Quantum Speedups for Stochastic Optimization with Heavy-Tailed Noise

Bin Luo
The Chinese University of Hong Kong
binluo@link.cuhk.edu.hk

Chengchang Liu*
The Chinese University of Hong Kong
7liuchengchang@gmail.com

Jonathan Allcock
Tencent Quantum Laboratory
jonallcock@tencent.com

Shengyu Zhang
Tencent Quantum Laboratory
shengyzhang@tencent.com

John C.S. Lui
The Chinese University of Hong Kong
cslui@cse.cuhk.edu.hk

Abstract

We study stochastic optimization with heavy-tailed gradient noise. We first propose a novel quantum mean estimator for multivariate heavy-tailed random variables that achieves lower query complexity than optimal classical estimators in the low-dimensional regime. We further develop an unbiased quantum mean estimator by applying a generalized multilevel Monte Carlo technique. We prove quantum lower bounds showing that, when the dimension d of the random vector is small and can be viewed as a constant, our quantum estimators are optimal up to logarithmic factors. These lower bounds also show that no quantum speedup is available compared to optimal classical mean estimators in the high-dimensional, low-precision regime. Based on these estimators, we propose a quantum normalized stochastic gradient descent method (QNSGD), which finds an ϵ-stationary point of a nonconvex objective using $\tilde{\mathcal{O}}\big(d^{\frac{p}{4(p-1)}}\epsilon^{-\frac{5p-4}{2p-2}}\big)$ queries to the stochastic gradient, where $p \in (1,2]$ is the tail index. For a convex objective function, we propose a quantum projected stochastic gradient descent method (QPSGD), which computes an ϵ-optimal solution with query complexity $\tilde{\mathcal{O}}\big(d^{\frac{p}{4(p-1)}}\epsilon^{-\frac{3p-2}{2p-2}}\big)$. Our results improve upon the classical lower bounds $\Omega\big(\epsilon^{-\frac{3p-2}{p-1}}\big)$ for nonconvex problems and $\Omega\big(\epsilon^{-\frac{p}{p-1}}\big)$ for convex problems, demonstrating a quantum speedup when d is relatively small.

1 Introduction

We consider the stochastic optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}\subseteq\mathbb{R}^d} f(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathbb{P}}\left[F(\mathbf{x};\xi)\right], \tag{1}$$

where f may be non-convex. Here, $F(\mathbf{x};\xi)$ is a random loss function whose expectation equals $f(\mathbf{x})$.

Traditionally, convergence analyses of stochastic first-order methods rely on the bounded-variance assumption for stochastic gradients [19]. However, empirical evidence from image classification [47], large language models [1, 58], and reinforcement learning [18] indicates that stochastic gradients

*Chengchang Liu is the corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

often exhibit heavy-tailed behavior, thereby violating the bounded-variance assumption. Moreover, the hardness example in [58] shows that standard stochastic gradient descent (SGD) with norm-agnostic step-size rules may fail to converge under heavy-tailed gradient noise. Motivated by these observations, we study the stochastic optimization problem (1) under a heavy-tailed stochastic first-order oracle, which we formalize below.

Assumption 1.1. *We assume the stochastic gradient $\nabla F(\mathbf{x};\xi)$ follows the heavy-tailed distribution such that $\mathbb{E}_{\xi\sim\mathbb{P}}[\nabla F(\mathbf{x};\xi)] = \nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$ and*

$$\mathbb{E}_{\xi\sim\mathbb{P}}\left[\|\nabla F(\mathbf{x};\xi) - \nabla f(\mathbf{x})\|^p\right] \le \sigma^p, \tag{2}$$

for some $\sigma > 0$ and $p \in (1, 2]$.

