2T^4} \le \frac{1}{36\mathscr{T}T^4}, \tag{38}$$

indicating that $|\psi_t\rangle$ only has a very little overlap with the subspace $W_{t;\perp}$ defined in (19), outside of which the columns $\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)}\}$ of U has no impact on the function value and derivatives of $\tilde{f}_{T;U}$. Thus,

$$\begin{aligned}\delta(\tilde{n}+1) &\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big[\|(\widetilde{O}_{\mathbf{g};U} - \widetilde{O}_{\mathbf{g};U_t})\,|\psi_t\rangle\,\|\big] + \delta(\tilde{n})\\ &\le \mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_x\}}[\|\Pi_{t;\perp}\,|\psi_t\rangle\,\|^2] + \delta(\tilde{n}) \le \frac{\tilde{n}+1}{36\mathscr{T}T^4},\end{aligned}$$

indicating that Eq. (34) also holds for $n = \tilde{n} + 1$. □

Lemma D.10 (A_t and A_{t-1} have similar outputs)**.** *For a hard instance $\tilde{f}_{T;U}(\mathbf{x})\colon \mathbb{R}^d \to \mathbb{R}$ defined on $\mathbb{B}(\mathbf{0}, 2\beta\sqrt{T})$ with $d \ge 2\mathscr{T}T\log\mathscr{T}$, let A_t for $t \in [K]$ be the sequence unitaries defined in Eq. (25). Then*

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big(\|A_t\,|\mathbf{0}\rangle - A_{t-1}\,|\mathbf{0}\rangle\,\|^2\big) \le \frac{1}{72T^4}.$$

Proof. From the definition of the unitaries in Eq. (33), we have

$$\|A_t\,|\mathbf{0}\rangle - A_{t-1}\,|\mathbf{0}\rangle\,\| = \|(\mathscr{A}(\mathscr{T}/2) - \hat{\mathscr{A}_t}(\mathscr{T}/2))\,|\psi\rangle\,\|$$

for some fixed quantum state $|\psi\rangle$ dependent on the vectors $\{\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(t-1)}\}$, where

$$\mathscr{A}(\mathscr{T}/2) = \widetilde{O}_{\mathbf{g};U}V_{\mathscr{T}/2}\widetilde{O}_{\mathbf{g};U}\cdots\widetilde{O}_{\mathbf{g};U}V_2\widetilde{O}_{\mathbf{g};U}V_1,$$

and

$$\hat{\mathscr{A}_t}(\mathscr{T}/2) = \widetilde{O}_{\mathbf{g};U_t}V_{\mathscr{T}/2}\widetilde{O}_{\mathbf{g};U_t}\cdots\widetilde{O}_{\mathbf{g};U_t}V_2\widetilde{O}_{\mathbf{g};U_t}V_1.$$

By Lemma D.9, we have

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big(\|(\mathscr{A}(T/2) - \hat{\mathscr{A}_t}(T/2))\,|\psi\rangle\,\|^2\big) \le \frac{1}{36\mathscr{T}T^4}\cdot\frac{\mathscr{T}}{2} = \frac{1}{72T^4},$$

which leads to

$$\mathbb{E}_{\{\mathbf{u}^{(t)},\ldots,\mathbf{u}^{(T)},M_{\mathbf{x}}\}}\big(\|A_t\,|\mathbf{0}\rangle - A_{t-1}\,|\mathbf{0}\rangle\,\|^2\big) \le \frac{1}{72T^4}.$$

□

Proposition D.11. *Consider the d-dimensional function $\tilde{f}_{T;U}(\mathbf{x})\colon \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T}) \to \mathbb{R}$ defined in (6) with the rotation matrix U being chosen arbitrarily and the dimension $d \ge 2\mathscr{T}T\log\mathscr{T}$ and $\mathcal{T} \ge T$. Consider the truncated sequence $A_{\text{quan}}^{(K\mathscr{T}/2)}$ of any possible quantum algorithm A_{quan} containing $K\mathscr{T}/2$ queries to the quantum stochastic gradient oracle $\widetilde{O}_{\mathbf{g};U}$ of $\mathbf{g}(\mathbf{x}, j)$ defined in Eq. (31) with $K < T$, and let p_U be the probability distribution over $\mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T})$ obtained by measuring the state $A_{\text{quan}}^{(K\mathscr{T}/2)}\,|0\rangle$, which is related to the rotation matrix U. Then,*

$$\Pr_{U,\{M_{\mathbf{x}}\},\mathbf{x}_{out}\sim p_U}\big[\|\nabla\tilde{f}_{T;U}(\mathbf{x}_{out})\| \le \alpha/\beta\big] \le \frac{1}{3},$$

where the probability is subject to all possible orthogonal rotation matrices U, and all possible matrices $\{M_{\mathbf{x}}\}$ in the quantum stochastic gradient function $\mathbf{g}(\mathbf{x}, j)$ for any $\mathbf{x}$.

