

Gradient Testing and Estimation by Comparisons

Xiwen Tao^{1,3,*} Chenyi Zhang^{2,*} Helin Wang^{1,3} Yexin Zhang^{3,4} Tongyang Li^{3,4,†}

¹ School of Electronics Engineering and Computer Science, Peking University

² Computer Science Department, Stanford University

³ Center on Frontiers of Computing Studies, Peking University

⁴ School of Computer Science, Peking University

Abstract

We study gradient testing and gradient estimation of smooth functions using only a comparison oracle that, given two points, indicates which one has the larger function value. For any smooth $f: \mathbb{R}^n \rightarrow \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$, and $\varepsilon > 0$, we design a gradient testing algorithm that determines whether the normalized gradient $\nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$ is ε -close or 2ε -far from a given unit vector $\mathbf{v}$ using $O(1)$ queries, as well as a gradient estimation algorithm that outputs an ε -estimate of $\nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$ using $O(n \log(1/\varepsilon))$ queries which we prove to be optimal. Furthermore, we study gradient estimation in the quantum comparison oracle model where queries can be made in superpositions, and develop a quantum algorithm using $O(\log(n/\varepsilon))$ queries.

1 Introduction

Optimization is foundational to machine learning, since the training of a neural network is equivalent to the minimization of its loss function. However, there are common scenarios where access to gradients is infeasible or computationally prohibitive, such as black-box adversarial attack on neural networks [14, 37, 40] and policy search in reinforcement learning [17, 45]. This motivates the study of optimization algorithms that operate with zeroth-order information, i.e., function-value evaluations. A series of work has established rigorous convergence and complexity guarantees for such methods in both convex optimization [21, 39] and nonconvex optimization [4, 22, 25, 27, 28, 50, 53].

More recently, optimization algorithms are soliciting for even less information. In particular, an intriguing setting is to only have access to comparisons of function values. Formally, for a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, its comparison oracle is defined as $O_f^{\text{Comp}}: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \{-1, 1\}$ that for any pair of inputs $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n$, it satisfies

$$O_f^{\text{Comp}} = \begin{cases} 1 & f(\mathbf{x}) \geq f(\mathbf{y}) \\ -1 & f(\mathbf{x}) \leq f(\mathbf{y}) \end{cases}. \quad (1)$$

(When $f(\mathbf{x}) = f(\mathbf{y})$, it is allowed to output either 1 or -1 .)

Several studies have developed optimization algorithms under this comparison setting. As a class of gradient-free optimization methods (see the survey by [32]), these works used comparisons once or

*Equal contribution.

†Corresponding author. Email: tongyangli@pku.edu.cn

This work subsumes the note “Comparisons are all You need for optimizing smooth functions” [52].

multiple times per iteration to approximately estimate the direction of the gradient. This idea can be traced back to [38], which iteratively updates a simplex of points based on comparisons to seek lower objective values. However, the Nelder-Mead method may fail to converge to a stationary point for smooth functions [20]. Recent works, including the Stochastic Three-Point (STP) method [6] and dueling optimization [44], update the optimization direction by randomly sampling directions per iteration and using comparisons to determine the relative ordering of function values.

Furthermore, Ref. [16] estimated a descent direction using comparisons between candidate points based on randomized probing at each iteration. Ref. [30] showed how to use comparison queries or direction preference queries inside ellipsoid methods for solving smooth convex problems. In a different line of work, [10] leveraged ideas from one-bit compressed sensing to recover gradient information from noisy comparison feedback. More recently, Ref. [48] introduced a rank-based stochastic gradient estimator that uses the partial ordering of the best candidates to infer a descent direction.

Despite the importance of gradients in optimization, prior work has not pinned down the optimal algorithm for gradient estimation using comparison queries. An existing result is due to [30], which achieved $O(n^2 \log n)$ comparison queries per gradient direction estimate at a fixed angular accuracy. Another related work is dueling optimization [44], which provided convergence guarantees based on update directions that are aligned with the true gradient in expectation; however, their analysis does not characterize the query complexity required to estimate gradient directions to a designated accuracy.

Beyond that, gradient testing is also a natural problem to explore. In general, property testing decides whether a given object has a certain property or is significantly different from any object that has the property. It typically brings more efficient algorithms than property estimation, and may circumvent significant costs when the input size is large or an exact algorithm is expensive. Therefore, property testing is a main algorithmic topic [26, 43], especially on discrete problems. However, its study on continuous optimization is scarce, which motivates us to study gradient testing together with gradient estimation.

Recently, following the rapid advancement of quantum computing [41, 42], quantum algorithms that provide speedup for solving optimization problems have been systematically investigated. These include constrained optimization problems covering linear programs [7, 11, 23], second-order cone programming [24, 31], and semidefinite programs [1, 3, 8, 9]. General convex optimization [2, 12, 46] and nonconvex optimization [15, 33, 34, 36, 51] are also extensively studied. See the survey by [18] for more details. In the quantum setting, it is also very natural to study optimization by comparisons, particularly because current quantum computers are noisy and simpler information directly benefits implementation of quantum algorithms in the near term. In particular, we study the setting where we are given a quantum comparison oracle

$$O_{f,Q}^{\text{Comp}} |\mathbf{x}\rangle |\mathbf{y}\rangle |z\rangle \rightarrow |\mathbf{x}\rangle |\mathbf{y}\rangle |z \oplus O_f^{\text{Comp}}(\mathbf{x}, \mathbf{y})\rangle \quad (2)$$

which is the quantum generalization of the comparison oracle O_f^{Comp} in (1), where $\oplus$ means XOR.

Main results. In this paper, we systematically study gradient testing and estimation, formally stated as follows:

Problem 1 (Gradient testing). *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be L -smooth, $\varepsilon \in (0, 1/\sqrt{2})$ and $\gamma > 0$ be fixed parameters. Given a point $\mathbf{x} \in \mathbb{R}^n$ and a unit vector $\mathbf{v} \in \mathbb{R}^n$, the goal is to decide whether*

$$\left\| \frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|} - \mathbf{v} \right\| \leq \varepsilon \quad \text{or} \quad \left\| \frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|} - \mathbf{v} \right\| > 2\varepsilon,$$

where we are promised that one of the two cases holds and $\|\nabla f(\mathbf{x})\| \geq \gamma$.

Problem 2 (Gradient estimation). *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be L -smooth, $\varepsilon \in (0, 1/\sqrt{2})$ and $\gamma > 0$ be fixed parameters. Given a point $\mathbf{x} \in \mathbb{R}^n$ satisfying $\|\nabla f(\mathbf{x})\| \geq \gamma$, the goal is to output a unit vector $\mathbf{v} \in \mathbb{R}^n$ satisfying $\left\| \frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|} - \mathbf{v} \right\| \leq \varepsilon$.*

Both problems are scale-invariant: multiplying the objective by a positive constant does not change either the comparison oracle or the normalized gradient direction. The lower bound $\|\nabla f(\mathbf{x})\| \geq \gamma$ is therefore a nondegeneracy condition rather than an additional source of query complexity. It is necessary because the normalized direction is undefined at stationary points, and, when the gradient norm is arbitrarily small, the first-order term used by directional-preference comparisons can be dominated by the L -smoothness error. In our algorithms, γ only determines the probing radius used in each comparison; the number of comparison queries is independent of the numerical value of γ . Once an estimate $\mathbf{h} \approx \nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$ is obtained, it can be used directly in normalized gradient descent updates of the form $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{h}_t$, so the local estimation primitive can be plugged into comparison-based optimization procedures. Numerical experiment can be seen in [Section 5](#).

For these two problems, we develop classical and quantum algorithms with comparison query complexities as follows:

- Gradient testing ([Algorithm 2](#)): $O(1)$ queries to O_f^{Comp} .
- Classical gradient estimation ([Algorithm 4](#)): $O(n \log(1/\varepsilon))$ queries to O_f^{Comp} .
- Quantum gradient estimation ([Algorithm 5](#)): $O(\log(n/\varepsilon))$ queries to $O_{f,Q}^{\text{Comp}}$.

In addition, we prove that our classical algorithm for gradient estimation is optimal. Another note is that our classical algorithm for gradient testing is randomized. In contrast, we prove a tight $\Theta(n)$ bound for deterministic algorithms; the corresponding algorithm and lower bound proof are detailed in [Appendix B](#).

For quantum gradient estimation, we prove an $\Omega(\log(1/\varepsilon))$ lower bound, which shows that our algorithm is optimal up to a factor of $\log n$. See a summary in [Table 1](#).

Table 1: Summary of our main results: classical and quantum algorithms for gradient testing and estimation by comparisons.

	Gradient Testing	Gradient Estimation
Classical	$\Theta(1)$ (Theorem 1)	$\Theta(n \log(1/\varepsilon))$ (Theorem 3 and Theorem 5)
Quantum		$O(\log(n/\varepsilon))$ (Theorem 4), $\Omega(\log(1/\varepsilon))$ (Theorem 6)

2 Preliminaries

2.1 Notation

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable L -smooth function, i.e.,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \quad (3)$$

For any vector $\mathbf{v}$, $\|\mathbf{v}\|$ represents its ℓ_2 -norm. For any $\mathbf{x} \in \mathbb{R}^n$, we use $\mathbf{g}(\mathbf{x}) := \nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$ to denote the normalized gradient at $\mathbf{x}$. We omit $\mathbf{x}$ when it is clear from context. We denote $\mathbf{e}_i$

to be the unit vector of the i -th coordinate component. We denote $S_n := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$ to be the unit sphere in $\mathbb{R}^n$, and $\mathbf{x} \sim S_n$ means that $\mathbf{x}$ is drawn from the uniform distribution on S_n . The orthogonal group $G(n)$ is the group of all real $n \times n$ matrices Q satisfying $Q^\top Q = I$, which represents all orthonormal transformations of $\mathbb{R}^n$. The Haar measure on $G(n)$ is the unique probability measure that is invariant under left and right multiplication, and sampling from the Haar measure corresponds to choosing an orthogonal matrix uniformly at random.

2.2 Inner Product Concentration for Random Vectors

Throughout the paper, we use the following estimation of the inner product between a fixed unit vector and a unit vector selected from S_n uniformly at random:

Lemma 1. *Let $n \geq 5$. For any $\mathbf{x} \in \mathbb{R}^n$, and any constant $c > 0$, there exist constants p_1 and p_2 such that*

$$p_1 \leq \Pr_{\mathbf{y} \sim S_n} [|\langle \mathbf{y}, \mathbf{x} \rangle| \leq \|\mathbf{x}\|/(c\sqrt{n})] \leq p_2,$$

where $\mathbf{y}$ is chosen from S_n uniformly at random. In particular, we have the inequalities below:

- $\Pr_{\mathbf{y} \sim S_n} [|\langle \mathbf{y}, \mathbf{x} \rangle| \leq 24\|\mathbf{x}\|/(25\sqrt{n})] \geq 3/5$
- $\Pr_{\mathbf{y} \sim S_n} [|\langle \mathbf{y}, \mathbf{x} \rangle| \leq 18\|\mathbf{x}\|/(25\sqrt{n})] \leq 11/20$
- $\Pr_{\mathbf{y} \sim S_n} [|\langle \mathbf{y}, \mathbf{x} \rangle| \geq \|\mathbf{x}\|/(5\sqrt{n})] \geq 4/5$

The lemma is adapted from Theorem 3.4.5 in [49]. The proof of [Lemma 1](#) is deferred to [Appendix A.1](#).

2.3 Basic Notation in Quantum Computing

We use the Dirac notation $|\cdot\rangle$ to represent quantum states, which can be seen as column vectors. The basic unit of quantum state, qubit, is represented by a normalized vector in $\mathbb{C}^2$, which has the superposition form $|\phi\rangle = \alpha|0\rangle + \beta|1\rangle$ on the computational basis $\{|0\rangle, |1\rangle\}$ with $\alpha, \beta \in \mathbb{C}$ and $|\alpha|^2 + |\beta|^2 = 1$. An n -qubit system is represented in a 2^n -dimensional complex vector space, with basis states $\{|0\rangle, |1\rangle\}^{\otimes n}$. A general n -qubit state can be expressed as $|\psi\rangle = \sum_{x \in \{0,1\}^n} \alpha_x |x\rangle$ with $\sum_x |\alpha_x|^2 = 1$. Quantum measurement on the computational basis collapses the state $|\psi\rangle = \sum_x \alpha_x |x\rangle$ to one of the basis states $|x\rangle$ with probability $|\alpha_x|^2$. Quantum operations are modeled as unitary transformations, which are linear operators U satisfying $U^\dagger U = I$, where $U^\dagger$ is the conjugate transpose of U .

A quantum algorithm can access a function via queries to a quantum oracle. For a classical function g , the oracle O_g is a unitary transformation that maps $|x\rangle|b\rangle$ to $|x\rangle|b \oplus g(x)\rangle$. This allows the oracle to be queried on superpositions of inputs, producing corresponding superpositions of outputs.

Quantum Fourier transform (QFT) is the quantum analogue of the discrete Fourier transform, but operates on amplitudes of quantum states. Formally, given the computational basis $\{|0\rangle, |1\rangle, \dots, |N-1\rangle\}$, it acts as follows:

$$\text{QFT}: |x\rangle \mapsto \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} e^{2\pi i x k / N} |k\rangle.$$

A key application of QFT is Jordan's gradient estimation algorithm [29]. Specifically, if the phase of each basis state is given by an inner product, then applying the inverse QFT and measuring in the computational basis allows one to recover the underlying vector with high probability.

Proposition 1 (informal version of [Proposition 3](#)). *Let $n \geq 2$ and $t \in \mathbb{Z}_+$. There exists a quantum procedure that, given an input state encoding $\mathbf{x} \in [0, 1]^n$ as $|\psi_{\mathbf{x}}\rangle = \frac{1}{\sqrt{(t+1)^n}} \sum_{\mathbf{y} \in \{0,1,\dots,t\}^n} e^{2\pi i \langle \mathbf{y}, \mathbf{x} \rangle} |\mathbf{y}\rangle$, outputs an estimate $\hat{\mathbf{v}} \in \{0, 1/t, \dots, 1\}^n$ such that*

$$\Pr \left[\|\hat{\mathbf{v}} - \mathbf{x}\|_2 = O(n^{1.5}/t) \right] \geq 2/3.$$

Given input state $|\psi\rangle$ satisfying $\| |\psi\rangle - |\psi_{\mathbf{x}}\rangle \| \leq \varepsilon$, the success probability decreases by at most $O(\varepsilon)$.

Compared with Lemma 2.2 of [12], which gives a per-coordinate error bound for the inverse QFT, the proposition here instead bounds the overall error in ℓ_2 norm. A formal statement with proof is deferred to [Appendix C](#).

3 Classical Algorithms

In this section, we present our algorithms for [Problem 1](#) and [Problem 2](#) given query access to a comparison oracle O_f^{Comp} defined in (1). We begin by developing a subroutine called “directional preference” that is used in both algorithms.

3.1 Directional Preference

Here we show that given $\mathbf{x} \in \mathbb{R}^n$ and a unit vector $\mathbf{v} \in \mathbb{R}^n$, we can use one comparison query to understand whether the inner product $\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle$ is roughly positive or negative, which determines whether $\mathbf{v}$ is following or against the direction of $\nabla f(\mathbf{x})$. This is also known as *directional preference* (DP) in [30].

Lemma 2. *Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be an L -smooth function. Given a point $\mathbf{x} \in \mathbb{R}^d$, a unit vector $\mathbf{v} \in S_d$, and precision $\Delta > 0$ for directional preference. Then we have:*

- If $O_f^{\text{Comp}}(\mathbf{x} + \frac{2\Delta}{L}\mathbf{v}, \mathbf{x}) = 1$, then $\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle \geq -\Delta$.
- If $O_f^{\text{Comp}}(\mathbf{x} + \frac{2\Delta}{L}\mathbf{v}, \mathbf{x}) = -1$, then $\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle \leq \Delta$.