Inspired by the importance of stochastic optimization under heavy-tailed gradient noise, numerous classical methods have been proposed in recent years. [58] showed that SGD with gradient clipping can find an ϵ-stationary point in expectation. Subsequent works [12, 41, 45] developed refined clipping-based algorithms that establish high-probability convergence guarantees. In addition, [28] proved that normalized SGD achieves the same oracle complexity as gradient clipping in the non-convex setting. More recently, [35] studied online convex optimization with heavy-tailed noise and employed the online-to-nonconvex (O2NC) framework [13] to recover optimal oracle complexity. Despite these advances, under the stochastic first-order oracle model, all existing methods are subject to the lower bounds $\Omega\big(\epsilon^{-\frac{3p-2}{p-1}}\big)$ for non-convex optimization and $\Omega\big(\epsilon^{-\frac{p}{p-1}}\big)$ for convex optimization, as established in [44, 58].

Recently, quantum computing has demonstrated provable speedups for a broad range of optimization problems, including convex [6, 30, 51, 59], non-convex [21, 46], non-smooth [6, 31, 33, 43], and minimax optimization [17, 32, 34, 53]. In particular, under the bounded-variance setting, [46] showed quantum speedups for stochastic optimization by using quantum mean estimators [10] to construct gradient estimators with a smaller batch size.

However, it remains an open question *whether one can demonstrate a quantum speedup on stochastic optimization with heavy-tailed noise*. First, **no** quantum mean estimation algorithms are currently available for multivariate random variables with heavy-tailed distributions. Although [4, 56] proposed a quantum mean estimator for heavy-tailed data, their method applies only to univariate random variables. For the multivariate case, all existing quantum mean estimators are designed for the bounded-variance setting [10, 46, 49]. However, these estimators may not apply to the heavy-tailed distributions with only a bounded p-th central moment for $p \in (1, 2)$, because the variance of such a heavy-tailed random vector need not exist or be finite. Second, most existing classical algorithms use only a single stochastic gradient per iteration [11, 16, 35, 45, 58], which limits the applicability of quantum mean estimators to improve the sample complexity.

In this paper, we address this open question by developing novel quantum mean estimators for multivariate heavy-tailed random variables and leveraging them to design quantum stochastic optimization algorithms. We summarize our contributions as follows.

- We propose a novel quantum heavy-tailed mean estimator, termed QHTME, for estimating the mean of a random vector $\mathbf{X}$ satisfying $\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|^p\right] \le \sigma^p$ for some $\sigma > 0$ and $p \in (1, 2]$, given access to a quantum sampling oracle (cf. Definition 2.2). QHTME generalizes the multivariate bounded-variance estimators in [10, 49] and the univariate heavy-tailed estimator in [4], corresponding to the special cases $p = 2$ and $d = 1$, respectively. Moreover, QHTME achieves a provable quantum speedup over optimal classical heavy-tailed mean estimators in the low-dimensional regime. Unlike the quantum estimator in [4], which requires a known bound on $\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|^p\right]$ as an input, our QHTME avoids this requirement. A detailed comparison with existing classical and quantum estimators is given in Table 1.
- We further refine QHTME to develop a quantum unbiased heavy-tailed mean estimator (QUHTME). This estimator extends the unbiased quantum mean estimator in [46] from the bounded-variance setting to general heavy-tailed distributions ($p \in (1, 2]$). A comparison with existing quantum mean estimators is provided in Table 2.
- We establish quantum lower bounds for obtaining an ϵ-accurate estimate of the mean of a d-dimensional random variable $\mathbf{X}$ with heavy-tailed distribution in Section 3.3. These lower bounds show that our quantum estimators QHTME and QUHTME are optimal in their dependence on σ and ϵ, up to logarithmic factors when d is small. Moreover, they imply

that no provable quantum speedup over the optimal classical heavy-tailed mean estimator is possible in the high-dimensional, low-precision regime.