Proof. We first demonstrate that the sequence of unitaries A_K defined in Eq. (33) cannot find an α/β-approximate stationary point with high probability. In particular, let p_{U_K} be the probability distribution over $\mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T})$ obtained by measuring the output state $A_K |0\rangle$. Then,

$$\Pr_{U_K, \{M_{\mathbf{x}}\}, \mathbf{x}_{\text{out}} \in p_{U_K}} \big[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \le \alpha/\beta\big] \le \Pr_{\{\mathbf{u}^{(k+1)}, \ldots, \mathbf{u}^{(T)}\}} \big[\|\nabla \tilde{f}_{T;U}(\mathbf{x})\| \le \alpha/\beta\big].$$

for any fixed $\mathbf{x}$. Then by Lemma B.2 we have

$$\Pr_{U_K, M_{\mathbf{x}}, \mathbf{x}_{\text{out}} \in p_{U_K}} \big[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \le \alpha/\beta\big] \le \frac{1}{6}.$$

Moreover, by Lemma D.10 and Cauchy-Schwarz inequality, we have

$$\mathbb{E}_U \big[\|A_K |0\rangle - A_0 |0\rangle\|^2\big] \le K \cdot \mathbb{E}_U \Big[\sum_{t=1}^{K-1} \|A_{t+1} |0\rangle - A_t |0\rangle\|^2\Big] \le \frac{1}{72T^2}.$$

Then by Markov's inequality,

$$\Pr_U \Big[\|A_{K-1} |0\rangle - A_0 |0\rangle\|^2 \ge \frac{1}{6T}\Big] \le \frac{1}{6T},$$

since both norms are at most 1. Thus, the total variance distance between the probability distribution p_U obtained by measuring $A_0 |0\rangle$ and the probability distribution p_{U_K} obtained by measuring $A_K |0\rangle$ is at most

$$\frac{1}{6T} + \frac{1}{6T} = \frac{1}{3T} \le \frac{1}{6}.$$

Hence, we can conclude that

$$\Pr_{U, \{M_{\mathbf{x}}\}, \mathbf{x}_{\text{out}} \sim p_U^{(t)}} \big[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \le \alpha/\beta\big] \le \Pr_{U_K, \{M_{\mathbf{x}}\}, \mathbf{x}_{\text{out}} \sim p_{U_K}} \big[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \le \alpha/\beta\big] + \frac{1}{6} = \frac{1}{3}.$$

□

Proposition D.12. *Suppose Δ, $\bar{L}$, σ, and ϵ are positive. Then,*

$$\mathcal{T}_\epsilon^{\text{stoc}}\big(\mathcal{A}_{\text{quan}}, \tilde{\mathcal{F}}_1(\Delta, \bar{L}, \mathcal{R}), \sigma\big) = \Omega\Big(\frac{\Delta \bar{L} \sigma}{\epsilon^3}\Big),$$

if we further assume the stochastic gradient function $\mathbf{g}(\mathbf{x})$ satisfies Assumption 1.3 with mean-squared smoothness parameter $\bar{L}$, where $\mathcal{R} = \sqrt{\frac{\hat{\ell}\sigma\Delta}{6\ell\bar{L}\gamma\epsilon}}$, the complexity measure $\mathcal{T}_\epsilon^{\text{stoc}}(\cdot)$ is defined in Eq. (14), and the function class $\tilde{\mathcal{F}}_1(\Delta, \bar{L}, \mathcal{R})$ is defined in Definition C.1. The lower bound holds even if we restrict $\tilde{\mathcal{F}}_1(\Delta, \bar{L}, \mathcal{R})$ to functions whose domain has dimension

$$\tilde{\Theta}\Big(\frac{\Delta \bar{L} \sigma}{\epsilon^3}\Big).$$

Proof. We set up the scaling parameters α and β in the hard instance $\tilde{f}_{T;U} : \mathbb{R}^d \to \mathbb{R}$ defined in Eq. (6) as

$$\alpha = \frac{L\beta^2}{\ell}, \qquad \beta = \frac{2\ell\epsilon}{L},$$

where ℓ is the gradient Lipschitz constant of $\bar{f}_T$ whose value is given in Lemma A.3, and the parameter $L \le \bar{L}$ is specified later. We also set the parameters

$$T = \frac{L\Delta}{48\ell\epsilon^2}, \qquad \mathscr{T} = \frac{\sigma^2\beta^2}{4\gamma^2\alpha^2} = \frac{\sigma^2}{4\gamma^2\epsilon^2}.$$