Algorithm 1: DP($\mathbf{x}, \mathbf{v}, \Delta$)

Input: Comparison oracle O_f^{comp} for $f: \mathbb{R}^n \rightarrow \mathbb{R}$, point $\mathbf{x} \in \mathbb{R}^n$, unit vector $\mathbf{v} \in S_d$, parameter $\Delta > 0$

- 1 **if** $O_f^{\text{comp}}(\mathbf{x} + \frac{2\Delta}{L}\mathbf{v}, \mathbf{x}) = 1$ **then**
- 2 **return** “ $\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle \geq -\Delta$ ”
- 3 **else**
- 4 **return** “ $\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle \leq \Delta$ ”

Proof. Since f is an L -smooth differentiable function,

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \frac{1}{2}L\|\mathbf{y} - \mathbf{x}\|^2$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Take $\mathbf{y} = \mathbf{x} + \frac{2\Delta}{L}\mathbf{v}$, this gives

$$f(\mathbf{y}) - f(\mathbf{x}) - \frac{2\Delta}{L}\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle \leq \frac{1}{2}L\left(\frac{2\Delta}{L}\right)^2 = \frac{2\Delta^2}{L}.$$

Therefore, if $O_f^{\text{comp}}(\mathbf{y}, \mathbf{x}) = 1$, i.e., $f(\mathbf{y}) \geq f(\mathbf{x})$,

$$\frac{2\Delta}{L}\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle \geq \frac{2\Delta}{L}\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle + f(\mathbf{x}) - f(\mathbf{y}) \geq -\frac{2\Delta^2}{L},$$

and hence

$$\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle \geq -\Delta.$$

On the other hand, if $O_f^{\text{comp}}(\mathbf{y}, \mathbf{x}) = -1$, i.e., $f(\mathbf{y}) \leq f(\mathbf{x})$, we have

$$\frac{2\Delta}{L}\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle \leq f(\mathbf{y}) - f(\mathbf{x}) + \frac{2\Delta^2}{L} \leq \frac{2\Delta^2}{L},$$

and hence $\langle \nabla f(\mathbf{x}), \mathbf{v} \rangle \leq \Delta$. □

3.2 Gradient Testing

Here we present our algorithm for gradient testing ([Problem 1](#)). Without loss of generality, we assume the given unit vector $\mathbf{v}$ satisfies $\mathbf{v} = (1, 0, \dots, 0)^\top$.

Algorithm 2: Classical Gradient Testing

Input: testing direction $\mathbf{v}$, lower bound γ on $\|\nabla f\|$

- 1 Set $\Delta = \varepsilon\gamma/(25\sqrt{2}n)$, $\delta = 1/3$, $T = \lceil 3200 \ln(1/\delta) \rceil$
- 2 Initialize $N = 0$
- 3 **for** $i = 1$ to T **do**
- 4 Randomly choose $\mathbf{y} \sim S_{n-1}$ and define $\alpha_{\mathbf{y}} = (-\varepsilon/\sqrt{(n-1)(1-\varepsilon^2)}, \mathbf{y})$
- 5 Call [Algorithm 1](#) with input $(\mathbf{x}, \alpha_{\mathbf{y}}/\|\alpha_{\mathbf{y}}\|, \Delta)$
- 6 **if** $\langle \nabla f, \alpha_{\mathbf{y}}/\|\alpha_{\mathbf{y}}\| \rangle \leq \Delta$ **then**
- 7 $N \leftarrow N + 1$
- 8 **if** $N/T \geq 63/80$ **then**
- 9 **return** Yes
- 10 **else**
- 11 **return** No

The high-level idea of [Algorithm 2](#) is as follows. For our convenience, we decompose $\mathbf{g}$ as $\mathbf{g} = (g_1, \tilde{\mathbf{g}})$ and denote $\mathbf{u} = \mathbf{v} - \mathbf{g}$. Note that $\langle \tilde{\mathbf{g}}, \mathbf{y} \rangle$ is related to the length of the norm $\|\tilde{\mathbf{g}}\|$. If we set a comparison standard k , then for different $\|\tilde{\mathbf{g}}\|$, the probability of $\mathbf{y}$ such that the inner product is less than k is also different. Specifically, we set $k = \varepsilon/\sqrt{n-1}$ here and in this case, there is a constant difference between the different probability. As a result, we can randomly choose a constant number of $\mathbf{y}$ to distinguish them.

Theorem 1. *Algorithm 2 solves Problem 1 using $O(1)$ queries to a comparison oracle O_f^{Comp} (1) with success probability at least $2/3$.*

Proof of Theorem 1. Since $\|\mathbf{g}\| = 1$, we have

$$\|\mathbf{g} - \mathbf{e}_1\|^2 = (1 - g_1)^2 + g_2^2 + \dots + g_n^2 = (1 - g_1)^2 + 1 - g_1^2 = 2 - 2g_1.$$

If $\|\mathbf{g} - \mathbf{e}_1\| \leq \varepsilon$, we can get $g_1 \geq 1 - \varepsilon^2/2$, which means

$$\|\tilde{\mathbf{g}}\|^2 = 1 - g_1^2 \leq 1 - (1 - \varepsilon^2/2)^2 = \varepsilon^2 - \varepsilon^4/4 \leq \varepsilon^2, \quad (4)$$

and

$$\varepsilon \langle \mathbf{v}, \mathbf{g} \rangle / \sqrt{1 - \varepsilon^2} = \varepsilon g_1 / \sqrt{1 - \varepsilon^2} = \varepsilon \sqrt{(1 - \|\tilde{\mathbf{g}}\|^2) / (1 - \varepsilon^2)} \geq \|\tilde{\mathbf{g}}\|. \quad (5)$$

If $\|\mathbf{g} - \mathbf{e}_1\| \geq 2\varepsilon$, we can get $g_1 \leq 1 - 2\varepsilon^2$, which means

$$\|\tilde{\mathbf{g}}\|^2 = \sum_{i=2}^n g_i^2 = 1 - g_1^2 \geq 1 - (1 - 2\varepsilon^2)^2 = 4\varepsilon^2 - 4\varepsilon^4 \geq 2\varepsilon^2, \quad (6)$$

and

$$\varepsilon \langle \mathbf{v}, \mathbf{g} \rangle / \sqrt{1 - \varepsilon^2} = \varepsilon \sqrt{(1 - \|\tilde{\mathbf{g}}\|^2) / (1 - \varepsilon^2)} \leq \|\tilde{\mathbf{g}}\| / \sqrt{2}, \quad (7)$$

since $\varepsilon \leq 1/\sqrt{2}$. Consider our task to determine whether $\|\mathbf{u}\| \leq \varepsilon$ or $\|\mathbf{u}\| \geq 2\varepsilon$. By (4) and (6), it suffices to determine whether $\|\tilde{\mathbf{g}}\| \leq \varepsilon$ or $\|\tilde{\mathbf{g}}\| \geq \sqrt{2}\varepsilon$. For any $\mathbf{y} \in S_{n-1}$, if we call Algorithm 1 with input $(\mathbf{x}, \alpha_{\mathbf{y}}/\|\alpha_{\mathbf{y}}\|, \Delta)$ and get $\langle \nabla f(\mathbf{x}), \alpha_{\mathbf{y}}/\|\alpha_{\mathbf{y}}\| \rangle \leq \Delta$, we have

$$\langle \mathbf{y}, \tilde{\mathbf{g}} \rangle = \langle (0, \mathbf{y}), \mathbf{g} \rangle \leq \frac{\varepsilon}{\sqrt{(n-1)(1-\varepsilon^2)}} \langle \mathbf{v}, \mathbf{g} \rangle + \Delta \|\alpha_{\mathbf{y}}\| \leq \frac{\varepsilon}{\sqrt{(n-1)(1-\varepsilon^2)}} \langle \mathbf{v}, \mathbf{g} \rangle + \frac{\varepsilon}{25n}.$$

Otherwise, we have

$$\langle \mathbf{y}, \tilde{\mathbf{g}} \rangle \geq \frac{\varepsilon}{\sqrt{(n-1)(1-\varepsilon^2)}} \langle \mathbf{v}, \mathbf{g} \rangle - \frac{\varepsilon}{25n}.$$

Hence we can obtain

$$\begin{aligned} & \Pr_{\mathbf{y} \sim S_{n-1}} \left[\langle \mathbf{y}, \tilde{\mathbf{g}} \rangle \leq \frac{1}{\sqrt{n-1}} \cdot \frac{\varepsilon \langle \mathbf{v}, \mathbf{g} \rangle}{\sqrt{1-\varepsilon^2}} - \frac{\varepsilon}{25n} \right] \\ & \leq \Pr_{\mathbf{y} \sim S_{n-1}} [\text{Algorithm 1 outputs “}\langle \nabla f, \alpha_{\mathbf{y}}/\|\alpha_{\mathbf{y}}\| \rangle \leq \Delta\text{” in Line 5}] \\ & \leq \Pr_{\mathbf{y} \sim S_{n-1}} \left[\langle \mathbf{y}, \tilde{\mathbf{g}} \rangle \leq \frac{1}{\sqrt{n-1}} \cdot \frac{\varepsilon \langle \mathbf{v}, \mathbf{g} \rangle}{\sqrt{1-\varepsilon^2}} + \frac{\varepsilon}{25n} \right] \end{aligned}$$

By Lemma 1, if $\|\mathbf{u}\| \leq \varepsilon$, which implies $\|\tilde{\mathbf{g}}\| \leq \varepsilon$ and (5), we can get

$$\begin{aligned} \mathbb{E}[N/T] & \geq \Pr_{\mathbf{y} \sim S_{n-1}} \left[\langle \mathbf{y}, \tilde{\mathbf{g}} \rangle \leq \frac{1}{\sqrt{n-1}} \cdot \frac{\varepsilon \langle \mathbf{v}, \mathbf{g} \rangle}{\sqrt{1-\varepsilon^2}} - \frac{\varepsilon}{25n} \right] \\ & \geq \Pr_{\mathbf{y} \sim S_{n-1}} \left[\langle \mathbf{y}, \tilde{\mathbf{g}} \rangle \leq \frac{\|\tilde{\mathbf{g}}\|}{\sqrt{n-1}} - \frac{\|\tilde{\mathbf{g}}\|}{25(n-1)} \right] \geq \frac{1}{2} + \frac{1}{2} \cdot \frac{3}{5} = \frac{4}{5}. \end{aligned}$$

By Hoeffding's inequality,

$$\Pr \left[\frac{4}{5} - \frac{N}{T} \geq \frac{1}{80} \right] \leq e^{-2T/80^2} \leq \delta = 1/3,$$

which means, from [Line 8](#), we at least have the probability of $2/3$ to get the right answer Yes. Otherwise, if $\|\mathbf{u}\| \geq 2\varepsilon$, which implies $\|\tilde{\mathbf{g}}\| \geq \sqrt{2}\varepsilon$ and [\(7\)](#), we have

$$\begin{aligned}\mathbb{E}[N/T] &\leq \Pr_{\mathbf{y} \sim S_{n-1}} \left[\langle \mathbf{y}, \tilde{\mathbf{g}} \rangle \leq \frac{1}{\sqrt{n-1}} \cdot \frac{\varepsilon \langle \mathbf{v}, \mathbf{g} \rangle}{\sqrt{1-\varepsilon^2}} + \frac{\varepsilon}{25n} \right] \\ &\leq \Pr_{\mathbf{y} \sim S_{n-1}} \left[\langle \mathbf{y}, \tilde{\mathbf{g}} \rangle \leq \frac{\|\tilde{\mathbf{g}}\|}{\sqrt{2(n-1)}} + \frac{\|\tilde{\mathbf{g}}\|}{25(n-1)} \right] \leq \frac{1}{2} + \frac{1}{2} \cdot \frac{11}{20} = \frac{31}{40}.\end{aligned}$$

Similarly by Hoeffding's inequality, we have

$$\Pr \left[\frac{N}{T} - \frac{31}{40} \geq \frac{1}{80} \right] \leq e^{-2T/80^2} \leq \delta = 1/3,$$

which means, from [Line 10](#), we at least have the probability of $2/3$ to get the right answer No. $\square$

Finally, we note that in the deterministic setting, gradient testing has a tight query complexity of $\Theta(n)$:

Theorem 2. *The classical deterministic query complexity for solving [Problem 1](#) with queries to a comparison oracle O_f^{Comp} [\(1\)](#) is $\Theta(n)$.*

Specifically, given the gradient $\mathbf{g}$ and target direction $\mathbf{v} = \mathbf{e}_1$, our algorithm (details given in [Appendix B.1](#)) estimates the ratio $g_i/g_1 = \langle \mathbf{g}, \mathbf{e}_i \rangle / \langle \mathbf{g}, \mathbf{v} \rangle$ for the remaining $n-1$ orthogonal directions $\mathbf{e}_i$ within a constant multiplicative factor. This is achieved via multiplicative search for a parameter β such that $\mathbf{g}$ is nearly orthogonal to $\beta\mathbf{e}_1 - \mathbf{e}_i$, ensuring that the total query complexity across all $n-1$ directions remains $O(n)$. Then we distinguish the two instances by verifying whether $\sum_{i=2}^n |g_i/g_1|^2$ satisfies the geometric constraint of the YES case. The $\Omega(n)$ lower bound proof (details given in [Appendix B.2](#)) builds a hard distribution where the unknown gradient is either exactly along a target direction or is slightly “tilted” toward one randomly chosen orthogonal direction. Any single query almost never detects this tilt, and gradient testing has low success probability unless $\Omega(n)$ queries are made.

3.3 Gradient Estimation

Here we present our algorithm for gradient estimation ([Problem 2](#)). As a warm-up, we first present an algorithm that finds a constant approximation of the normalized gradient.

Algorithm 3: Classical Gradient Estimation with Constant Precision

Input: lower bound γ on $\|\nabla f\|$

Output: normalized vector $\mathbf{u}$ such that $\langle \mathbf{u}, \mathbf{g} \rangle \geq 1/10$

- 1 Set $\Delta_1 = \gamma/n$
 - 2 Let U be drawn from the Haar measure on $G(n)$, where $G(n)$ denotes the orthogonal group in $\mathbb{R}^n$. Initialize orthonormal frame $(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n) = (U\mathbf{e}_1, U\mathbf{e}_2, \dots, U\mathbf{e}_n)$
 - 3 **for** $i = 1$ to n **do**
 - 4 Call [Algorithm 1](#) with input $(\mathbf{x}, \mathbf{v}_i, \Delta_1)$.
 - 5 **if** $\langle \mathbf{v}_i, \nabla f \rangle < \Delta_1$ **then**
 - 6 $\mathbf{v}_i \leftarrow -\mathbf{v}_i$
 - 7 **return** $\mathbf{u} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{v}_i$
-

Proposition 2. *Algorithm 3* outputs a normalized vector $\mathbf{u}$ satisfying $\langle \mathbf{u}, \mathbf{g} \rangle \geq 1/10$ using $O(n)$ queries to a comparison oracle O_f^{Comp} (1) with success probability at least $2/3$.

Proof. In Line 6, if $\langle \mathbf{v}_i, \mathbf{g} \rangle < \Delta_1/\gamma$, we flip $\mathbf{v}_i$. Hence, when the algorithm terminates we have

$$\langle \mathbf{v}_i, \mathbf{g} \rangle \geq -\Delta_1/\gamma \geq -1/n. \quad (8)$$

Under (8), we can promise that the vector $\mathbf{u}$ in Line 7 satisfies $\mathbb{E}(\langle \mathbf{u}_1, \mathbf{g} \rangle) \geq 0.7$ (this is formally stated as Lemma 4 with full proofs deferred to Appendix A.2). Define $p = \Pr[\langle \mathbf{u}_1, \mathbf{g} \rangle \geq 1/10]$. Because of $\langle \mathbf{u}_1, \mathbf{g} \rangle \leq 1$, we can get $p \geq 2/3$ by Markov's inequality. $\square$

Building upon Algorithm 3, we obtain our Algorithm 4 that solves Problem 2 to any given precision ε .

Algorithm 4: Classical Gradient Estimation with Arbitrary Given Precision

Input: accuracy ε , lower bound γ on $\|\nabla f\|$

- 1 Set $\Delta_1 = \gamma/n$, $\Delta_2 = \varepsilon\gamma/400\sqrt{n}$ and define $w_i(\beta) = \beta\mathbf{e}_1 - \mathbf{e}_i$
- 2 Call Algorithm 3 and get an estimation $\mathbf{u}$ satisfying $\langle \mathbf{u}, \mathbf{g} \rangle \geq 1/10$
- 3 Apply an orthogonal change of basis so that $\mathbf{u}$ is mapped to $\mathbf{e}_1$.
- 4 For any $i = 2, 3, \dots, n$, initialize all $\ell_i = 1$, $t_i = 1$.
- 5 **for** $i = 2$ to n **do**
- 6 Call Algorithm 1 with input $(\mathbf{x}, \mathbf{e}_i, \Delta_1)$
- 7 **if** $\langle \nabla f, \mathbf{e}_i \rangle < \Delta_1$ **then**
- 8 $t_i \leftarrow -t_i$
- 9 Let $D = \text{diag}(1, t_2, \dots, t_n)$ and apply the orthogonal transformation D to the coordinate frame.
- 10 **for** $i = 2$ to n **do**
- 11 **repeat**
- 12 $\beta \leftarrow \ell_i/\sqrt{n}$
- 13 Call Algorithm 1 with input $(\mathbf{x}, w_i(\beta)/\|w_i(\beta)\|, \Delta_2)$
- 14 **if** $\langle \nabla f, w_i(\beta)/\|w_i(\beta)\| \rangle \leq \Delta_2$ **then**
- 15 $\ell_i \leftarrow 2\ell_i$
- 16 **until** $\langle \nabla f, w_i(\beta)/\|w_i(\beta)\| \rangle > -\Delta_2$
- 17 $\alpha_{i,1} \leftarrow -\ell_i/\sqrt{n}$, $\alpha_{i,2} \leftarrow \ell_i/\sqrt{n}$
- 18 **repeat**
- 19 $\alpha_i \leftarrow (\alpha_{i,1} + \alpha_{i,2})/2$
- 20 Call Algorithm 1 with input $(\mathbf{x}, w_i(\alpha_i)/\|w_i(\alpha_i)\|, \Delta_2)$
- 21 **if** $\langle \nabla f, w_i(\alpha_i)/\|w_i(\alpha_i)\| \rangle \leq \Delta_2$ **then**
- 22 $\alpha_{i,1} \leftarrow \alpha_i$
- 23 **else**
- 24 $\alpha_{i,2} \leftarrow \alpha_i$
- 25 **until** $\alpha_{i,2} - \alpha_{i,1} < \varepsilon/4\sqrt{n}$
- 26 **return** $\mathbf{h} = (\sum_i \alpha_i t_i) / \sqrt{\sum_i \alpha_i^2}$ where $\alpha_1 = 1$

The intuition of Algorithm 4 can be summarized as follows. To estimate the gradient direction $\mathbf{g}$ within error ε , a natural approach is to estimate each coordinate to relative accuracy $\varepsilon/\sqrt{n}$.