- For non-convex stochastic optimization, we propose a quantum normalized stochastic gradient descent method, `QNSGD`, which integrates `QHTME` with a gradient normalization technique. Under heavy-tailed noise, `QNSGD` finds an ϵ-stationary point using $\tilde{\mathcal{O}}(d^{\frac{p}{4(p-1)}}\epsilon^{-\frac{5p-4}{2p-2}})$ queries to the quantum stochastic gradient oracle (cf. Definition 2.4). In the low-dimensional regime, this result improves upon the classical lower bound $\Omega(\epsilon^{-\frac{3p-2}{p-1}})$ by a factor of $\epsilon^{-\frac{p}{2(p-1)}}$. A comparison with existing classical and quantum methods is provided in Table 3.
- For convex stochastic optimization, we propose a quantum projected stochastic gradient descent method, `QPSGD`, which leverages the unbiased quantum estimator `QUHTME`. Under heavy-tailed noise, `QPSGD` obtains an ϵ-optimal solution using $\tilde{\mathcal{O}}(d^{\frac{p}{4(p-1)}}\epsilon^{-\frac{3p-2}{2(p-1)}})$ expected queries to the quantum stochastic gradient oracle. When d is relatively small, this result breaks the classical lower bound $\Omega(\epsilon^{-\frac{p}{p-1}})$ for $p \in (1,2)$. A comparison with existing classical methods is given in Table 4.

While our results are primarily theoretical, we briefly discuss their practicality and potential applications in Appendix B.

Table 1: Comparison of estimators for producing an estimate $\tilde{\boldsymbol{\mu}}$ of the mean $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}]$ of a d-dimensional random vector $\mathbf{X}$ satisfying $\mathbb{E}[\|\mathbf{X}-\boldsymbol{\mu}\|^p] \le \sigma^p$ such that $\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\| \le \epsilon$.

	References	Assumption on p	Dimension d	Queries	Bias
Classical	[36]	$p=2$	$d \ge 1$	$\mathcal{O}((\sigma/\epsilon)^2)$	0
	Theorem 2.7	$p \in (1,2]$	$d \ge 1$	$\tilde{\mathcal{O}}((\sigma/\epsilon)^{\frac{p}{p-1}})$	ϵ
	[36]	$p \in (1,2]$	$d \ge 1$	$\Omega((\sigma/\epsilon)^{\frac{p}{p-1}})$	–
Quantum	[49]	$p=2$	$d \ge 1$	$\tilde{\mathcal{O}}(d^{\frac{1}{2}}\sigma/\epsilon)$	ϵ
	[4]	$p \in (1,2]$	$d=1$	$\tilde{\mathcal{O}}((\sigma/\epsilon)^{\frac{p}{2(p-1)}})$ †	ϵ
	Theorem 3.4	$p \in (1,2]$	$d \ge 1$	$\tilde{\mathcal{O}}((d^{\frac{1}{2}}\sigma/\epsilon)^{\frac{p}{2(p-1)}})$	ϵ
	[10]	$p=2$	$1 \le d \le (\sigma/\epsilon)^2$	$\Omega(d^{\frac{1}{2}}\sigma\epsilon^{-1})$	–
			$d \ge (\sigma/\epsilon)^2$	$\Omega(\sigma^2\epsilon^{-2})$	–
	Theorem 3.9	$p \in (1,2]$	$1 \le d \le (\sigma/\epsilon)^{\frac{2}{p-1}}$	$\Omega((\sigma/\epsilon)^{\frac{p}{2(p-1)}})$	–
			$d \ge (\sigma/\epsilon)^{\frac{2}{p-1}}$	$\Omega((\sigma/\epsilon)^{\frac{p}{p-1}})$‡	–

† These results require prior knowledge of σ.
‡ This lower bound is established under the low-precision condition stated in Theorem 3.9.

Table 2: Comparison of estimators for producing an estimate $\tilde{\boldsymbol{\mu}}$ of the mean $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}]$ of a d-dimensional random vector $\mathbf{X}$ satisfying $\mathbb{E}[\|\mathbf{X}-\boldsymbol{\mu}\|^p] \le \sigma^p$ such that $\mathbb{E}[\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^p] \le \epsilon^p$.