Then by Lemma A.3, we know that $\tilde{f}_{T;U}$ is L-smooth and thus $\bar{L}$-smooth since $L \le \bar{L}$, and

$$\tilde{f}_{T;U}(\mathbf{0}) - \inf_{\mathbf{x}} \tilde{f}_{T;U}(\mathbf{x}) = \alpha\big(\bar{f}_T(\mathbf{0}) - \inf_{\mathbf{x}} \bar{f}_T(\mathbf{x})\big) \le \frac{12L\beta^2}{\ell} \cdot T \le \Delta,$$

indicating that $\tilde{f}_{T;U} \in \tilde{\mathcal{F}}(\Delta, L_p, \mathcal{R})$ for arbitrary dimension d and rotation matrix U. Moreover, for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, by Lemma D.7 we have

$$\mathbb{E}\big[\|\mathbf{g}(\mathbf{x}, j) - \nabla \tilde{f}_{T;U}(\mathbf{x})\|^2\big] \le 4\alpha^2\gamma^2\mathscr{T}/\beta^2 \le \delta^2,$$

indicating that the variance of the stochastic gradient function $\mathbf{g}$ defined in Eq. (31) is bounded by σ^2, and

$$\mathbb{E}\big\|\hat{\mathbf{g}}(\mathbf{x}, j) - \hat{\mathbf{g}}(\mathbf{y}, j)\big\|^2 \le \frac{\hat{\ell}^2\mathscr{T}\alpha^2\|\mathbf{x}-\mathbf{y}\|^2}{\beta^2} = \frac{\hat{\ell}^2 L^2\mathscr{T}}{\ell^2} \cdot \|\mathbf{x}-\mathbf{y}\|^2.$$

Hence, to guarantee that Assumption 1.3 is satisfied, we set

$$L = \frac{\ell}{\hat{\ell}\sqrt{\mathscr{T}}} \cdot \bar{L} = \frac{2\ell\gamma\epsilon}{\hat{\ell}\sigma} \cdot \bar{L}.$$

Furthermore, we notice that the radius $\mathcal{R}$ satisfies

$$\mathcal{R} = 2\beta\sqrt{T} = \frac{4\ell\epsilon}{L} \cdot \sqrt{\frac{L\Delta}{48\ell\epsilon^2}} = \sqrt{\frac{\ell\Delta}{3L}} = \sqrt{\frac{\hat{\ell}\sigma\Delta}{6\ell\bar{L}\gamma\epsilon}}.$$

To guarantee that $\mathscr{T} \ge T$, ϵ has to satisfy

$$\frac{\sigma^2}{4\gamma^2\epsilon^2} \ge \frac{\Delta}{48\ell\epsilon^2} \cdot \frac{2\ell\gamma\epsilon\bar{L}}{\hat{\ell}\sigma},$$

indicating

$$\epsilon \le \frac{\sigma^2}{4\gamma^2} \cdot \frac{24\hat{\ell}\sigma}{\Delta\gamma\bar{L}} = \frac{6\sigma^2\hat{\ell}}{4\gamma^3\Delta\bar{L}}.$$

By Proposition D.4, for any truncated sequence $A_{\mathrm{quan}}^{(K\mathscr{T}/2)}$ of any possible quantum algorithm A_{quan} containing $K\mathscr{T}/2 < T\mathscr{T}/2$ queries to the oracle $O_f^{(p)}$ on input domain $\mathbb{B}(0, \mathcal{R})$, we have

$$\Pr_{U, \mathbf{x}_{\text{out}} \sim p_U}\big[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \le \alpha/\beta\big] = \Pr_{U, \mathbf{x}_{\text{out}} \sim p_U}\big[\|\nabla \tilde{f}_{T;U}(\mathbf{x}_{\text{out}})\| \le \epsilon\big] \le \frac{1}{3},$$

where p_U is the probability distribution over $\mathbf{x} \in \mathbb{B}(\mathbf{0}, 2\beta\sqrt{T}) = \mathbb{B}(\mathbf{0}, \mathcal{R})$ obtained by measuring the state $A_{\mathrm{quan}}^{(K\mathscr{T}/2)}|0\rangle$, given that the dimension d satisfies

$$d \ge 2\mathscr{T}T\log\mathscr{T} = \tilde{\Theta}\Big(\frac{\Delta\bar{L}\sigma}{\epsilon^3}\Big).$$