In particular, we call [Algorithm 1](#) with input direction $\alpha_j \mathbf{e}_i - \alpha_i \mathbf{e}_j$ to roughly compare between α_i/α_j and g_i/g_j . Assuming without loss of generality that g_1 has the largest magnitude among all coordinates, we can then perform a binary search over $[-1, 1]$ to estimate each ratio g_i/g_1 to accuracy $\varepsilon/\sqrt{n}$.

However, the approach above requires $\Theta(n \log(n/\varepsilon))$ queries. To improve upon this, we adopt an idea similar to that used in [Algorithm 6](#), which fully exploits the upper bound on $\sum_{i=1}^n |g_i/g_1|^2$: the quantity $|g_i/g_1|$ cannot be large for all coordinates simultaneously. This is achieved by identifying an estimation $\mathbf{u}$ such that $\mathbf{u}$ and $\mathbf{g}$ has a constant overlap using [Algorithm 3](#). Then, we rotate the frame so that $\mathbf{u}$ become the first ordinate, which ensures that the upper bound on $\sum_{i=1}^n |g_i/g_1|^2$ is not too large. Then, we compute an upper bound on each quantity $\sqrt{n} \cdot |g_i/g_1|$, denoted by ℓ_i , which is obtained in Lines 11–16. For each coordinate i , we perform a binary search over the interval $[-\ell_i/\sqrt{n}, \ell_i/\sqrt{n}]$ to estimate g_i/g_1 . Since $\sum_{i=1}^n \ell_i^2/n = O(1)$, the number of iterations in our binary search $\sum_{i=1}^n \log(\ell_i/\varepsilon)$ is at most $O(n \log(1/\varepsilon))$.

We note that the random frame does not need to be fully Haar distributed. The proof only uses the constant moment guarantee summarized in [Lemma 4](#); hence an orthogonal 2-design, or any sufficiently accurate approximation satisfying the same guarantee, can in principle replace the Haar-random frame.

Theorem 3. *[Algorithm 4](#) solves [Problem 2](#) using $O(n \log(1/\varepsilon))$ queries to a comparison oracle O_f^{Comp} with success probability at least $2/3$. Moreover, we can increase success probability to at least $1 - \eta$ using $O(n \log(1/\varepsilon) \log(1/\eta))$ queries for any $\eta \in (0, 1)$.*

Proof. By [Proposition 2](#), in the rest of the proof, we consider the case that the event

$$g_1 = \langle \mathbf{u}, \mathbf{g} \rangle \geq 1/10 \tag{9}$$

holds. In Lines 11–16, we call [Algorithm 1](#) with input direction $w_i(\ell_i/\sqrt{n})$, which can judge the sign of $\ell_i/\sqrt{n} - g_i/g_1$ (similar to (10)) within error Δ_2 . If the sign is negative, we multiply ℓ_i by 2. As a result, when the loop terminates we have

$$\begin{aligned} \langle \nabla f, w_i(\ell_i/\sqrt{n}) \rangle &\geq -\Delta_2, \\ \frac{\ell_i}{\sqrt{n}} g_1 - g_i &\geq -\Delta_2 \sqrt{1 + \ell_i^2/n/\gamma} \geq -\frac{10\Delta_2}{\gamma}, \end{aligned}$$

and similarly

$$\begin{aligned} \langle \nabla f, w_i(\ell_i/2\sqrt{n}) \rangle &\leq \Delta_2, \\ \frac{\ell_i}{2\sqrt{n}} g_1 - g_i &\leq \Delta_2 \sqrt{1 + \ell_i^2/4n/\gamma} \leq \frac{10\Delta_2}{\gamma}. \end{aligned}$$

Using the value we set for Δ_2 , the error from [Algorithm 1](#) satisfies

$$\frac{10\Delta_2}{\gamma \cdot g_1} \leq \frac{100\Delta_2}{\gamma} \leq \frac{\varepsilon}{4\sqrt{n}},$$

hence we have

$$|g_i/g_1| - \frac{\varepsilon}{4\sqrt{n}} \leq \ell_i/\sqrt{n} \leq 2|g_i/g_1| + \frac{\varepsilon}{2\sqrt{n}}.$$

As a result, in Lines 11–16, we are promised to get the smallest ℓ_i which is a power of 2 and satisfies $\ell_i/\sqrt{n} > g_i/g_1$ up to a small error above. With (9) holding, for $\|\mathbf{g}\| = 1$, we have $\sum_{i=2}^n g_i^2/g_1^2 = O(1)$. Consequently, to find all ℓ_i , the number of queries we need is

$$\sum_{i=2}^n 2(1 + \log \ell_i) \leq \sum_{i=2}^n 2\ell_i \leq \sum_{i=2}^n 2\ell_i^2 = O(n).$$

In Lines 18–25, our binary search needs $O(\log(\ell_i/\varepsilon))$ iterations to reach accuracy $\varepsilon/4\sqrt{n}$. If we both have

$$\langle \nabla f, w_i(\alpha_{i,1}) \rangle \leq \Delta_2 \quad \langle \nabla f, w_i(\alpha_{i,2}) \rangle \geq -\Delta_2,$$

we will then call Algorithm 1 with input $(y, w_i((\alpha_{i,1} + \alpha_{i,2})/2), \Delta_2)$, which gives

$$\left| \alpha_i - \frac{g_i}{g_1} \right| \leq \varepsilon/(4\sqrt{n}) + \frac{\Delta_2 \sqrt{1 + \ell_i^2/n}}{\gamma \cdot g_1} \leq \varepsilon/(2\sqrt{n}),$$

and

$$\|\mathbf{h} - \mathbf{g}\|^2 \leq 2|g_1|^2 \sum_{i=1}^n (\alpha_i - |g_i/g_1|)^2 \leq \varepsilon^2/2 < \varepsilon^2.$$

During binary search, we use

$$\sum_i \log \left(\frac{4\ell_i}{\varepsilon} \right) \leq n \log \frac{1}{\varepsilon} + \sum_i \log(4\ell_i) = O \left(n \log \frac{1}{\varepsilon} \right)$$

queries. As a result, we use $O(n + n \log(1/\varepsilon)) = O(n \log(1/\varepsilon))$ queries with success probability at least $2/3$ (following that of Algorithm 3).

Finally, we show how to boost the success probability to at least $1 - \eta$ with an overhead of $O(\log \frac{1}{\eta})$ in query complexity. Consider the $m = \lceil 18 \ln(1/\eta) \rceil$ independent base runs in the algorithm. Define $\mathbf{h}^{(r)}$ to be the output of the r -th iteration and set $\mathcal{N}_r = \{s \in \{1, \dots, m\} : \|\mathbf{h}^{(r)} - \mathbf{h}^{(s)}\| \leq 2\varepsilon/3\}$. We finally return any $\mathbf{h}^{(r)}$ with $|\mathcal{N}_r| \geq m/2$ and prove that it satisfies our requirements with success probability at least $1 - \eta$.

Let X_r be the indicator that the r -th candidate satisfies $\|\mathbf{h}^{(r)} - \mathbf{g}\| \leq \varepsilon_0$. Since $\Pr[X_r = 1] \geq 2/3$, Hoeffding's inequality gives

$$\Pr \left[\sum_{r=1}^m X_r \leq m/2 \right] \leq \exp(-m/18) \leq \eta.$$

Conditioned on the complementary event, more than half of the candidates are good. Any two good candidates are within $2\varepsilon_0 = 2\varepsilon/3$ of each other, so every good candidate satisfies the selection condition $|\mathcal{N}_r| \geq m/2$. Thus the algorithm can select at least one candidate. Moreover, for any selected candidate $\mathbf{h}^{(r)}$, the set $\mathcal{N}_r$ intersects the set of good candidates because both sets have size greater than $m/2$ or at least $m/2$ with a strict majority of good candidates. Therefore, for some good $\mathbf{h}^{(s)}$,

$$\|\mathbf{h}^{(r)} - \mathbf{g}\| \leq \|\mathbf{h}^{(r)} - \mathbf{h}^{(s)}\| + \|\mathbf{h}^{(s)} - \mathbf{g}\| \leq 2\varepsilon/3 + \varepsilon/3 = \varepsilon.$$

The total query complexity is $m \cdot O(n \log \frac{1}{\varepsilon}) = O(n \log \frac{1}{\varepsilon} \log \frac{1}{\eta})$. $\square$

4 Quantum Algorithm for Gradient Estimation

In this section, we present our quantum algorithm for gradient estimation.

As introduced in Section 2.3, Ref. [29] assumed quantum query access to $\langle \mathbf{y}, \nabla f \rangle$ for each $\mathbf{y}$ and then use QFT to estimate ∇f . However, we cannot directly estimate $\langle \mathbf{y}, \mathbf{g} \rangle$ in our comparison setting. Instead, we choose a direction $\mathbf{e}_1$ and approximate $\langle \mathbf{y}, \mathbf{g} \rangle / \langle \mathbf{e}_1, \mathbf{g} \rangle$. To achieve this, we apply Algorithm 1 with input direction $\mathbf{v} = (k, \tilde{\mathbf{y}})$, where we assume $\mathbf{y} = (y_1, \tilde{\mathbf{y}})$ and apply binary search on k such that $\mathbf{v}$ is roughly orthogonal to $\mathbf{g}$. This indicates that the overlap between $\tilde{\mathbf{y}}$ and $\mathbf{g}$ is about $-k \langle \mathbf{g}, \mathbf{e}_1 \rangle$. Then, we add y_1 , obtain $(y_1 - k) \langle \mathbf{e}_1, \mathbf{g} \rangle$ as $\langle \mathbf{y}, \mathbf{g} \rangle$, and kick this back to the phase. Finally, we apply QFT to get the estimation.

Algorithm 5: Quantum Gradient Estimation

Input: $\mathbf{x} \in \mathbb{R}^d$, accuracy ε , lower bound γ on $\|\nabla f\|$

- 1 Set $\Delta = \gamma\varepsilon^2/(48\pi n^3)$ and the quantum transformation unitary U such that
$$U|\mathbf{y}\rangle = \frac{1}{\sqrt{N}} \sum_{\mathbf{z} \in M} \mathbf{e}^{-2\pi i \langle \mathbf{z}, \mathbf{y} \rangle t} |\mathbf{z}\rangle$$
- 2 Randomly sample a unit vector $\mathbf{v}$, and rotate the frame so that $\mathbf{v}$ becomes the first basis vector $\mathbf{e}_1$
- 3 Initialize $|\psi\rangle = \frac{1}{\sqrt{N}} \sum_{\mathbf{y} \in M} |\mathbf{y}\rangle |\phi\rangle$, where $M = \{\mathbf{y} \mid y_i = 0, \frac{1}{t}, \frac{2}{t}, \dots, 1\}$, $t = \lceil 10n^2/\varepsilon \rceil$, and
$$|\phi\rangle = \frac{1}{\sqrt{t^2+1}} \sum_{j=0}^{t^2} \mathbf{e}^{-2\pi i j/t} |j\rangle$$
- 4 Set $k_1 = -5n$, and $k_2 = 5n$
- 5 **repeat**
 - 6 Set $k = (k_1 + k_2)/2$
 - 7 For each $\mathbf{y}$, call **Algorithm 1** with input $(\mathbf{x}, (k, \tilde{\mathbf{y}})/\|(k, \tilde{\mathbf{y}})\|, \Delta)$
 - 8 **if** $\langle \nabla f, (k, \tilde{\mathbf{y}})/\|(k, \tilde{\mathbf{y}})\| \rangle < \Delta$ **then**
 - 9 Set $k_1 = k$
 - 10 **else**
 - 11 Set $k_2 = k$
- 12 **until** $k_2 - k_1 < \varepsilon^2/(8\pi n^{1.5})$
- 13 Set $h(\mathbf{y}) = y_1 - k$, and output $\lfloor h(\mathbf{y}) \cdot t^2/\sqrt{n} \rfloor$ on the second register
- 14 Apply the quantum transformation U on our first register.
- 15 **return** the measurement result of the first register in the computational basis.

Theorem 4. *Algorithm 5 solves Problem 2 using $O(\log(n/\varepsilon))$ queries to a quantum comparison oracle $O_{f,Q}^{\text{Comp}}$ (2) with success probability at least $8/15 - 2\varepsilon$. Moreover, we can increase success probability to $1 - \eta$ using $O(\log \frac{n}{\varepsilon} \log \frac{1}{\eta})$ queries for any $\eta \in (0, 1)$.*

Proof. By Lemma 1, we have

$$\Pr_{\mathbf{v} \sim S_n} [|\langle \mathbf{v}, \mathbf{g} \rangle| \geq 1/(5\sqrt{n})] \geq 4/5.$$

Therefore, without loss of generality, we can assume $|\langle \mathbf{v}, \mathbf{g} \rangle| \geq 1/(5\sqrt{n})$, which means for any $\mathbf{y} \in M$,

$$|\langle \mathbf{g}, \mathbf{y} \rangle| \leq \sqrt{n} \leq 5n |\langle \mathbf{g}, \mathbf{v} \rangle|,$$

indicating that there exists $k \in [-n, n]$ such that

$$\langle \mathbf{g}, (k, \tilde{\mathbf{y}}) \rangle = 0.$$

During our binary search in Lines 5–12, for each $\mathbf{y}$, define $\alpha_{\mathbf{y}}(k) = (k, \tilde{\mathbf{y}})$. If we already obtain

$$\begin{aligned} \langle \nabla f, \alpha_{\mathbf{y}}(k_1)/\|\alpha_{\mathbf{y}}(k_1)\| \rangle &\leq \Delta \\ \langle \nabla f, \alpha_{\mathbf{y}}(k_2)/\|\alpha_{\mathbf{y}}(k_2)\| \rangle &\geq -\Delta, \end{aligned}$$

we will then call **Algorithm 1** with input $(\mathbf{x}, \alpha_{\mathbf{y}}((k_1 + k_2)/2)/\|\alpha_{\mathbf{y}}((k_1 + k_2)/2)\|, \Delta)$. Finally, after $O(\log(n/\varepsilon))$ iterations, we get

$$\begin{aligned} \langle \nabla f, (k, \tilde{\mathbf{y}}) \rangle &\leq \|(k, \tilde{\mathbf{y}})\| \Delta \leq (5n + n) \Delta = 6n \Delta, \\ \langle \nabla f, (k + \varepsilon^2/(8\pi n^{1.5}), \tilde{\mathbf{y}}) \rangle &\geq -6n \Delta, \end{aligned}$$

which also means

$$\begin{aligned}\langle \mathbf{g}, \mathbf{y} \rangle &\leq (y_1 - k) \langle \mathbf{g}, \mathbf{v} \rangle + 6n\Delta/\gamma, \\ \langle \mathbf{g}, \mathbf{y} \rangle &\geq (y_1 - k - \varepsilon^2/(8\pi n^{1.5})) \langle \mathbf{g}, \mathbf{v} \rangle - 6n\Delta/\gamma.\end{aligned}$$

And then we have

$$\left| h(\mathbf{y}) - \frac{\langle \mathbf{g}, \mathbf{y} \rangle}{\langle \mathbf{g}, \mathbf{v} \rangle} \right| \leq \varepsilon^2/(8\pi n^{1.5}) + \frac{6n\Delta}{\gamma \langle \mathbf{g}, \mathbf{v} \rangle} \leq \varepsilon^2/(4\pi n^{1.5}),$$

which means

$$\left| e^{2\pi i h(\mathbf{y}) \cdot \frac{t}{5\sqrt{n}}} - e^{2\pi i \langle \frac{\mathbf{g}}{\langle \mathbf{g}, \mathbf{v} \rangle}, \mathbf{y} \rangle \cdot \frac{t}{5\sqrt{n}}} \right| \leq |e^{i\varepsilon} - 1| \leq \varepsilon.$$

In [Line 13](#), we can get the quantum state

$$|\psi\rangle = \frac{1}{\sqrt{N}} \sum_{\mathbf{y} \in M} e^{2\pi i h(\mathbf{y}) \cdot \frac{t}{5\sqrt{n}}} |\mathbf{y}\rangle |\phi\rangle.$$

As a result, we get

$$\left\| |\psi\rangle - \frac{1}{\sqrt{N}} \sum_{\mathbf{y} \in M} e^{2\pi i \langle \frac{\mathbf{g}}{\langle \mathbf{g}, \mathbf{v} \rangle}, \mathbf{y} \rangle \cdot \frac{t}{5\sqrt{n}}} |\mathbf{y}\rangle \right\| \leq \varepsilon.$$

By [Proposition 3](#), we can output the result $\mathbf{z}$ after the transformation in [Line 14](#) such that

$$\left\| \mathbf{z} - \frac{\mathbf{g}}{5\sqrt{n} \cdot \langle \mathbf{g}, \mathbf{v} \rangle} \right\| \leq 2n^{1.5}/t \leq \varepsilon/(10\sqrt{n})$$

with probability at least $2/3 - 2\varepsilon$. This further implies

$$\left\| \frac{\mathbf{z}}{\|\mathbf{z}\|} - \mathbf{g} \right\| \leq 2 \sin \theta \leq 10\sqrt{n} \left\| \mathbf{z} - \frac{\mathbf{g}}{5\sqrt{n} \cdot \langle \mathbf{g}, \mathbf{v} \rangle} \right\| \leq \varepsilon,$$

where $\theta = \arg(\frac{\mathbf{z}}{\|\mathbf{z}\|}, \mathbf{g})$. In all, we can succeed with probability at least $\frac{4}{5} \cdot (\frac{2}{3} - 2\varepsilon) \geq \frac{8}{15} - 2\varepsilon$.