	References	Assumption on p	Dimension d	Queries	Bias
Classical	Empirical Mean	$p \in (1,2]$	$d \ge 1$	$\tilde{\mathcal{O}}((\sigma/\epsilon)^{\frac{p}{p-1}})$	0
	[36]	$p \in (1,2]$	$d \ge 1$	$\Omega((\sigma/\epsilon)^{\frac{p}{p-1}})$	–
Quantum	[8]	$p=2$	$d=1$	$\tilde{\mathcal{O}}(\sigma/\epsilon)$	$\delta\sigma$
	[46]	$p=2$	$d \ge 1$	$\tilde{\mathcal{O}}(d^{\frac{1}{2}}\sigma/\epsilon)$	0
	Theorem 3.7	$p \in (1,2]$	$d \ge 1$	$\tilde{\mathcal{O}}((d^{\frac{1}{2}}\sigma/\epsilon)^{\frac{p}{2(p-1)}})$	0
	[46]	$p=2$	$1 \le d \le (\sigma/\epsilon)^2$	$\Omega(d^{\frac{1}{2}}\sigma\epsilon^{-1})$	–
	Theorem 3.9	$p \in (1,2]$	$1 \le d \le (\sigma/\epsilon)^{\frac{2}{p-1}}$	$\Omega((\sigma/\epsilon)^{\frac{p}{2(p-1)}})$	–
			$d \ge (\sigma/\epsilon)^{\frac{2}{p-1}}$	$\Omega((\sigma/\epsilon)^{\frac{p}{p-1}})$‡	–

‡ This lower bound is established under the low-precision condition stated in Theorem 3.9.

Table 3: Comparison of different methods to find an ϵ-stationary point, i.e. $\mathbb{E}\left[\|\nabla f(\mathbf{x})\|\right] \leq \epsilon$ in stochastic nonconvex optimization. p is the index of the heavy-tailed assumption on the stochastic gradient in Assumption 1.1.

	References	Assumption on p	Oracle Complexity
Classical	[19]	$p = 2$	$\mathcal{O}(\epsilon^{-4})$
	[28, 35, 58]	$p \in (1, 2]$	$\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$
	Lower Bound [58]	$p \in (1, 2]$	$\Omega\big(\epsilon^{-\frac{3p-2}{p-1}}\big)$
Quantum	[46]	$p = 2$	$\tilde{\mathcal{O}}(d^{\frac{1}{2}}\epsilon^{-3})$
	Theorem 4.4	$p \in (1, 2]$	$\tilde{\mathcal{O}}(d^{\frac{p}{4(p-1)}}\epsilon^{-\frac{5p-4}{2p-2}})$

Table 4: Comparison of different methods to find an ϵ optimal solution, i.e. $\mathbb{E}\left[f(\mathbf{x}) - f^*\right] \leq \epsilon$ in stochastic convex optimization. p is the index in Assumption 1.1.

	References	Assumption on p	Oracle Complexity
Classical	[19]	$p = 2$	$\mathcal{O}(\epsilon^{-2})$
	[16, 35]	$p \in (1, 2]$	$\mathcal{O}(\epsilon^{-\frac{p}{p-1}})$
	Lower Bound [44]	$p \in (1, 2]$	$\Omega(\epsilon^{-\frac{p}{p-1}})$
Quantum	**Theorem 5.3**	$p \in (1, 2]$	$\tilde{\mathcal{O}}(d^{\frac{p}{4(p-1)}}\epsilon^{-\frac{3p-2}{2(p-1)}})$

2 Preliminaries

We use $\|\cdot\|$ to denote the spectral norm for matrices and the Euclidean norm for vectors, respectively. We denote by $\mathrm{Geom}(\rho)$ the geometric distribution with parameter ρ, and by $\Pi_{\mathcal{X}}(\cdot)$ the projection operator onto the set $\mathcal{X}$. $\tilde{\mathcal{O}}(\cdot)$ denotes the big-$\mathcal{O}$ notation that suppresses polylogarithmic factors in d, δ, n and ϵ. All logarithms are taken in base 2.