Then according to Definition 2.2 we can conclude that

$$\mathcal{T}_\epsilon^{\mathrm{stoc}}\big(\mathcal{A}_{\mathrm{quan}}, \tilde{\mathcal{F}}_1(\Delta, \bar{L}, \mathcal{R}, \sigma)\big) \ge \frac{\mathscr{T}T}{2} = \tilde{\Omega}\Big(\frac{\Delta\bar{L}\sigma}{\epsilon^3}\Big).$$

□

D.4.3. Lower Bound with Unbounded Input Domain

In this subsection, we extend the quantum lower bound proved in Proposition D.12 to the function class $\mathcal{F}(\Delta, \bar{L})$ with unbounded input domain via similar scaling techniques adopted in Arjevani et al. (2022), Appendix C.2, and Appendix D.3.2. In particular, we consider the scaled hard instance $\hat{f}_{T;U}$ introduced in Carmon et al. (2020a) and also used in Arjevani et al. (2022),

$$\hat{f}_{T;U}(\mathbf{x}) := \tilde{f}_{T;U}(\chi(\mathbf{x})) + \frac{\alpha}{10} \cdot \frac{\|\mathbf{x}\|^2}{\beta^2},$$

where

$$\chi(\mathbf{x}) := \frac{\mathbf{x}}{\sqrt{1 + \|\mathbf{x}\|^2/\hat{\mathcal{R}}^2}},$$

with the following parameters

$$\alpha = \frac{L\beta^2}{\ell} \qquad \beta = \frac{2\ell\epsilon}{L}, \qquad T = \frac{L\Delta}{48\ell\epsilon}, \qquad L = \frac{2\ell\gamma\epsilon\bar{L}}{\hat{\ell}\sigma}, \qquad \hat{\mathcal{R}} = 230\beta\sqrt{T},$$

whose values are also adopted in the proof of Proposition D.12. The constants in $\hat{f}_{T;U}$ are chosen carefully such that stationary points of $\hat{f}_{T;U}$ are in one-to-one correspondence to stationary points of the hard instance $\tilde{f}_{T;U}$ concerning the setting with bounded input domain. Quantitatively,

Lemma D.13 (Arjevani et al. 2022, Section 4)**.** *Let Δ, $\bar{L}$, and ϵ be positive constants. Then, under the following choice of parameters*

$$\alpha = \frac{L\beta^2}{\ell}, \qquad \beta = \frac{2\ell\epsilon}{L}, \qquad T = \frac{L\Delta}{48\ell\epsilon}, \qquad L = \frac{2\ell\gamma\epsilon\bar{L}}{\hat{\ell}\sigma},$$

where ℓ is the gradient Lipschitz parameter of $\bar{f}_T$ whose value is given in Lemma A.3, such that for any function pairs $(\tilde{f}_{T;U}, \hat{f}_{T;U}) \in \tilde{\mathcal{F}}_1(\Delta, \bar{L}, \mathcal{R}) \times \mathcal{F}_1(\Delta, \bar{L})$ with dimension $d \geq 400T\log T$ and the same rotation matrix U, there exists a bijection between the ϵ-stationary points of $\tilde{f}_{T;U}$ and the ϵ-stationary points of $\hat{f}_{T;U}$ that is independent from U.

Equipped with Lemma D.13, we are now ready to prove Theorem 3.8.

Proof of Theorem 3.8. Note that one quantum query to the stochastic gradient of $\hat{f}_{T;U}$ can be implemented by one quantum query to the stochastic gradient of $\tilde{f}_{T;U}$ with the same rotation U, if we directly scale the stochastic gradient function of $\tilde{f}_{T;U}$ to $\hat{f}_{T;U}$, which will not increase the variance of the stochastic gradient function, and the mean-squared smoothness condition in Assumption 1.3 is still preserved with the same mean-squared smoothness parameter $\bar{L}$. Combined with Lemma D.13, we can note that the problem of finding ϵ-stationary points of $\tilde{f}_{T;U}$ with unknown U can be reduced to the problem of finding ϵ-stationary points of $\hat{f}_{T;U}$ with no additional overhead in terms of query complexity. Then by Proposition D.12, we can conclude that

$$\mathcal{T}_\epsilon^{\mathrm{stoc}}\big(\mathcal{A}_{\mathrm{quan}}, \mathcal{F}_1(\Delta, \bar{L}), \sigma\big) \geq \mathcal{T}_\epsilon^{\mathrm{stoc}}\big(\mathcal{A}_{\mathrm{quan}}, \tilde{\mathcal{F}}_1(\Delta, \bar{L}, \mathcal{R}), \sigma\big) = \Omega\Big(\frac{\Delta\bar{L}\sigma}{\epsilon^3}\Big),$$

if we further assume the stochastic gradient function satisfies Assumption 1.3. The dimension dependence is the same as Proposition D.12. □