In addition, the success probability can also be boosted to $1 - \eta$ with any $\eta \in (0, 1)$ similar to the proof of [Theorem 3](#). $\square$

5 Numerical Experiments

We complement our theoretical results with numerical experiments that (i) validate the empirical query complexity of the proposed algorithms, and (ii) demonstrate that plugging our gradient direction estimator into a NGD optimizer yields convergence comparable to access to the exact normalized gradient, and better than existing comparison based baselines under matched per-iteration query budgets. All experiments are run on an Apple MacBook Air (M2, 2022) with 8-core Apple M2 chip and 16 GB memory.

Setup. We use three test functions: a general strongly-convex quadratic $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top U^\top \Lambda U \mathbf{x}$ with U Haar-orthogonal and Λ a diagonal matrix whose eigenvalues are evenly spread over $[1, 10]$; a sparse quadratic function in which the first $s = 10$ coordinates carry weights evenly spread over $[1, 10]$ and the remaining coordinates carry a vanishing weight; and the extended Rosenbrock function $f(\mathbf{x}) = \sum_{i=1}^{n-1} [100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2]$.

Table 2: Estimation accuracy and average query count for [Algorithm 4](#) across dimensions, $\varepsilon = 0.2$, 100 trials per cell. Success = fraction of trials with ℓ_2 error below ε .

Function	n	Success	Avg. error	Avg. queries
Gen. Quadratic	10	1.00	0.0068	85.3
	50	1.00	0.0071	461.0
	100	1.00	0.0071	929.6
Sparse Quadratic	10	1.00	0.0068	86.4
	50	1.00	0.0071	458.5
	100	1.00	0.0071	922.3
Ext. Rosenbrock	10	1.00	0.0068	85.4
	50	1.00	0.0072	459.8
	100	1.00	0.0072	925.7

Table 3: Estimation accuracy and average query count for [Algorithm 4](#) on the extended Rosenbrock function at $n = 100$ as the target precision ε varies, 100 trials per cell.

ε	Success	Avg. error	Max. error	Avg. queries
0.20	1.00	0.0072	0.0081	925.7
0.10	1.00	0.0036	0.0041	1024.9
0.05	1.00	0.0018	0.0020	1124.5
0.01	1.00	0.00045	0.00052	1320.7

Correctness of our algorithms. We evaluate the three classical algorithms developed in this paper: the randomized $O(1)$ gradient testing algorithm ([Algorithm 2](#)), the deterministic $O(n)$ gradient testing algorithm ([Algorithm 6](#)), and the gradient estimation algorithm ([Algorithm 4](#)). For testing, we construct an extreme YES instance at distance 0.95ε from $\nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$ and an extreme NO instance at distance 2.05ε , and run 100 independent trials for each (n, ε) configuration with $n \in \{10, 50, 100\}$ and $\varepsilon = 0.2$. Both testing algorithms attain an empirical success probability of 100% on every configuration, far exceeding the constant probability guarantee in our theoretical analysis. For estimation, [Table 2](#) reports the dependence on dimension at $\varepsilon = 0.2$ across all three test functions, and [Table 3](#) reports the dependence on the target precision ε on the extended Rosenbrock function at $n = 100$. The estimator returns a unit vector whose error is below ε in nearly 100% of trials, with average error well below the target precision. This confirms that the worst-case query bound is not tight in typical instances.

Empirical query complexity. [Theorem 3](#) predicts $Q = O(n \log(1/\varepsilon))$ comparison queries for our estimation algorithm. We verify this by sweeping a 10×10 grid of (n, ε) values on the extended Rosenbrock function with $\gamma = 0.05$ and 100 trials per cell. [Figure 1a](#) shows the joint dependence of the average query count on n and $\ln(1/\varepsilon)$, following the bilinear model

$$Q \approx w_1 n \ln(1/\varepsilon) + w_2 n + w_3 \ln(1/\varepsilon) + b.$$

Fitting this model jointly across the grid gives $w_1 \approx 1.38$, $w_2 \approx 7.06$, $w_3 \approx -1.41$, $b \approx -5.28$ with $R^2 = 0.999$, in tight agreement with the theoretical scaling.

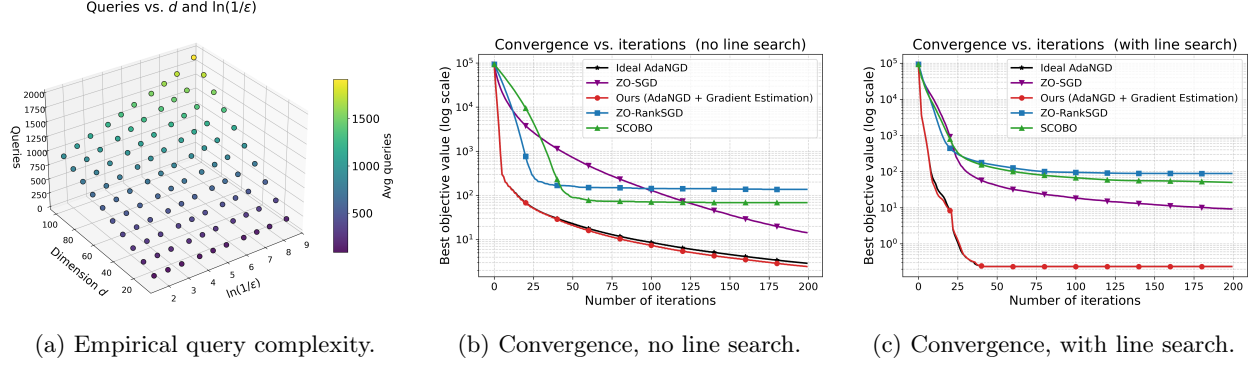


Figure 1: (a) Average comparison queries used by Algorithm 4 on the extended Rosenbrock function over a 10×10 grid of $(n, \ln(1/\epsilon))$ values; the fitted bilinear surface is $Q \approx 1.38 n \ln(1/\epsilon) + 7.06 n - 1.41 \ln(1/\epsilon) - 5.28$ with $R^2 = 0.999$. (b), (c) Convergence on the extended Rosenbrock function in $n = 100$, averaged over 10 random initializations, without and with line search respectively. Under per-iteration comparison budgets of the same order, ADANGD+OURS tracks the noise-free IDEAL ADANGD curve, while the baselines stall orders of magnitude above the optimum.

Optimization with the estimated direction. We feed the direction returned by Algorithm 4 into adaptive NGD [35] with step size $\eta_t = R/\sqrt{2t}$, $R = 5$, giving ADANGD+OURS. Baselines on the extended Rosenbrock function in $n = 100$ (averaged over 10 random initializations) are: IDEAL ADANGD, the noise-free reference using the true normalized gradient; ZO-SGD [39], a two-point Gaussian smoothing estimator with $m = 100$ probes; ZO-RANKSGD [48], an (m, k) ranking oracle with $m = k = 100$; SCOBO [10], $m = 100$ pairwise sign comparisons per iteration. We take $\epsilon = 0.2$ for our estimator, where the cost is of the same order as the baselines. We also report a line-search variant ($\ell = 5$ shrinking trials, factor 0.1). The results are shown in Figure 1b and Figure 1c.

After 200 iterations ADANGD+OURS reaches objective 0.23 with line search and 2.6 without, essentially matching IDEAL ADANGD (0.23 and 3.1); while the other three baselines still remain over 10. At a comparable per-iteration query budget, our estimator returns a direction whose angular deviation is small enough that adaptive NGD behaves as if it had exact gradient direction, whereas the other baselines spend the same budget on a much coarser direction.

6 Lower Bounds on Gradient Estimation

We also establish lower bounds on gradient estimation, both in the classical and quantum settings.

Theorem 5. Any classical algorithm that solves Problem 2 with success probability $2/3$ takes $\Omega(n \log(1/\epsilon))$ queries to a comparison oracle O_f^{Comp} (1).

The proof is built upon an ϵ -net argument on the sphere, and a gradient estimator must effectively identify which one is present. Each comparison query yields only a constant number of outcomes, conveys limited information, implying that the number of queries should scale linearly with dimension. The detailed proof can be found in Appendix D.1.

Theorem 6. Any quantum algorithm that solves Problem 2 with success probability $2/3$ takes $\Omega(\log(1/\epsilon))$ queries to a quantum comparison oracle $O_{f,Q}^{\text{Comp}}$ (2).

This quantum lower bound argument applies the hybrid argument [5], where we consider a hard instance with binary partial derivatives and prove that we need enough number of quantum queries to estimate the gradient. The detailed proof can be found in Appendix D.2.

7 Conclusions

In this paper, we study gradient testing and gradient estimation using comparison queries. For any smooth $f: \mathbb{R}^n \rightarrow \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$, and $\varepsilon > 0$, our classical gradient testing algorithm determines whether $\nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$ is ε -close or 2ε -far from a given unit vector $\mathbf{v}$ using $O(n)$ queries, as well as a gradient estimation algorithm that outputs an ε -estimate of $\nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$ using $O(n \log(1/\varepsilon))$ queries. Both bounds are shown to be optimal. Furthermore, we study these problems in the quantum setting, and develop quantum algorithms for gradient testing and gradient estimation using $O(1)$ and $O(\log(n/\varepsilon))$ queries, respectively.

Technically, after an appropriate rotation, there exists a dominant coordinate whose inner product with the gradient captures a constant fraction of the gradient magnitude. We isolate this coordinate and express inner product as a scale of this constant fraction, and also optimized the application of binary search in other coordinates to achieve optimality. In terms of quantum algorithms, we leverage the quantum Fourier transform (QFT) to coherently encode directional derivatives across coordinates into relative phases of a single quantum state. This phase encoding enables the simultaneous extraction of relative inner products via inverse QFT.

Our work raises several open directions for future research. First, our quantum algorithm for gradient estimation is not yet optimal, and we conjecture that there exists an algorithm matching our lower bound. Second, it is also natural to study the gradient testing and estimation problems with a stochastic comparison oracle. Third, it would be valuable to investigate how the proposed gradient testing and estimation procedures can be incorporated into optimization solvers with improved performance.

Acknowledgments

We thank Xingyu Zhao for helpful discussions at the early stage of this paper. We thank the anonymous reviewers for their constructive feedback. X.T., H.W., Y.Z., and T.L. were supported by the National Natural Science Foundation of China (Grant Numbers 62372006 and 92365117).

References

- [1] Joran van Apeldoorn and András Gilyén. Improvements in quantum sdp-solving with applications. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019. arXiv:[1804.05058](#)
- [2] Joran van Apeldoorn, András Gilyén, Sander Gribling, and Ronald de Wolf. Convex optimization using quantum oracles. *Quantum*, 4:220, 2020. arXiv:[1809.00643](#)
- [3] Joran van Apeldoorn, András Gilyén, Sander Gribling, and Ronald de Wolf. Quantum SDP-solvers: Better upper and lower bounds. *Quantum*, 4:230, 2020. arXiv:[1705.01843](#)
- [4] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pages 1–42, 2022. arXiv:[1809.06474](#)
- [5] Charles H. Bennett, Ethan Bernstein, Gilles Brassard, and Umesh Vazirani. Strengths and weaknesses of quantum computing. *SIAM Journal on Computing*, 26(5):1510–1523, 1997. arXiv:[quant-ph/9701001](#)

- [6] El Houcine Bergou, Eduard Gorbunov, and Peter Richtárik. Stochastic three points method for unconstrained smooth minimization. *SIAM Journal on Optimization*, 30(4):2726–2749, 2020. arXiv:[1902.03591](#)
- [7] Adam Bouland, Yosheb M Getachew, Yujia Jin, Aaron Sidford, and Kevin Tian. Quantum speedups for zero-sum games via improved dynamic Gibbs sampling. In *International Conference on Machine Learning*, pages 2932–2952. PMLR, 2023. arXiv:[2301.03763](#)
- [8] Fernando G. S. L. Brandão and Krysta M. Svore. Quantum speed-ups for solving semidefinite programs. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 415–426. IEEE, 2017. arXiv:[1609.05537](#)
- [9] Fernando G.S.L. Brandão, Amir Kalev, Tongyang Li, Cedric Yen-Yu Lin, Krysta M. Svore, and Xiaodi Wu. 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)*, pages 27:1–27:14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019. arXiv:[1710.02581](#)
- [10] HanQin Cai, Daniel McKenzie, Wotao Yin, and Zhenliang Zhang. A one-bit, comparison-based gradient estimator. *Applied and Computational Harmonic Analysis*, 60:242–266, 2022. arXiv:[2010.02479](#)
- [11] Pablo A. M. Casares and Miguel Angel Martin-Delgado. A quantum interior-point predictor–corrector algorithm for linear programming. *Journal of physics A: Mathematical and Theoretical*, 53(44):445305, 2020. arXiv:[1902.06749](#)
- [12] Shouvanik Chakrabarti, Andrew M. Childs, Tongyang Li, and Xiaodi Wu. Quantum algorithms and lower bounds for convex optimization. *Quantum*, 4:221, 2020. arXiv:[1809.01731](#)
- [13] Sourav Chatterjee. A note about the uniform distribution on the intersection of a simplex and a sphere. *Journal of Topology and Analysis*, 9(4):717–738, 2017. arXiv:[1011.4043](#)
- [14] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017. arXiv:[1708.03999](#)
- [15] Zherui Chen, Yuchen Lu, Hao Wang, Yizhou Liu, and Tongyang Li. Quantum Langevin dynamics for optimization. *Communications in Mathematical Physics*, 406(3):52, 2025. arXiv:[2311.15587](#)
- [16] Minhao Cheng, Simranjit Singh, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-OPT: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2019. arXiv:[1909.10773](#)
- [17] Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. In *International Conference on Machine Learning*, pages 970–978. PMLR, 2018. arXiv:[1804.02395](#)
- [18] Alexander M Dalzell, Sam McArdle, Mario Berta, Przemysław Bienias, Chi-Fang Chen, András Gilyén, Connor T. Hann, Michael J. Kastoryano, Emil T Khabiboulline, Aleksander Kubica, Grant Salton, Samson Wang, and Fernando G. S. L. Brandão. Quantum algorithms: A survey of applications and end-to-end complexities, 2023. arXiv:[2310.03011](#)