The quantum access models are expressed using the Dirac notation of quantum computing [14, 42, 55]. We use $\{|i\rangle\}_{i=0}^{n-1}$ to denote the computational basis of $\mathbb{C}^n$, where $|0\rangle$ denotes the first standard basis vector. A normalized vector $\mathbf{v} \in \mathbb{C}^n$ can be written as a quantum state $|\mathbf{v}\rangle = \sum_{i=0}^{n-1} v_i |i\rangle$.

Throughout this paper, real numbers are encoded in the computational basis using a fixed-point binary representation. Specifically, a real number x is encoded as a basis state $|x\rangle = |x_1 \dots x_q\rangle \in \mathbb{C}^{2^q}$, where the binary string $x_1 \dots x_q$ corresponds to a fixed-point encoding of x. Similarly, for $\mathbf{x} \in \mathbb{R}^d$, the notation $|\mathbf{x}\rangle$ denotes the computational-basis encoding of the coordinates of $\mathbf{x}$, rather than an amplitude encoding of the vector $\mathbf{x}$. We assume access to sufficiently many qubits to avoid overflow for all real numbers appearing in the algorithms.

With these conventions, we first formalize the classical and quantum access models for a random variable $\mathbf{X} \in \mathbb{R}^d$.

Definition 2.1 (Classical sampling oracle)**.** Given a random variable $\mathbf{X} : \Omega \to E \subseteq \mathbb{R}^d$ on a probability space $(\Omega, 2^\Omega, \mathbb{P})$, we define a classical sampling oracle as the process of drawing a sample $\omega \in \Omega$ according to $\mathbb{P}$ and observing the value of $\mathbf{X}(\omega) \in \mathbb{R}^d$.

Definition 2.2 (Quantum sampling oracle)**.** Let $\mathbf{X}$ be a d-dimensional finite-valued random variable taking values in $E \subset \mathbb{R}^d$, and let $p_{\mathbf{X}}(\mathbf{x}) := \Pr_{\omega\sim\mathbb{P}}[\mathbf{X}(\omega) = \mathbf{x}]$ for each $\mathbf{x} \in E$. A quantum sampling oracle for $\mathbf{X}$ is a unitary operator $O_{\mathbf{X}}$ satisfying

$$O_{\mathbf{X}} : |0\rangle |0\rangle \to \sum_{\mathbf{x}\in E} \sqrt{p_{\mathbf{X}}(\mathbf{x})}\ |\mathbf{x}\rangle |\psi_{\mathbf{x}}\rangle , \tag{3}$$

where $|0\rangle$ denotes fixed reference states, and each $|\psi_{\mathbf{x}}\rangle$ is a normalized auxiliary state.

Note that measuring the first register of $O_{\mathbf{X}} |0\rangle |0\rangle$ in the computational basis yields a classical sample distributed as $\mathbf{X}$, so the classical sampling oracle can be simulated by $O_{\mathbf{X}}$. Although Definition 2.2 is stated for finite-valued random variables, it should be viewed as a finite-precision representation. For

heavy-tailed distributions, this corresponds to a suitably truncated and discretized approximation. We also note that prior work [10, 46] adopts different notions of quantum sampling oracle. A comparison and a discussion of compatibility with heavy-tailed assumptions are deferred to Appendix A.1.

We now define the classical and quantum access models to the stochastic gradient $\nabla F(\mathbf{x};\xi)$ satisfying Assumption 1.1.

Definition 2.3 (Classical stochastic gradient oracle)**.** We define a classical stochastic gradient oracle as the process of implementing a random function that, when queried at $\mathbf{x}$, samples a random seed ξ from a probability distribution $\mathbb{P}$ and returns $\nabla F(\mathbf{x};\xi)$ satisfying Assumption 1.1.

Definition 2.4 (Quantum stochastic gradient oracle)**.** Let $\nabla F(\mathbf{x};\xi)$ be the stochastic gradient at a query point $\mathbf{x}$, and let $p_{\mathbf{x}}(\mathbf{v}) := \Pr_{\xi\sim\mathbb{P}}[\nabla F(\mathbf{x};\xi) = \mathbf{v}]$, where $\mathbf{v} \in E$ and $E \subset \mathbb{R}^d$ is the finite range of $\nabla F(\mathbf{x};\xi)$. A quantum stochastic gradient oracle at $\mathbf{x}$ is a unitary operator $O_{\nabla F,\mathbf{x}}$ satisfying

$$O_{\nabla F,\mathbf{x}} : |0\rangle\,|0\rangle \to \sum_{v\in E} \sqrt{p_{\mathbf{x}}(\mathbf{v})}\;|\mathbf{v}\rangle\,|\psi_{\mathbf{x},\mathbf{v}}\rangle\,,$$

where $|0\rangle$ denotes fixed reference states, and each $|\psi_{\mathbf{x},\mathbf{v}}\rangle$ is a normalized auxiliary state.

We note that [46] also defined a quantum stochastic gradient oracle, but their model requires coherent access to the query point $\mathbf{x}$. A detailed comparison is deferred to Appendix A.2.

Below, we define the notion of a quantile and introduce a quantum algorithm for estimating it, which will be used as a subroutine in our subsequent algorithm.

Definition 2.5 (Quantile)**.** Let $\mathbf{X}$ be a finite-valued real random variable and let $\tau \in (0,1]$. The quantile of order τ is defined as $Q(\tau) := \sup\{x \in \mathbb{R} : \Pr[\mathbf{X} \geq x] \geq \tau\}$.

Lemma 2.6 (Theorem 3.4, [23])**.** *Let $\mathbf{X}$ be a univariate random variable. Given $\tau, \delta \in (0,1)$, the quantile estimation algorithm $\texttt{Quantile}(\mathbf{X}, \tau, \delta)$ outputs an approximate quantile $\tilde{Q}$ such that $Q(\tau) \leq \tilde{Q} \leq Q(c\tau)$ with probability at least $1-\delta$, where $c \in (0,1)$ is a universal constant. The algorithm uses $\mathcal{O}\left(\frac{\log(1/\delta)}{\sqrt{\tau}}\right)$ queries to the quantum sampling oracle $O_{\mathbf{X}}$.*

We now present the optimal classical heavy-tailed mean estimator. Related work on classical heavy-tailed mean estimation, together with the proof of Theorem 2.7, is provided in Appendix C.1.

Theorem 2.7 (Classical heavy-tailed multivariate mean estimation)**.** *Let $\mathbf{X}$ be a d-dimensional random variable with mean $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}]$ and finite p-th central moment $\mathbb{E}[\|\mathbf{X}-\boldsymbol{\mu}\|^p]$ for some $p \in (1,2]$. Let $\delta \in (0,1)$ and $k = \lceil 24\ln(1/\delta)\rceil$. For an integer n such that $n \geq 2k$, with probability at least $1-\delta$, the classical heavy-tailed mean estimator $\texttt{CHTME}(\mathbf{X}, n, \delta)$ (Algorithm 5) outputs a mean estimate $\tilde{\boldsymbol{\mu}}$ such that*

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq 3(8C_p)^{1/p}\left(\mathbb{E}[\|\mathbf{X}-\boldsymbol{\mu}\|^p]\right)^{1/p}\left(\frac{2\lceil 24\ln(1/\delta)\rceil}{n}\right)^{\frac{p-1}{p}},$$

by using n queries to the classical sampling oracle, where C_p is a p-dependent constant.

3 Quantum Mean Estimation for Heavy-Tailed Random Variables

In this section, we develop novel quantum mean estimators for d-dimensional heavy-tailed random variables. We begin by formalizing the estimation tasks of interest.

Problem 3.1. *Let $\mathbf{X}$ be a d-dimensional random variable with mean $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}]$ satisfying $\mathbb{E}[\|\mathbf{X}-\boldsymbol{\mu}\|^p] \leq \sigma^p$ for some $p \in (1,2]$, and some constant $\sigma > 0$. For any $\epsilon > 0$, the goal is to output an estimate $\tilde{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ satisfying $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \epsilon$.*

We also consider the construction of unbiased estimators whose estimation error has a controlled p-th raw moment.