- [19] Anindya De, Ilias Diakonikolas, Vitaly Feldman, and Rocco A. Servedio. Nearly optimal solutions for the Chow parameters problem and low-weight approximation of halfspaces. *Journal of the ACM*, 61(2):11:1–11:36, 2014. arXiv:[1206.0985](#)
- [20] John E. Dennis, Jr and Virginia Torczon. Direct search methods on parallel machines. *SIAM Journal on Optimization*, 1(4):448–474, 1991.
- [21] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015. arXiv:[1312.2139](#)
- [22] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018. arXiv:[1807.01695](#)
- [23] Minbo Gao, Zhengfeng Ji, Tongyang Li, and Qisheng Wang. Logarithmic-regret quantum learning algorithms for zero-sum games. *Advances in Neural Information Processing Systems*, 36:31177–31203, 2023. arXiv:[2304.14197](#)
- [24] M. Garrido, Alexander M. Dalzell, and Sam McArdle. Quantum and classical algorithms for SOCP based on the multiplicative weights update method, 2025. arXiv:[2507.14127](#)
- [25] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for non-convex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. arXiv:[1309.5549](#)
- [26] Oded Goldreich. *Introduction to property testing*. Cambridge University Press, 2017.
- [27] Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International Conference on Machine Learning*, pages 3100–3109. PMLR, 2019. arXiv:[1910.12166](#)
- [28] Chi Jin, Lydia T. Liu, Rong Ge, and Michael I. Jordan. On the local minima of the empirical risk. *Advances in Neural Information Processing Systems*, 31, 2018. arXiv:[1803.09357](#)
- [29] Stephen P. Jordan. Fast quantum algorithm for numerical gradient estimation. *Physical Review Letters*, 95(5), July 2005.
- [30] Mustafa O. Karabag, Cyrus Neary, and Ufuk Topcu. Smooth convex optimization using sub-zeroth-order oracles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):3815–3822, 2021. arXiv:[2103.00667](#)
- [31] Iordanis Kerenidis, Anupam Prakash, and Dániel Szilágyi. Quantum algorithms for second-order cone programming and support vector machines. *Quantum*, 5:427, 2021. arXiv:[1908.06720](#)
- [32] Jeffrey Larson, Matt Menickelly, and Stefan M. Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019. arXiv:[1904.11585](#)
- [33] Jiaqi Leng, Ethan Hickman, Joseph Li, and Xiaodi Wu. Quantum Hamiltonian descent, 2023. arXiv:[2303.01471](#)
- [34] Jiaqi Leng, Kewen Wu, Xiaodi Wu, and Yufan Zheng. (sub) exponential quantum speedup for optimization, 2025. arXiv:[2504.14841](#)

- [35] Levy, Kfir. Online to offline conversions, universality and adaptive minibatch sizes. *Advances in Neural Information Processing Systems*, 30, 2017. arXiv:[1705.10499](#)
- [36] Yizhou Liu, Weijie J. Su, and Tongyang Li. On quantum speedups for nonconvex optimization via quantum tunneling walks. *Quantum*, 7:1030, 2023. arXiv:[2209.14501](#)
- [37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. arXiv:[1706.06083](#)
- [38] John A. Nelder and Roger Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.
- [39] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.
- [40] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519, 2017. arXiv:[1602.02697](#)
- [41] John Preskill. Quantum computing in the NISQ era and beyond. *Quantum*, 2:79, 2018. arXiv:[1801.00862](#)
- [42] John Preskill. Beyond NISQ: The megaquop machine. *ACM Transactions on Quantum Computing*, 6(3):1–7, 2025. arXiv:[2502.17368](#)
- [43] Dana Ron. Algorithmic and analysis techniques in property testing. *Foundations and Trends® in Theoretical Computer Science*, 5(2):73–205, 2010.
- [44] Aadirupa Saha, Tomer Koren, and Yishay Mansour. Dueling convex optimization. In *International Conference on Machine Learning*, pages 9245–9254. PMLR, 2021.
- [45] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning, 2017. arXiv:[1703.03864](#)
- [46] Aaron Sidford and Chenyi Zhang. Quantum speedups for stochastic optimization. *Advances in Neural Information Processing Systems*, 37, 2023. arXiv:[2308.01582](#)
- [47] Steven Soojin Kim and Kavita Ramanan. A conditional limit theorem for high-dimensional ℓ^p -spheres. *Journal of Applied Probability*, 55(4):1060–1077, 2018. arXiv:[1509.05442](#)
- [48] Zhiwei Tang, Dmitry Rybin, and Tsung-Hui Chang. Zeroth-order optimization meets human feedback: Provable learning via ranking oracles. In *The Twelfth International Conference on Learning Representations*, 2024. arXiv:[2303.03751](#)
- [49] Roman Vershynin. *High-dimensional probability*. Cambridge University Press, 2009.
- [50] Emmanouil-Vasileios Vlastakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Efficiently avoiding saddle points with zero order methods: No gradients required. *Advances in Neural Information Processing Systems*, 32, 2019. arXiv:[1910.13021](#)
- [51] Chenyi Zhang, Jiaqi Leng, and Tongyang Li. Quantum algorithms for escaping from saddle points. *Quantum*, 5:529, 2021. arXiv:[2007.10253](#)

- [52] Chenyi Zhang and Tongyang Li. Comparisons are all you need for optimizing smooth functions. arXiv:[2405.11454v1](#), 2024.
- [53] Hualin Zhang, Huan Xiong, and Bin Gu. Zeroth-order negative curvature finding: Escaping saddle points without gradients. *Advances in Neural Information Processing Systems*, 35:38332–38344, 2022. arXiv:[2210.01496](#)

A Auxiliary Lemmas for Classical Upper Bounds

A.1 Inner Product Concentration for Random Vectors on Sphere

In this subsection, we give a formal proof of the inner product concentration for random vectors on sphere stated in [Section 2.2](#), restated below:

Lemma 3. *Let $n \geq 5$. For any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$, and any constant $c > 0$, there exists constant p_1 and p_2 which is independent of n , such that*

$$p_1 \leq \Pr_{\mathbf{y} \sim S_n} [|\langle \mathbf{y}, \mathbf{x} \rangle| \leq \|\mathbf{x}\|/(c\sqrt{n})] \leq p_2.$$

where $\mathbf{y}$ is chosen from S_n uniformly at random. In particular, we have the inequalities below:

- $\Pr_{\mathbf{y} \sim S_n} [|\langle \mathbf{y}, \mathbf{x} \rangle| \leq \|\mathbf{x}\| \cdot \frac{24}{25\sqrt{n}}] \geq 3/5$
- $\Pr_{\mathbf{y} \sim S_n} [|\langle \mathbf{y}, \mathbf{x} \rangle| \leq \|\mathbf{x}\| \cdot \frac{18}{25\sqrt{n}}] \leq 11/20$
- $\Pr_{\mathbf{y} \sim S_n} [|\langle \mathbf{y}, \mathbf{x} \rangle| \geq \|\mathbf{x}\|/(5\sqrt{n})] \geq 4/5$

Proof. The existence of an absolute constant $p_1 > 0$ such that

$$\Pr_{\mathbf{y} \sim \text{Unif}(S_n)} \left[|\langle \mathbf{y}, \mathbf{x} \rangle| \leq \frac{\|\mathbf{x}\|}{c\sqrt{n}} \right] \geq p_1$$

for all $n \geq 5$ and all $\mathbf{x} \neq \mathbf{0}$ follows from standard anti-concentration results for one-dimensional marginals of the uniform distribution on the sphere; see Theorem 3.4.5 in [\[49\]](#). We focus on establishing the remaining bounds and the stated explicit constants.

By rotational invariance of the uniform distribution on S_n , there exists an orthogonal matrix Q such that $Q\mathbf{x} = \|\mathbf{x}\|\mathbf{e}_1$. Since $Q\mathbf{y} \sim \text{Unif}(S_n)$ and $\langle \mathbf{y}, \mathbf{x} \rangle = \langle Q\mathbf{y}, Q\mathbf{x} \rangle$, we assume without loss of generality that $\mathbf{x} = \|\mathbf{x}\|\mathbf{e}_1$. Then

$$\frac{\langle \mathbf{y}, \mathbf{x} \rangle}{\|\mathbf{x}\|} = y_1.$$

Define the rescaled coordinate $T_n := \sqrt{n}y_1$. The marginal density of y_1 is

$$f_n(t) = c_n(1-t^2)^{\frac{n-3}{2}}, \quad t \in [-1, 1], \quad c_n = \frac{\Gamma(n/2)}{\sqrt{\pi} \Gamma((n-1)/2)}.$$

By the change of variables $t = s/\sqrt{n}$, the density of T_n is

$$h_n(s) = c_n\sqrt{n} \left(1 - \frac{s^2}{n}\right)^{\frac{n-3}{2}} \mathbf{1}\{|s| \leq \sqrt{n}\}.$$

For $t \geq 0$, define $F_n(t) := \Pr(|T_n| \leq t) = 2 \int_0^t h_n(s) ds$. For each fixed $s \in [0, 1]$, $h_n(s)$ is nondecreasing in n for all $n \geq 5$, which implies that $F_n(t)$ is nondecreasing in n for every fixed $t \in [0, 1]$. On the other hand, letting $g \sim \mathcal{N}(0, I_n)$ and $\mathbf{y} = \mathbf{g}/\|\mathbf{g}\|$, we have $\mathbf{y} \sim \text{Unif}(S_n)$ and

$$T_n = \frac{g_1}{\sqrt{\frac{1}{n} \sum_{i=1}^n g_i^2}}.$$

By the law of large numbers and Slutsky's theorem, T_n converges in distribution to $Z \sim \mathcal{N}(0, 1)$, hence $F_n(t) \rightarrow 2\Phi(t) - 1$ for every fixed $t \geq 0$. Using monotonicity and the explicit density for $n = 5$, one obtains

$$F_n(0.96) \geq F_5(0.96) \geq \frac{3}{5},$$

which proves the first displayed inequality. Similarly,

$$F_n(0.72) \leq \lim_{m \rightarrow \infty} F_m(0.72) = 2\Phi(0.72) - 1 < \frac{11}{20},$$

and

$$\Pr(|T_n| \geq 1/5) = 1 - F_n(1/5) \geq \lim_{m \rightarrow \infty} 2(1 - \Phi(1/5)) > \frac{4}{5},$$

which completes the proof. $\square$

A.2 Constant Overlap via Basis Averaging

In this subsection, we prove that for a fixed unit vector in $\mathbb{R}^n$, assuming that its inner product with all basis vectors of a Haar-random orthonormal basis is at least $-1/n$, the given vector has a constant overlap with the uniform average of all basis vectors. Similar asymptotic bounds are standard in high-dimensional spherical measures; see, e.g., [13, 47].

Lemma 4. *Fix $n \geq 3$ and a unit vector $\mathbf{v} \in \mathbb{R}^n$. Let $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ be a Haar-random orthonormal basis in $\mathbb{R}^n$, conditioned on the event*

$$E_n := \left\{ \langle \mathbf{u}_i, \mathbf{v} \rangle \geq -\frac{1}{n}, \forall i \in [n] \right\}.$$

Define

$$W_n := \frac{1}{\sqrt{n}} \left\langle \sum_{i=1}^n \mathbf{u}_i, \mathbf{v} \right\rangle.$$

Then for all $n \geq 500$,

$$\mathbb{E}[W_n \mid E_n] > 0.7.$$

Proof. Denote

$$X := U^\top \mathbf{v}, \quad X_i = \langle \mathbf{u}_i, \mathbf{v} \rangle.$$

Since U is Haar-distributed and $\mathbf{v}$ is fixed, X is uniformly distributed on S^{n-1} . Thus

$$W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i, \quad E_n = \left\{ X_i \geq -\frac{1}{n} \forall i \in [n] \right\}.$$

Now let $G = (G_1, \dots, G_n) \sim N(0, I_n)$. From Eq. 3.15 of [49], $G/\|G\|$ is also uniformly distributed on S^{n-1} . Therefore,

$$\mathbb{E}[W_n \mid E_n] = \mathbb{E} \left[\frac{\sum_{i=1}^n G_i}{nR} \mid G_i \geq -\frac{R}{\sqrt{n}} \forall i \right].$$

Define $\alpha_n := 1/\sqrt{n}$. Let

$$m(\alpha) := \frac{\varphi(\alpha)}{\Phi(\alpha)}, \quad \varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \quad \Phi(t) = \int_{-\infty}^t \varphi(s) ds.$$

A standard Taylor expansion at $\alpha = 0$ gives

$$m(\alpha) \geq \sqrt{\frac{2}{\pi}} - \frac{2}{\pi}\alpha \quad \text{for all } \alpha \in [0, 1].$$

In particular, for all $n \geq 3$,

$$m(\alpha_n) \geq \sqrt{\frac{2}{\pi}} - \frac{2}{\pi} \frac{1}{\sqrt{n}}.$$

Next, to control the error introduced by replacing R^{-1} with 1, recall that

$$R = \frac{\|G\|}{\sqrt{n}}, \quad W_n = \frac{\frac{1}{n} \sum_{i=1}^n G_i}{R},$$

and we can get

$$|W_n - m(\alpha_n)| = \left| \frac{\sum_{i=1}^n G_i}{nR} - \frac{\sum_{i=1}^n G_i}{n} \right| = \frac{\sum_{i=1}^n G_i}{n} |R^{-1} - 1| \leq \frac{\sqrt{\sum_{i=1}^n G_i^2}}{\sqrt{n}} |R^{-1} - 1| = |R - 1|.$$

So we only need to bound $\mathbb{E}[|R - 1| \mid \mathbf{E}_n]$ explicitly. Note that R is independent with $\mathbf{E}_n$, we have

$$\mathbb{E}[|R - 1| \mid \mathbf{E}_n] = \mathbb{E}[|R - 1|].$$

Since $R = \|G\|/\sqrt{n}$ with $\|G\|^2 \sim \chi_n^2$, where χ_n^2 is the n -dimension chi-square distribution, we have $\text{Var}(R) \leq 1/n$ for all $n \geq 3$, and thus by Cauchy-Schwarz inequality we have

$$\mathbb{E}[|R - 1|] \leq \sqrt{\text{Var}(R)} \leq \frac{1}{\sqrt{2n}}.$$

Consequently,

$$\mathbb{E}[W_n \mid \mathbf{E}_n] \geq m(\alpha_n) - \frac{1}{\sqrt{2n}} \quad \text{for all } n \geq 3.$$

Combining the two inequalities,

$$\mathbb{E}[W_n \mid \mathbf{E}_n] \geq \sqrt{\frac{2}{\pi}} - \left(\frac{2}{\pi} + \frac{1}{\sqrt{2}} \right) \frac{1}{\sqrt{n}} \geq \sqrt{\frac{2}{\pi}} - \frac{1.35}{\sqrt{n}},$$

Finally, if $n \geq 225$ then $1.35/\sqrt{n} \leq 0.09$, which gives $\mathbb{E}[W_n \mid \mathbf{E}_n] \geq 0.7$. $\square$

B Classical Deterministic Gradient Testing

In this section, prove the tight $\Theta(n)$ bound for classical deterministic gradient testing.

B.1 Classical Deterministic Gradient Testing Algorithm

We first give a classical deterministic algorithm for gradient testing.

Theorem 7. *Algorithm 6 solves Problem 1 using $O(n)$ queries to a comparison oracle O_f^{Comp} (1).*

Denote $\alpha_i = |g_i|/|g_1|$ for any $i = 2, 3, \dots, n$, where $\mathbf{g}$ is the normalized gradient (see Section 2.1). Define $\delta = \sqrt{1/(1 - \varepsilon^2/2)^2 - 1}$. The distance $\|\mathbf{g} - \mathbf{v}\|$ captures angular deviation, while the ratios α_i capture relative coordinate imbalance. The following lemma establishes that, to solve Problem 1, it suffices for us to estimate each α_i within a constant multiplicative error.

Algorithm 6: Classical Deterministic Gradient Testing

Input: A point $\mathbf{x} \in \mathbb{R}^n$, testing direction $\mathbf{v}$, accuracy $\varepsilon \in (0, 1/\sqrt{2})$, lower bound γ on $\|\nabla f(\mathbf{x})\|$

Output: Test whether $\|\mathbf{g} - \mathbf{v}\| \leq \varepsilon$ (YES case) or $\|\mathbf{g} - \mathbf{v}\| \geq 2\varepsilon$ (NO case)

- 1 Set $\delta = \sqrt{1/(1 - \varepsilon^2/2)^2 - 1}$, $\Delta_1 = \gamma/(7n)$, $\Delta_2 = \frac{\gamma}{8n^2}$, $\Delta_3 = \frac{\gamma\delta}{30\sqrt{14}n^{1.5}}$ and define $w_i(\beta) := \beta \mathbf{e}_1 - \mathbf{e}_i$
- 2 For any $i = 2, 3, \dots, n$ initialize all $\ell_i = 1$, $t_i = 1$
- 3 **for** $i = 2$ to n **do**
- 4 Call **Algorithm 1** with input $(\mathbf{x}, \mathbf{e}_i, \Delta_1)$
- 5 **if** $\langle \nabla f, \mathbf{e}_i \rangle < \Delta_1$ **then**
- 6 $t_i \leftarrow -t_i$
- 7 Let $D = \text{diag}(1, t_2, \dots, t_n)$ and apply the orthogonal transformation D to the coordinate frame.
- 8 Set $\mathbf{y} = (2n, -1, \dots, -1)$ and call **Algorithm 1** with input $(\mathbf{x}, \mathbf{y}/\|\mathbf{y}\|, \Delta_2)$
- 9 **if** $\langle \nabla f, \mathbf{y}/\|\mathbf{y}\| \rangle \leq \Delta_2$ **then**
- 10 **return** No
- 11 **for** $i = 2$ to n **do**
- 12 **repeat**
- 13 Set $\beta = \delta \ell_i / \sqrt{7n}$, and call **Algorithm 1** with input $(\mathbf{x}, w_i(\beta)/\|w_i(\beta)\|, \Delta_3)$
- 14 **if** $\langle \nabla f, w_i(\beta)/\|w_i(\beta)\| \rangle \leq \Delta_3$ **then**
- 15 $\ell_i \leftarrow 3\ell_i/2$
- 16 **else**
- 17 **break**
- 18 **if** $\sum_{j=2}^n \ell_j^2 \geq 21n$ **then**
- 19 **break and return** No
- 20 **return** Yes

Lemma 5. If $\|\mathbf{g} - \mathbf{v}\| \leq \varepsilon$, we have $\sum_{i=2}^n \alpha_i^2 \leq \delta^2$ and $g_1^2 \sum_{i=2}^n \alpha_i^2 \leq \varepsilon^2$. If $\|\mathbf{g} - \mathbf{v}\| \geq 2\varepsilon$, we have $\sum_{i=2}^n \alpha_i^2 \geq 4\delta^2$ and $g_1^2 \sum_{i=2}^n \alpha_i^2 \geq 2\varepsilon^2$.