Problem 3.2. *Let $\mathbf{X}$ be a d-dimensional random variable with mean $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}]$ satisfying $\mathbb{E}[\|\mathbf{X}-\boldsymbol{\mu}\|^p] \leq \sigma^p$ for some $p \in (1,2]$, and some constant $\sigma > 0$. For any $\epsilon > 0$, the goal is to output an **unbiased** estimate $\tilde{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ satisfying $\mathbb{E}[\tilde{\boldsymbol{\mu}}] = \boldsymbol{\mu}$ and $\mathbb{E}[\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^p] \leq \epsilon^p$.*

Algorithm 1 Quantum Heavy-Tailed Mean Estimator ($\texttt{QHTME}(\mathbf{X}, n, p, d, \delta)$)

Input: random variable $\mathbf{X}$, query parameter n, tail index p, dimension d, failure probability δ.

$k \leftarrow \lceil 2\log n \rceil$, $n' \leftarrow \left\lceil n \cdot \frac{2^{1/p}\gamma_p \log(10kd/\delta)}{c^{1/p}\log(2d/\delta)} \right\rceil$ and $n'' \leftarrow \left\lceil 2\lceil 24\ln(2/\delta)\rceil 3^{\frac{p}{p-1}} (8C_p)^{\frac{1}{p-1}} \right\rceil$, where

$$\gamma_p \leftarrow \begin{cases} \left(1 - 2^{-(1/p - 1/2)}\right)^{-1}, & p \in (1,2), \\ k, & p = 2, \end{cases}$$

$\boldsymbol{\zeta} \leftarrow \texttt{CHTME}(\mathbf{X}, n'', \delta/2)$, $a_0 \leftarrow 0$, and define $\mathbf{Y} \leftarrow \mathbf{X} - \boldsymbol{\zeta}$

for $\ell = 1, \ldots, k$ **do**

  $a_\ell \leftarrow \max\left\{a_{\ell-1}, \texttt{Quantile}\left(\|\mathbf{Y}\|, 2^{-\ell}, \delta/(10k)\right)\right\}$

  $$\tilde{\boldsymbol{\mu}}_\ell \leftarrow \begin{cases} 0, & a_{\ell-1} = a_\ell, \\ \texttt{QBounded}_d\left(\frac{\mathbf{Y} \cdot \mathbf{1}\{a_{\ell-1} < \|\mathbf{Y}\| \le a_\ell\}}{a_\ell}, 2^{-(\ell-1)}, n', \frac{\delta}{10k}\right), & a_{\ell-1} < a_\ell. \end{cases}$$

end for

Output: $\tilde{\boldsymbol{\mu}} \leftarrow \boldsymbol{\zeta} + \sum_{\ell=1}^{k} a_\ell \tilde{\boldsymbol{\mu}}_\ell$.

3.1 Quantum Heavy-Tailed Mean Estimator

Here we propose a quantum mean estimator to solve Problem 3.1. We first introduce an existing quantum mean estimator for bounded random variables.

Lemma 3.3 (Theorem 3.3, [10])**.** *Let $\mathbf{X}$ be a d-dimensional random variable such that $\|\mathbf{X}\| \le 1$. Given $L \in (0,1], \delta \in (0,1)$ and $n \ge 1$ such that $\mathbb{E}[\|\mathbf{X}\|] \le L$, Algorithm 1 of [10], denoted by $\texttt{QBounded}_d(\mathbf{X}, L, n, \delta)$, outputs an estimate $\tilde{\boldsymbol{\mu}}$ of $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ satisfying $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \le \frac{\sqrt{L}\log(d/\delta)}{n}$ with probability at least $1-\delta$, by using $\tilde{\mathcal{O}}(n)$ queries to the quantum sampling oracle $O_{\mathbf{X}}$.*