Proof. Due to $\mathbf{v} = \mathbf{e}_1 = (1, 0, \dots, 0)$ and $\|\mathbf{g}\| = 1$, we have

$$\|\mathbf{g} - \mathbf{v}\|^2 = (1 - g_1)^2 + g_2^2 + \dots + g_n^2 = (1 - g_1)^2 + 1 - g_1^2 = 2 - 2g_1.$$

If $\|\mathbf{g} - \mathbf{v}\| \leq \varepsilon$, we can get $g_1 \geq 1 - \varepsilon^2/2$, which means

$$\sum_{i=2}^n \alpha_i^2 = (1 - g_1^2)/g_1^2 \leq 1/(1 - \varepsilon^2/2)^2 - 1 \leq \delta^2.$$

If $\|\mathbf{g} - \mathbf{v}\| \geq 2\varepsilon$, we can get $g_1 \leq 1 - 2\varepsilon^2$, which means

$$\sum_{i=2}^n \alpha_i^2 = (1 - g_1^2)/g_1^2 \geq 1/(1 - 2\varepsilon^2)^2 - 1 \geq 4\delta^2.$$

□

Intuitively, [Algorithm 6](#) proceeds as follows: We first choose a suitable $\mathbf{y}$ to exclude the case when $\langle \mathbf{v}, \mathbf{g} \rangle$ is too small, which implies $\mathbf{v}$ belongs to the NO case. Then we call [Algorithm 1](#) and use $w_i(\beta) := \beta \mathbf{e}_1 - \mathbf{e}_i$ to compare β and $|g_i/g_1|$ for each i . Note that if we call [Algorithm 1](#) with input direction $w_i(\beta)$, we can roughly get the sign of

$$\langle \nabla f, w_i(\beta) / \|w_i(\beta)\| \rangle = \|\nabla f\|(\beta g_1 - g_i) / \|w_i(\beta)\|, \quad (10)$$

which is determined by $\beta - g_i/g_1$. Therefore we can roughly get the sign of $\beta - g_i/g_1$ by it.

If β is roughly less than $|g_i/g_1|$, it is multiplied by $3/2$, and finally we find $\beta = \delta \ell_i / \sqrt{7n}$ which satisfies $2\beta/3 < |g_i/g_1| < \beta$ approximately. And then we can count $\sum_{i=2}^n \ell_i^2$ to get the range of $\|\mathbf{v} - \mathbf{g}\|$.

Proof of [Theorem 7](#). In [Line 9](#), we exclude the extreme case where $g_1 = \langle \mathbf{g}, \mathbf{v} \rangle$ is very small. On the one hand, if

$$\langle \nabla f, \mathbf{y} / \|\mathbf{y}\| \rangle \leq \Delta,$$

we have

$$2ng_1 \leq \sum_{i=2}^n g_i + \frac{4n\Delta_2}{\gamma} \leq \sqrt{n \sum_{i=2}^n g_i^2} + \frac{1}{2n} = \sqrt{n(1 - g_1^2)} + \frac{1}{2n} \leq \sqrt{n} + \frac{1}{2n},$$

which means

$$\langle \mathbf{v}, \mathbf{g} \rangle = g_1 \leq \frac{1}{\sqrt{n}}.$$

We can get that $\mathbf{v}$ belongs to the NO case. On the other hand, if

$$\langle \nabla f, \mathbf{y} / \|\mathbf{y}\| \rangle \geq -\Delta,$$

we have

$$2ng_1 \geq \sum_{i=2}^n g_i - \frac{4n\Delta_2}{\gamma} \geq \frac{1}{2} \sum_{i=2}^n g_i^2 - \frac{1}{2n} = \frac{1}{2}(1 - g_1^2) - \frac{1}{2n},$$

which means $g_1 \geq 1/10n$. Therefore, in the rest of the proof, we assume

$$g_1 = \langle \mathbf{g}, \mathbf{v} \rangle \geq \frac{1}{10n}. \quad (11)$$

With (10), in our [Lines 12–19](#), if we get $\delta \ell_i / \sqrt{7n} - |\alpha_i|$ is negative within error Δ_3 , we will multiply ℓ_i by $3/2$, and if ℓ_i remains unchanged, we promise that we find the smallest integer ℓ_i which is an integer power of $3/2$ and satisfies $\delta \ell_i / \sqrt{7n} \geq |\alpha_i|$ (up to Δ_3 , the precision of calling [Algorithm 1](#) in [Line 13](#)). Consequently, for any $i = 2, 3, \dots, n$, we can find such an ℓ_i using $\lceil 1 + \log \ell_i \rceil$ comparison queries.

On the one hand, we prove that if $\sum_{j=2}^n \ell_j^2 \geq 21n$, we will have $\|\mathbf{g} - \mathbf{v}\| \geq \varepsilon^2$. Denote $S = \{i \mid i \in \{2, 3, \dots, n\}, \ell_i > 1\}$. Note that if $i \in S$, it means that we have started the recursive setting of powers of $3/2$ in [Lines 12–19](#) above, and we can finally get ℓ_i which do not satisfy the judge in [Line 14](#). It means that we have

$$\langle \nabla f, w_i(2\delta \ell_i / 3\sqrt{7n}) / \|w_i(2\delta \ell_i / 3\sqrt{7n})\| \rangle \leq \Delta_3, \quad (12)$$

$$\langle \nabla f, w_i(\delta \ell_i / \sqrt{7n}) / \|w_i(\delta \ell_i / \sqrt{7n})\| \rangle \geq -\Delta_3. \quad (13)$$

Combine (10) and (12), we can get

$$\begin{aligned} \frac{2\delta\ell_i}{3\sqrt{7n}}g_1 - g_i &\leq \frac{\Delta_3}{\gamma} \sqrt{1 + \frac{4\delta^2\ell_i^2}{63n}} \leq \sqrt{2}\Delta_3/\gamma, \\ |\alpha_i| = |g_i/g_1| &\geq \frac{2\delta\ell_i}{3\sqrt{7n}} - \sqrt{2}\Delta_3/(\gamma g_1). \end{aligned}$$

By our choice of Δ_3 in Line 1 and (11), we have

$$|\alpha_i| \geq \frac{2\delta\ell_i}{3\sqrt{7n}} - \frac{\delta}{3\sqrt{7n}}.$$

Applying the following Cauchy-Schwarz inequality,

$$\left(|\alpha_i|^2 + \frac{\delta^2}{9n}\right) \cdot \left(1 + \frac{1}{7}\right) \geq \left(|\alpha_i| + \frac{\delta}{3\sqrt{7n}}\right)^2,$$

we obtain

$$|\alpha_i|^2 \geq \frac{\delta^2\ell_i^2}{18n} - \frac{\delta^2}{9n}.$$

Therefore, if $\sum_{i=2}^n \ell_i^2 \geq 21n$, we have $\sum_{i \in S} \ell_i^2 \geq 20n$, and

$$\sum_{i \in S} |\alpha_i|^2 \geq \sum_{i \in S} \left(\frac{\delta^2\ell_i^2}{18n} - \frac{\delta^2}{9n} \right) \geq \delta^2 \left(\frac{20}{18} - \frac{1}{9} \right) = \delta^2.$$

Under the promise of Problem 1 and Lemma 5, this implies that $\mathbf{x}$ belongs to the NO case.

On the other hand, for the case where $\sum_{i=2}^n \ell_i^2 \leq 21n$. From (13), we have

$$|\alpha_i| \leq \frac{\ell_i}{\sqrt{7n}} + \sqrt{1 + \frac{\delta^2\ell_i^2}{7n}} \Delta_3/(\gamma g_1).$$

Use the value of Δ_3 we set in Line 1 and (11), we have

$$|\alpha_i| \leq \frac{\delta\ell_i}{\sqrt{7n}} + \frac{\delta}{3\sqrt{7n}}.$$

By Cauchy-Schwarz inequality,

$$|\alpha_i|^2 \leq \left(1 + \frac{1}{6}\right) \left(\frac{\delta^2\ell_i^2}{7n} + \frac{\delta^2}{63n}\right) \leq \frac{\delta^2\ell_i^2}{6n} + \frac{\delta^2}{9n}.$$

If $\sum_{i=2}^n \ell_i^2 \leq 21n$, we have

$$\sum_{i=2}^n |\alpha_i|^2 \leq \sum_{i=2}^n \left(\frac{\delta^2\ell_i^2}{6n} + \frac{\delta^2}{9n} \right) \leq \delta^2 \left(\frac{21}{6} + \frac{1}{9} \right) < 4\delta^2.$$

Under the promise of Problem 1 and Lemma 5, this implies that $\mathbf{x}$ belongs to the YES case.

Next, we bound the query complexity of Algorithm 6. Note that when Algorithm 6 terminates, all ℓ_i satisfy

$$\sum_{i=2}^n \ell_i \leq \sum_{i=2}^n \ell_i^2 \leq \frac{9}{4} \cdot \sum_{i=2}^n \max\{2\ell_i/3, 1\}^2 = O(n).$$

Based on our recursive setting in Lines 12–19, using

$$\sum_{i=1}^n (1 + \log \ell_i) \leq n + \sum_{i=1}^n \ell_i \tag{14}$$

queries, we can find all values of ℓ_i with $O(n)$ queries. $\square$

B.2 Classical Lower Bound on Gradient Testing

Here, we prove our classical deterministic lower bound for gradient testing, which implies that our [Algorithm 6](#) is optimal. Our lower bound is obtained by considering the function $f(\mathbf{x}) = \langle \mathbf{g}, \mathbf{x} \rangle + b$ for some $\mathbf{g} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, whose comparison query oracle satisfies

$$O_f^{\text{Comp}}(\mathbf{x}, \mathbf{x} + \mathbf{y}) = \text{sgn}(\langle \nabla f(\mathbf{x}), \mathbf{y} \rangle) = \text{sgn}(\langle \mathbf{g}, \mathbf{y} \rangle). \quad (15)$$

Theorem 8 (Deterministic lower bound for gradient testing). *Fix $\varepsilon \in (0, 1/10)$ and consider the objective*

$$f_{\mathbf{g},b}(\mathbf{x}) = \langle \mathbf{g}, \mathbf{x} \rangle + b.$$

Any deterministic algorithm that, for every $b \in \mathbb{R}$ and unit vectors $\mathbf{g}, \mathbf{v} \in \mathbb{R}^n$, correctly distinguishes

$$(\text{YES}) \quad \|\mathbf{g} - \mathbf{v}\|_2 \leq \varepsilon \quad \text{from} \quad (\text{NO}) \quad \|\mathbf{g} - \mathbf{v}\|_2 > 2\varepsilon$$

using access to $O_{f_{\mathbf{g},b}}^{\text{Comp}}$ must make at least $n - 1$ comparison queries.

Proof. By rotation invariance, assume $\mathbf{v} = \mathbf{e}_1$. Fix a deterministic algorithm $\mathcal{A}$ making at most $q \leq n - 2$ comparison queries. We construct two affine objectives, one YES and one NO, whose comparison transcripts are identical. Set $b = 0$ and first consider the YES gradient

$$\mathbf{g}_Y = \mathbf{e}_1.$$

Run $\mathcal{A}$ with the oracle for $f_Y(\mathbf{x}) = \langle \mathbf{g}_Y, \mathbf{x} \rangle$, let $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(q)}$ be the queries made on this execution path, and define

$$W := \text{span}\{\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(q)}\}.$$

Since $\dim W \leq q \leq n - 2$,

$$\dim(\mathbf{e}_1^\perp \cap W^\perp) \geq n - 1 - q \geq 1.$$

Choose a unit vector $\mathbf{u} \in \mathbf{e}_1^\perp \cap W^\perp$. Now define the NO gradient

$$\mathbf{g}_N := \frac{\mathbf{e}_1 + \delta \mathbf{u}}{\sqrt{1 + \delta^2}}, \quad \delta := 3\varepsilon.$$

Then $\|\mathbf{g}_N\| = 1$ and

$$\|\mathbf{g}_N - \mathbf{e}_1\| \geq |\langle \mathbf{g}_N - \mathbf{e}_1, \mathbf{u} \rangle| = \frac{\delta}{\sqrt{1 + \delta^2}} > 2\varepsilon,$$

where the last inequality uses $\delta = 3\varepsilon$ and $\varepsilon < 1/10$, indicating that $\mathbf{g}_N$ is a NO instance.

It remains to show that $\mathbf{g}_N$ produces an identical transcript as $\mathbf{g}_Y$. Proceed by induction on the interaction. Suppose the transcripts agree up to query $t - 1$, for some $t < q$. Since $\mathcal{A}$ is deterministic, its next query on $\mathbf{g}_N$ is the same YES-path query $\mathbf{w}^{(t)}$. For this query, since $\mathbf{w}^{(t)} \in W$ and $\mathbf{u} \perp W$,

$$\langle \mathbf{g}_N, \mathbf{w}^{(t)} \rangle = \frac{\langle \mathbf{e}_1, \mathbf{w}^{(t)} \rangle + \delta \langle \mathbf{u}, \mathbf{w}^{(t)} \rangle}{\sqrt{1 + \delta^2}} = \frac{\langle \mathbf{e}_1, \mathbf{w}^{(t)} \rangle}{\sqrt{1 + \delta^2}}.$$

Hence

$$\text{sgn}(\langle \mathbf{g}_N, \mathbf{w}^{(t)} \rangle) = \text{sgn}(\langle \mathbf{e}_1, \mathbf{w}^{(t)} \rangle) = \text{sgn}(\langle \mathbf{g}_Y, \mathbf{w}^{(t)} \rangle).$$

Thus the oracle replies agree at query t , completing the induction. After the first q replies, $\mathcal{A}$ is in the same state as on the YES execution and therefore has same output. Since $\mathbf{g}_Y$ is a YES instance and $\mathbf{g}_N$ is a NO instance, $\mathcal{A}$ errs on one of them, and we can conclude that any correct deterministic algorithm needs at least $n - 1$ queries. $\square$

C Quantum Gradient Estimation

Here, we prove the following quantum gradient estimation algorithm adapted from [29] estimates gradients to certain precision assuming an approximate quantum evaluation oracle:

Proposition 3 (Quantum gradient estimation). *Let $n \geq 2$ and $t \in \mathbb{Z}_+$. Define $T := t + 1$ and*

$$M := \left\{0, \frac{1}{t}, \frac{2}{t}, \dots, 1\right\}^n, \quad N := |M| = T^n.$$

For any $\mathbf{x} \in [0, 1]^n$, define the quantum state

$$|\Phi_{\mathbf{x}}\rangle := \frac{1}{T^{n/2}} \sum_{\mathbf{y} \in \{0, 1, \dots, T-1\}^n} e^{2\pi i \langle \mathbf{y}, \mathbf{x} \rangle} |\mathbf{y}\rangle. \quad (16)$$

Let F_T denote the T -dimensional quantum Fourier transform over $\mathbb{Z}_T$, i.e., $F_T|j\rangle = \frac{1}{\sqrt{T}} \sum_{k=0}^{T-1} \mathbf{e}^{-2\pi i j k / T} |k\rangle$, and denote $F := F_T^{\otimes n}$. Consider the following procedure: apply $F^\dagger$ to the input state, which is $|\Phi_{\mathbf{x}}\rangle$ in ideal case and $|\psi\rangle$ in approximation case, measure in the computational basis obtaining $K \in \{0, 1, \dots, T-1\}^n$, and output $\hat{\mathbf{v}} := \frac{1}{t} K \in M$. Let

$$m := \left\lceil 2 + \frac{3}{2}n \right\rceil. \quad (17)$$

Then for every $\mathbf{x} \in [0, 1]^n$,

$$\Pr \left[\|\hat{\mathbf{v}} - \mathbf{x}\|_2 \leq \frac{\sqrt{n}(m+1)}{t} \right] \geq \frac{2}{3}. \quad (18)$$

Moreover, if the actual input state $|\psi\rangle$ satisfies $\| |\psi\rangle - |\Phi_{\mathbf{x}}\rangle \|_2 \leq \varepsilon$, then the output satisfies

$$\Pr \left[\|\hat{\mathbf{v}} - \mathbf{x}\|_2 \leq \frac{\sqrt{n}(m+1)}{t} \right] \geq \frac{2}{3} - 2\varepsilon. \quad (19)$$

Proof. We first analyze the ideal input $|\Phi_{\mathbf{x}}\rangle$. Since

$$e^{2\pi i \langle \mathbf{y}, \mathbf{x} \rangle} = \prod_{j=1}^n e^{2\pi i y_j x_j} \quad \text{and} \quad F = F_T^{\otimes n},$$

the state $|\Phi_{\mathbf{x}}\rangle$ decomposes as a tensor product across coordinates. Consequently, after applying $F^\dagger$, the measurement outcome $K = (K_1, \dots, K_n)$ consists of independent coordinates, where each K_j arises from measuring

$$|\phi_\alpha\rangle := \frac{1}{\sqrt{T}} \sum_{y=0}^{T-1} e^{2\pi i y \alpha} |y\rangle, \quad \alpha = x_j.$$

Applying $F_T^\dagger$ to $|\phi_\alpha\rangle$, the amplitude of observing $k \in \{0, \dots, T-1\}$ is

$$a_k = \langle k | F_T^\dagger | \phi_\alpha \rangle = \frac{1}{T} \sum_{y=0}^{T-1} e^{2\pi i y (\alpha - k/T)}.$$

Let $\delta_k := \alpha - k/T$. Summing the geometric series gives

$$a_k = \frac{1}{T} \cdot \frac{1 - e^{2\pi i T \delta_k}}{1 - e^{2\pi i \delta_k}} = \frac{1}{T} e^{\pi i (T-1) \delta_k} \frac{\sin(\pi T \delta_k)}{\sin(\pi \delta_k)}.$$

Hence

$$\Pr[K_j = k] = \frac{1}{T^2} \left(\frac{\sin(\pi T \delta_k)}{\sin(\pi \delta_k)} \right)^2.$$

Let $\theta := T\alpha$. Then $\delta_k = (\theta - k)/T$ and

$$\Pr[K_j = k] = \frac{1}{T^2} \left(\frac{\sin(\pi(\theta - k))}{\sin(\pi(\theta - k)/T)} \right)^2.$$

Using $|\sin(\pi z)| \leq 1$ and the inequality

$$|\sin(\pi u)| \geq 2|u| \quad \text{for } |u| \leq \frac{1}{2},$$

we obtain, for all k ,

$$\Pr[K_j = k] \leq \frac{1}{4(\theta - k)^2}.$$

Therefore, for any integer $m \geq 2$,

$$\Pr[|K_j - \theta| \geq m] \leq \sum_{r=m}^{\infty} \frac{1}{2r^2} \leq \frac{1}{2(m-1)}. \quad (20)$$

Let $\theta_j := Tx_j$. By (20),

$$\Pr[|K_j - \theta_j| \geq m] \leq \frac{1}{2(m-1)}.$$

Applying the union bound,

$$\Pr[\exists j : |K_j - \theta_j| \geq m] \leq \frac{n}{2(m-1)}.$$

With the choice of (17), this probability is at most $1/3$, hence

$$\Pr[\forall j : |K_j - \theta_j| < m] \geq \frac{2}{3}.$$

On this event,

$$|\hat{v}_j - x_j| \leq \frac{m+1}{t},$$

and thus

$$\|\hat{\mathbf{v}} - \mathbf{x}\|_2 \leq \frac{\sqrt{n}(m+1)}{t}.$$

This proves (18). Since $F^\dagger$ is unitary,

$$\|F^\dagger|\psi\rangle - F^\dagger|\Phi_{\mathbf{x}}\rangle\|_2 = \|\psi\rangle - |\Phi_{\mathbf{x}}\rangle\|_2 \leq \varepsilon.$$

Let P and Q denote the measurement distributions obtained from $F^\dagger|\psi\rangle$ and $F^\dagger|\Phi_{\mathbf{x}}\rangle$, respectively. For any event E ,

$$|P(E) - Q(E)| \leq \frac{1}{2} \|\psi\rangle\langle\psi| - |\Phi_{\mathbf{x}}\rangle\langle\Phi_{\mathbf{x}}|\|_1 \leq \|\psi\rangle - |\Phi_{\mathbf{x}}\rangle\|_2 \leq \varepsilon.$$

Let E be the event

$$\|\hat{\mathbf{v}} - \mathbf{x}\|_2 \leq \frac{\sqrt{n}(m+1)}{t}.$$

Since $Q(E) \geq 2/3$, it follows that

$$P(E) \geq \frac{2}{3} - \varepsilon \geq \frac{2}{3} - 2\varepsilon,$$

which proves (19). □

D Proof Details of Gradient Estimation Lower Bounds

D.1 Classical Lower Bound

In this subsection, we prove the classical lower bound on gradient estimation ([Theorem 5](#)). Similarly, this is also obtained by considering to the special case in [Appendix B.2](#) and our query oracle becomes [\(15\)](#). Our lower bound relies on the following two lemmas.

Lemma 6. *Let $n \geq 1$ and $0 < \varepsilon \leq 1$. There exists a set $M \subset S_n$ such that*

$$\forall \mathbf{x} \neq \mathbf{y} \in M, \quad \|\mathbf{x} - \mathbf{y}\| > \varepsilon, \quad \text{and} \quad |M| \geq \frac{2}{\varepsilon^n}.$$

Proof. We construct M greedily as a *maximal* ε -separated subset of S_n : start with $M = \emptyset$ and keep adding points $\mathbf{x} \in S_n$ with $\|\mathbf{x} - \mathbf{y}\| > \varepsilon$ for all $\mathbf{y} \in M$, until this is no longer possible. By construction, M is ε -separated. Maximality implies the covering property

$$S_n \subseteq \bigcup_{\mathbf{x} \in M} B(\mathbf{x}, \varepsilon), \quad B(\mathbf{x}, \varepsilon) := \{\mathbf{z} \in S^n : \|\mathbf{z} - \mathbf{x}\| \leq \varepsilon\},$$

because otherwise a point outside the union could be added to M , contradiction. Hence, writing $\sigma_n(\cdot)$ for the standard rotation-invariant surface measure on S_n , and equivalently, $\sigma_n(A)/\sigma_n(S_n)$ is the probability that a uniformly random point on S_n lies in A . Then we have

$$\sigma_n(S_n) \leq \sum_{\mathbf{x} \in M} \sigma_n(B(\mathbf{x}, \varepsilon)) \leq |M| \sup_{\mathbf{x} \in S_n} \sigma_n(B(\mathbf{x}, \varepsilon)).$$

So it suffices to show the following: for every $\mathbf{x} \in S_n$ and $0 < \varepsilon \leq 1$,

$$\sigma_n(B(\mathbf{x}, \varepsilon)) \leq \frac{1}{2} \varepsilon^n \sigma_n(S_n). \tag{21}$$

Assuming [\(21\)](#) is true, we get

$$\sigma_n(S_n) \leq |M| \cdot \frac{1}{2} \varepsilon^n \sigma_n(S_n) \Rightarrow |M| \geq \frac{2}{\varepsilon^n},$$

as claimed. In the rest of the proof, we prove [\(21\)](#). Fix $\mathbf{x} \in S_n$ and let Z be uniform on S_n . Then

$$\frac{\sigma_n(B(\mathbf{x}, \varepsilon))}{\sigma_n(S^n)} = \Pr[\|Z - \mathbf{x}\| \leq \varepsilon].$$

Let $T := \langle Z, \mathbf{x} \rangle \in [-1, 1]$. Since $\|Z - \mathbf{x}\|^2 = 2 - 2\langle Z, \mathbf{x} \rangle$, the event $\|Z - \mathbf{x}\| \leq \varepsilon$ is equivalent to

$$T \geq 1 - \frac{\varepsilon^2}{2}.$$

The density of T is

$$f_n(t) = c_n (1 - t^2)^{\frac{n-1}{2}}, \quad c_n := \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi} \Gamma(\frac{n}{2})}.$$

For $t \in [0, 1]$ we have $1 - t^2 = (1 - t)(1 + t) \leq 2(1 - t)$, so with $\delta := \varepsilon^2/2$,

$$\begin{aligned}
\Pr [T \geq 1 - \delta] &= \int_{1-\delta}^1 c_n (1 - t^2)^{\frac{n-1}{2}} dt \\
&\leq \int_{1-\delta}^1 c_n (2(1 - t))^{\frac{n-1}{2}} dt \\
&= c_n 2^{\frac{n-1}{2}} \int_0^\delta u^{\frac{n-1}{2}} du \quad (u = 1 - t) \\
&= c_n 2^{\frac{n-1}{2}} \cdot \frac{\delta^{\frac{n+1}{2}}}{\frac{n+1}{2}} = \frac{c_n}{n+1} \varepsilon^{n+1}.
\end{aligned}$$

Using the standard gamma-ratio bound

$$c_n = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi} \Gamma(\frac{n}{2})} \leq \sqrt{\frac{n}{2}}$$

we obtain

$$\Pr [\|Z - x\| \leq \varepsilon] = \Pr \left[T \geq 1 - \frac{\varepsilon^2}{2} \right] \leq \frac{1}{2} \varepsilon^{n+1} \leq \frac{1}{2} \varepsilon^n$$

for any $\varepsilon \leq 1$. This finishes the proof. $\square$

We need another lemma on adaptive query lower bounds. We prove that any query model with three possible outcomes, corresponding to $<$, $>$, $=$ in our comparison query model, need to use at least $\log n$ queries to distinguish n different results, no matter it is deterministic algorithm or random algorithm.

Lemma 7. *Let I be an unknown index uniformly distributed over $[n] := \{1, 2, \dots, n\}$. Consider a query model in which each query returns one of three possible outcomes, and each I corresponds to a type of query oracle. For different I , their oracles are also different. Any (possibly randomized) algorithm that identifies I with success probability at least $2/3$ requires $\Omega(\log n)$ queries in the worst case.*

Proof. Consider the uniform distribution over I . By Yao's minimax principle, there exists a fixing of the internal randomness of $\mathcal{A}$ such that the resulting deterministic algorithm $\mathcal{A}_r$ succeeds with probability at least $2/3$ when I is drawn uniformly from $[n]$. Therefore, it suffices to prove that any deterministic algorithm that makes at most T queries has success probability at most $O(3^T/n)$ under the uniform distribution.

Now consider an arbitrary deterministic algorithm making at most T adaptive queries. For such a determined algorithm, it can be represented as a ternary decision tree of depth at most T . Each internal node corresponds to a query, and each node has exactly three outgoing edges, one for each possible query outcome. Each leaf of the tree outputs a value in $[n]$.

A decision tree of depth at most T has at most 3^T leaves, and hence the algorithm can correctly identify at most 3^T values of I . Under the uniform distribution over I , the success probability of any such deterministic algorithm is therefore at most $3^T/n$. To ensure that the success probability is least $2/3$, we must have $T \geq \log_3 n - O(1)$. This proves that any randomized algorithm identifying I with constant success probability requires $\Omega(\log n)$ queries. $\square$

Proof of Theorem 5. By Lemma 6, we can get an ε -net M on S_n , which satisfies that for each $\mathbf{x}, \mathbf{y} \in M$, we have $\|\mathbf{x} - \mathbf{y}\| > \varepsilon$ and $|M| = \Omega((1/\varepsilon)^n)$.

Assume our gradient direction belongs to M . Note that each comparison query only has three types of possible responses. As a result, by [Lemma 7](#), if we want to distinguish each gradient direction in M , the query complexity is at least

$$\log_3 |M| = \Omega(n \log(1/\varepsilon)).$$

□

D.2 Quantum Lower Bound

In this subsection, we prove the quantum lower bound on gradient estimation ([Theorem 6](#)). It is still obtained by considering the special case in [Appendix B.2](#) and our quantum oracle becomes

$$O_{f,Q}^{\text{Comp}}|\mathbf{x}\rangle|\mathbf{x} + \mathbf{y}\rangle|z\rangle = |\mathbf{x}\rangle|\mathbf{y}\rangle|z \oplus \text{sgn}(\langle \nabla f(\mathbf{x}), \mathbf{y} \rangle)\rangle. \quad (22)$$

Our proof is decomposed into three parts: a bound on average sensitivity of halfspaces ([Appendix D.2.1](#)), a bound on oracle distinction ([Appendix D.2.2](#)), and finally the proof of [Theorem 6](#) ([Appendix D.2.3](#)).

D.2.1 Average sensitivity of halfspaces

Lemma 8 (Average k -flip sensitivity of halfspaces). *There exists a universal constant $C' > 0$ such that the following holds. Let*

$$h(\mathbf{z}) = \mathbf{1} \left[\sum_{i=1}^n w_i z_i \geq 0 \right], \quad \mathbf{z} \in \{-1, 1\}^n,$$

be a halfspace. Let Z be uniform on $\{-1, 1\}^n$, and let S be uniform over all k -subsets of $[n]$. Then

$$\mathbb{E}_{S: |S|=k} \Pr [h(Z) \neq h(Z^{\oplus S})] \leq C' \sqrt{\frac{k}{n}}.$$

Moreover, for all $k \leq n/2$ one may take C' universal (independent of n, k).

Proof. Scaling $\mathbf{w}$ by any $\lambda > 0$ does not change h , hence we may assume

$$\|\mathbf{w}\|_2^2 = \sum_{i=1}^n w_i^2 = 1.$$

Also, flipping the sign of a coordinate z_i and simultaneously flipping the sign of w_i leaves the distribution of Z invariant and does not change the probability of disagreement. Thus, without loss of generality we may assume $w_i \geq 0$ for all i . Fix $S \subseteq [n]$ with $|S| = k$. Write

$$X := \sum_{i=1}^n w_i Z_i, \quad B_S := \sum_{i \in S} w_i Z_i, \quad Y_S := \sum_{i \notin S} w_i Z_i.$$

Then $X = Y_S + B_S$. Flipping the coordinates in S sends $Z_i \mapsto -Z_i$ for $i \in S$, hence the corresponding linear form becomes

$$X' := \sum_{i=1}^n w_i (Z^{\oplus S})_i = Y_S - B_S.$$

Therefore,

$$h(Z) \neq h(Z^{\oplus S}) \iff \mathbf{1}[X \geq 0] \neq \mathbf{1}[X' \geq 0] \iff (X)(X') < 0 \text{ or } XX' = 0.$$

Since $XX' = (Y_S + B_S)(Y_S - B_S) = Y_S^2 - B_S^2$, we obtain the inclusion

$$\{h(Z) \neq h(Z^{\oplus S})\} \subseteq \{|Y_S| \leq |B_S|\}. \quad (23)$$

Let

$$\sigma_S^2 := \text{Var}(Y_S) = \sum_{i \notin S} w_i^2 = 1 - \sum_{i \in S} w_i^2 = 1 - \|w_S\|_2^2.$$

Consider the normalized sum $\tilde{Y}_S := Y_S/\sigma_S$, which has variance 1. A standard anti-concentration inequality for Rademacher sums (a consequence of a local limit theorem, see e.g., [19]) states that there exists a universal constant $C_0 > 0$ such that for every $t > 0$,

$$\Pr(|\tilde{Y}_S| \leq t) \leq C_0 t. \quad (24)$$

Equivalently, for every $u > 0$,

$$\Pr(|Y_S| \leq u) \leq C_0 \frac{u}{\sigma_S}. \quad (25)$$

Since the random variables Y_S and B_S depend on disjoint sets of coordinates of Z , we can get that Y_S and B_S are independent. Using (23), independence, and conditioning on B_S , we have

$$\begin{aligned} \Pr(h(Z) \neq h(Z^{\oplus S})) &\leq \Pr(|Y_S| \leq |B_S|) \\ &= \mathbb{E}\left[\Pr(|Y_S| \leq |B_S| \mid B_S)\right] \\ &\leq \mathbb{E}\left[C_0 \frac{|B_S|}{\sigma_S}\right] \quad (\text{by (25) with } u = |B_S|) \\ &= \frac{C_0}{\sigma_S} \mathbb{E}|B_S|. \end{aligned}$$

By Cauchy-Schwarz inequality, there exists a universal constant $C_1 > 0$ such that

$$\mathbb{E}|B_S| = \mathbb{E}\left|\sum_{i \in S} w_i Z_i\right| \leq C_1 \left(\sum_{i \in S} w_i^2\right)^{1/2} = C_1 \|w_S\|_2.$$

Hence

$$\Pr(h(Z) \neq h(Z^{\oplus S})) \leq C_0 C_1 \frac{\|w_S\|_2}{\sqrt{1 - \|w_S\|_2^2}}. \quad (26)$$

Define $a := \|w_S\|_2 \in [0, 1]$. We claim there is a universal $C_2 > 0$ such that for all $a \in [0, 1]$,

$$\min\left\{1, \frac{a}{\sqrt{1 - a^2}}\right\} \leq C_2 a. \quad (27)$$

Indeed, if $a \leq 1/2$ then $\sqrt{1 - a^2} \geq \sqrt{3}/2$ and $\frac{a}{\sqrt{1 - a^2}} \leq \frac{2}{\sqrt{3}}a$. If $a > 1/2$, then $\min\{1, \frac{a}{\sqrt{1 - a^2}}\} \leq 1 \leq 2a$. Thus (27) holds with $C_2 := 2$.

Combining (26) with the trivial bound $\Pr(h(Z) \neq h(Z^{\oplus S})) \leq 1$ and using (27), we get a universal constant $C_3 > 0$ such that for every fixed S ,

$$\Pr(h(Z) \neq h(Z^{\oplus S})) \leq C_3 \|w_S\|_2. \quad (28)$$

Take the expectation over uniform S with $|S| = k$:

$$\mathbb{E}_S \Pr(h(Z) \neq h(Z^{\oplus S})) \leq C_3 \mathbb{E}_S \|w_S\|_2.$$

By Jensen's inequality,

$$\mathbb{E}_S \|w_S\|_2 \leq \sqrt{\mathbb{E}_S \|w_S\|_2^2}.$$

Moreover,

$$\mathbb{E}_S \|w_S\|_2^2 = \mathbb{E}_S \sum_{i \in S} w_i^2 = \sum_{i=1}^n w_i^2 \Pr(i \in S) = \sum_{i=1}^n w_i^2 \frac{k}{n} = \frac{k}{n} \|w\|_2^2 = \frac{k}{n}.$$

Therefore,

$$\mathbb{E}_{S: |S|=k} \Pr[h(Z) \neq h(Z^{\oplus S})] \leq C_3 \sqrt{\frac{k}{n}}.$$

Setting $C' := C_3$ completes the proof. $\square$

D.2.2 Distinction bound

Lemma 9. *Let $\mathbf{z} \in \{-1, 1\}^n$ and consider the oracle*

$$O_{\mathbf{z}} : |\mathbf{x}, \mathbf{y}\rangle \mapsto |\mathbf{x}, \mathbf{y} \oplus f(\mathbf{x}, \mathbf{z})\rangle,$$

where $\mathbf{x} \in \mathbb{Z}^n$ (coordinates not necessarily ± 1) and

$$f(\mathbf{x}, \mathbf{z}) = \begin{cases} 1 & \text{if } \langle \mathbf{x}, \mathbf{z} \rangle \geq 0, \\ 0 & \text{if } \langle \mathbf{x}, \mathbf{z} \rangle < 0. \end{cases}$$

Let

$$M := \{(\mathbf{z}, \tilde{\mathbf{z}}) \in \{-1, 1\}^n \times \{-1, 1\}^n : H(\mathbf{z}, \tilde{\mathbf{z}}) = k\},$$

where H denotes the Hamming distance. Then the value

$$S_1 := \sum_{(\mathbf{z}, \tilde{\mathbf{z}}) \in M} \|O_{\mathbf{z}}|\psi\rangle - O_{\tilde{\mathbf{z}}}|\psi\rangle\|^2$$

for a normalized state $|\psi\rangle$ satisfies

$$S_1 = O\left(\sqrt{\frac{k}{n}} \binom{n}{k} 2^n\right).$$

Proof. Let the Hilbert space be spanned by $\{|\mathbf{x}, \mathbf{y}\rangle\}$, where $\mathbf{x}$ is the query register and $\mathbf{y}$ is the answer register. Any normalized state can be written as

$$|\psi\rangle = \sum_{\mathbf{x}} |\mathbf{x}\rangle |\psi_{\mathbf{x}}\rangle, \quad \sum_{\mathbf{x}} \|\psi_{\mathbf{x}}\|^2 = 1,$$

where $|\psi_{\mathbf{x}}\rangle$ is the (unnormalized) state on the second register conditional on query $\mathbf{x}$.

For each fixed $\mathbf{x}$ and $\mathbf{z}$, define the unitary on the answer register

$$U_{\mathbf{z}}^{(\mathbf{x})} : |\mathbf{y}\rangle \mapsto |\mathbf{y} \oplus f(\mathbf{x}, \mathbf{z})\rangle.$$

Then

$$O_{\mathbf{z}}|\psi\rangle - O_{\tilde{\mathbf{z}}}|\psi\rangle = \sum_{\mathbf{x}} |\mathbf{x}\rangle (U_{\mathbf{z}}^{(\mathbf{x})} - U_{\tilde{\mathbf{z}}}^{(\mathbf{x})}) |\psi_{\mathbf{x}}\rangle.$$

Let

$$D(\mathbf{z}, \tilde{\mathbf{z}}) := \{\mathbf{x} : f(\mathbf{x}, \mathbf{z}) \neq f(\mathbf{x}, \tilde{\mathbf{z}})\}.$$

If $\mathbf{x} \notin D(\mathbf{z}, \tilde{\mathbf{z}})$ then $U_{\mathbf{z}}^{(\mathbf{x})} = U_{\tilde{\mathbf{z}}}^{(\mathbf{x})}$ and that term vanishes. Therefore

$$\|O_{\mathbf{z}}|\psi\rangle - O_{\tilde{\mathbf{z}}}|\psi\rangle\|^2 = \left\| \sum_{\mathbf{x} \in D(\mathbf{z}, \tilde{\mathbf{z}})} |\mathbf{x}\rangle (U_{\mathbf{z}}^{(\mathbf{x})} - U_{\tilde{\mathbf{z}}}^{(\mathbf{x})}) |\psi_{\mathbf{x}}\rangle \right\|^2 = \sum_{\mathbf{x} \in D(\mathbf{z}, \tilde{\mathbf{z}})} \|(U_{\mathbf{z}}^{(\mathbf{x})} - U_{\tilde{\mathbf{z}}}^{(\mathbf{x})}) |\psi_{\mathbf{x}}\rangle\|^2,$$

Now, for any two unitaries U, V and any vector $|\mathbf{v}\rangle$,

$$\|(U - V)|\mathbf{v}\rangle\| \leq \|U|\mathbf{v}\rangle\| + \|V|\mathbf{v}\rangle\| = 2\|\mathbf{v}\|.$$

Hence,

$$\|(U_{\mathbf{z}}^{(\mathbf{x})} - U_{\tilde{\mathbf{z}}}^{(\mathbf{x})}) |\psi_{\mathbf{x}}\rangle\|^2 \leq 4\|\psi_{\mathbf{x}}\|^2, \quad \text{whenever } f(\mathbf{x}, \mathbf{z}) \neq f(\mathbf{x}, \tilde{\mathbf{z}}).$$

We conclude that for any $\mathbf{z}, \tilde{\mathbf{z}}$,

$$\|O_{\mathbf{z}}|\psi\rangle - O_{\tilde{\mathbf{z}}}|\psi\rangle\|^2 \leq 4 \sum_{\mathbf{x}: f(\mathbf{x}, \mathbf{z}) \neq f(\mathbf{x}, \tilde{\mathbf{z}})} \|\psi_{\mathbf{x}}\|^2. \quad (29)$$

Using (29),

$$\begin{aligned} S_1 &\leq 4 \sum_{(\mathbf{z}, \tilde{\mathbf{z}}) \in M} \sum_{\mathbf{x}: f(\mathbf{x}, \mathbf{z}) \neq f(\mathbf{x}, \tilde{\mathbf{z}})} \|\psi_{\mathbf{x}}\|^2 \\ &= 4 \sum_{\mathbf{x}} \|\psi_{\mathbf{x}}\|^2 \underbrace{|\{(\mathbf{z}, \tilde{\mathbf{z}}) \in M : f(\mathbf{x}, \mathbf{z}) \neq f(\mathbf{x}, \tilde{\mathbf{z}})\}|}_{=: N_{\mathbf{x}}}. \end{aligned}$$

Therefore,

$$S_1 \leq 4 \sum_{\mathbf{x}} \|\psi_{\mathbf{x}}\|^2 N_{\mathbf{x}} \leq 4 \left(\max_{\mathbf{x}} N_{\mathbf{x}} \right) \sum_{\mathbf{x}} \|\psi_{\mathbf{x}}\|^2 = 4 \max_{\mathbf{x}} N_{\mathbf{x}},$$

using $\sum_{\mathbf{x}} \|\psi_{\mathbf{x}}\|^2 = 1$. Hence, the problem reduces to bounding $N_{\mathbf{x}}$, where for fixed $\mathbf{x} \in \mathbb{Z}^n$ we define

$$g_{\mathbf{x}}(\mathbf{z}) := f(\mathbf{x}, \mathbf{z}) = \mathbf{1}[\langle \mathbf{x}, \mathbf{z} \rangle \geq 0], \quad \mathbf{z} \in \{-1, 1\}^n,$$

and

$$N_{\mathbf{x}} = |\{(\mathbf{z}, \tilde{\mathbf{z}}) \in M : g_{\mathbf{x}}(\mathbf{z}) \neq g_{\mathbf{x}}(\tilde{\mathbf{z}})\}|.$$

Note that the function $g_{\mathbf{x}}$ is a *halfspace* (linear threshold function) on the hypercube:

$$g_{\mathbf{x}}(\mathbf{z}) = \mathbf{1}\left[\sum_{i=1}^n w_i z_i \geq 0\right], \quad w_i := x_i \in \mathbb{Z}.$$

Applying Lemma 8 to $g_{\mathbf{x}}$, we obtain

$$N_{\mathbf{x}} = \binom{n}{k} 2^{n-1} \mathbb{E}_{S: |S|=k} \Pr[g_{\mathbf{x}}(Z) \neq g_{\mathbf{x}}(Z^{\oplus S})] \leq C_0 \sqrt{\frac{k}{n}} \binom{n}{k} 2^{n-1}.$$

Recalling that

$$S_1 \leq 4 \max_x N_{\mathbf{x}},$$

we conclude

$$S_1 \leq 4 \cdot C_0 \sqrt{\frac{k}{n}} \binom{n}{k} 2^{n-1} = 2C_0 \sqrt{\frac{k}{n}} \binom{n}{k} 2^n.$$

Thus there is an absolute constant $C := 2C_0 > 0$ such that

$$S_1 \leq C \sqrt{\frac{k}{n}} \binom{n}{k} 2^n = O\left(\sqrt{\frac{k}{n}} \binom{n}{k} 2^n\right).$$

□

D.2.3 Proof of Theorem 6

Now, we give the proof of Theorem 6. This is established by the following theorem:

Theorem 9. *Let $\mathbf{z} \in \{-1, 1\}^n$ be an unknown string and we have oracle $O_{\mathbf{z}}|\mathbf{x}, \mathbf{y}\rangle = |\mathbf{x}, \mathbf{y} \oplus f(\mathbf{x}, \mathbf{z})\rangle$, where $\mathbf{x} \in \mathbb{Z}^n$ and $f(\mathbf{x}, \mathbf{z}) = 1$ when $\langle \mathbf{x}, \mathbf{z} \rangle \geq 0$, $f(\mathbf{x}, \mathbf{z}) = 0$ when $\langle \mathbf{x}, \mathbf{z} \rangle < 0$. Then every algorithm output $\mathbf{z}$ within k errors uses at least $\frac{1}{4} \log \frac{n}{k}$ queries to $O_{\mathbf{z}}$.*

Proof. Define $M = \{(\mathbf{x}, \tilde{\mathbf{x}}) \mid H(\mathbf{x}, \tilde{\mathbf{x}}) = k\}$ and we can get $|M| = \binom{n}{k} \cdot 2^n$. By Lemma 9, we have

$$S_1 := \sum_{(\mathbf{x}, \tilde{\mathbf{x}}) \in M} \|O_{\mathbf{x}}|\psi\rangle - O_{\tilde{\mathbf{x}}}|\psi\rangle\|^2 \leq C \sqrt{\frac{k}{n}} \binom{n}{k} 2^n.$$

We hope to get a bound of

$$S_T := \sum_{(\mathbf{x}, \tilde{\mathbf{x}}) \in M} \|O_{\mathbf{x}}U_{T-1}O_{\mathbf{x}} \cdots U_1O_{\mathbf{x}}|\psi\rangle - O_{\tilde{\mathbf{x}}}U_{T-1}O_{\tilde{\mathbf{x}}} \cdots U_1O_{\tilde{\mathbf{x}}}|\psi\rangle\|^2.$$

In fact, we have

$$\sum_{(\mathbf{x}, \tilde{\mathbf{x}}) \in M} \|O_{\mathbf{x}}U_1O_{\mathbf{x}}|\psi\rangle - O_{\tilde{\mathbf{x}}}U_1O_{\tilde{\mathbf{x}}}|\psi\rangle\|^2 \leq 2 \sum_{(\mathbf{x}, \tilde{\mathbf{x}}) \in M} \|(O_{\mathbf{x}} - O_{\tilde{\mathbf{x}}})U_1O_{\mathbf{x}}|\psi\rangle\|^2 + \|O_{\mathbf{x}}|\psi\rangle - O_{\tilde{\mathbf{x}}}|\psi\rangle\|^2.$$

Assume that $|\psi\rangle = \sum_i \alpha_i |\psi_i\rangle |\phi_i\rangle$. Since our oracle can only judge, which means for different $\mathbf{x}$, $O_{\mathbf{x}}|\psi_i\rangle |\phi_i\rangle$ only have two results $|\psi_i\rangle |\phi_{i,0}\rangle$ and $|\psi_i\rangle |\phi_{i,1}\rangle$. So we can get

$$\begin{aligned} \sum_{(\mathbf{x}, \tilde{\mathbf{x}}) \in M} \|(O_{\mathbf{x}} - O_{\tilde{\mathbf{x}}})U_1O_{\mathbf{x}}|\psi\rangle\|^2 &\leq \sum_{(\mathbf{x}, \tilde{\mathbf{x}}) \in M} \sum_i \alpha_i^2 \|(O_{\mathbf{x}} - O_{\tilde{\mathbf{x}}})U_1O_{\mathbf{x}}|\psi_i\rangle |\phi_i\rangle\|^2 \\ &\leq \sum_{(\mathbf{x}, \tilde{\mathbf{x}}) \in M} \sum_i \alpha_i^2 (\|(O_{\mathbf{x}} - O_{\tilde{\mathbf{x}}})U_1|\psi_i\rangle |\phi_{i,0}\rangle\|^2 + \|(O_{\mathbf{x}} - O_{\tilde{\mathbf{x}}})U_1|\psi_i\rangle |\phi_{i,1}\rangle\|^2) \\ &= \sum_i \alpha_i^2 \sum_{(\mathbf{x}, \tilde{\mathbf{x}}) \in M} \|(O_{\mathbf{x}} - O_{\tilde{\mathbf{x}}})U_1|\psi_i\rangle |\phi_{i,0}\rangle\|^2 + \|(O_{\mathbf{x}} - O_{\tilde{\mathbf{x}}})U_1|\psi_i\rangle |\phi_{i,1}\rangle\|^2 \\ &\leq \sum_i 2\alpha_i^2 S_1 = 2S_1. \end{aligned}$$

Similarly, we can get

$$S_T \leq (2^{T+1} + 2)S_1 \leq (2^{T+1} + 2) \cdot \sqrt{\frac{k}{n}} \binom{n}{k} 2^n.$$

We let $T = \frac{1}{4} \log \frac{n}{k}$, and then there exist a $(\mathbf{x}, \tilde{\mathbf{x}}) \in M$ such that

$$\|O_{\mathbf{x}}U_{T-1}O_{\mathbf{x}} \cdots U_1O_{\mathbf{x}}|\psi\rangle - O_{\tilde{\mathbf{x}}}U_{T-1}O_{\tilde{\mathbf{x}}} \cdots U_1O_{\tilde{\mathbf{x}}}|\psi\rangle\|^2 \leq \frac{1}{(n/k)^{\frac{1}{4}}}.$$

□

Note that we can reduce [Problem 2](#) to the special case of assuming $\nabla f \in \{-1, 1\}^n$, and the result allows εn -bit error. By [Theorem 9](#), set $k = \varepsilon n$, we obtain the quantum lower bound $\Omega(\log(1/\varepsilon))$ for [Theorem 6](#).