To estimate $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}]$ under a heavy-tailed distribution, the quantum estimator $\texttt{QBounded}$ in Lemma 3.3 cannot be applied directly, as the random variable $\mathbf{X}$ is not necessarily bounded. Similar to other mean estimators [10, 24, 25], we decompose $\mathbf{X}$ into a sequence of truncated random variables $\mathbf{X}'_\ell := \mathbf{X} \cdot \mathbf{1}\{a_{\ell-1} < \|\mathbf{X}\| \le a_\ell\}$, for $\ell = 1, \ldots, k$, with $a_0 = 0$. We normalize them as $\mathbf{X}_\ell := \mathbf{X}'_\ell / a_\ell$. The quantum estimator $\texttt{QBounded}$ is then applied to each $\mathbf{X}_\ell$. A key challenge is that, unlike the univariate mean estimator for bounded random variables in [40], $\texttt{QBounded}$ requires an explicit upper bound on $\mathbb{E}[\|\mathbf{X}_\ell\|]$. Inspired by [10], we address this issue by determining the thresholds a_ℓ via the quantile estimator in Lemma 2.6. Specifically, a_ℓ are chosen to approximate quantiles of order $2^{-\ell}$, so that $\Pr(\|\mathbf{X}\| \ge a_\ell) \approx 2^{-\ell}$. This construction yields the bound $\mathbb{E}[\|\mathbf{X}_\ell\|] \le 2^{-(\ell-1)}$. Another challenge is that the resulting error bound depends on the p-th raw moment $\mathbb{E}[\|\mathbf{X}\|^p]$ rather than the central moment $\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]$ in this way. To resolve this, we first compute a rough estimate $\boldsymbol{\zeta}$ of $\boldsymbol{\mu}$ using the optimal classical heavy-tailed mean estimator in Theorem 2.7, incurring only a logarithmic increase in query complexity. We formalize the resulting procedure in Algorithm 1 and state its performance guarantees in Theorems C.5 and 3.4, whose proofs are deferred to Appendix C.2.

Theorem 3.4. *Algorithm 1 ($\texttt{QHTME}$) solves Problem 3.1 using $n(\epsilon) := \tilde{\mathcal{O}}\left(d^{\frac{p}{4(p-1)}} \sigma^{\frac{p}{2(p-1)}} \epsilon^{-\frac{p}{2(p-1)}}\right)$ queries to the quantum sampling oracle $O_{\mathbf{X}}$ with probability at least $1-\delta$.*

Remark 3.5. $\texttt{QHTME}$ applies to d-dimensional random vectors ($d \ge 1$) with bounded p-th central moment ($p \in (1,2]$), which generalizes the univariate setting considered in [4, Section 5.2] and the bounded-variance setting considered in [10, Theorem 3.4]. Moreover, unlike the method in [4], which requires prior knowledge of an upper bound on $\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]$ as an input, our method does not.

Remark 3.6. $\texttt{QHTME}$ has a provable advantage over optimal classical mean estimators in terms of *query complexity*. By Corollary C.4, optimal classical estimators $\texttt{CHTME}$ require $\tilde{\mathcal{O}}((\sigma/\epsilon)^{\frac{p}{p-1}})$ queries to the classical sampling oracle to solve Problem 3.1. In contrast, $\texttt{QHTME}$ requires only $\tilde{\mathcal{O}}((\sqrt{d}\sigma/\epsilon)^{\frac{p}{2(p-1)}})$ queries to the quantum sampling oracle. As a result, $\texttt{QHTME}$ achieves a quantum speedup over the optimal classical estimators in the low-dimensional regime, specifically when $d = \mathcal{O}((\sigma/\epsilon)^{2-\kappa})$ for any fixed $\kappa \in (0,2)$.

Algorithm 2 Quantum Unbiased Heavy-Tailed Mean Estimator ($\texttt{QUHTME}(\mathbf{X}, \sigma, p, d, \epsilon)$)

Input: random variable $\mathbf{X}$, heavy-tailed parameter σ, tail index p, problem dimension d, error ϵ.
Set $c \leftarrow 24/(2^{\frac{p-1}{2}} - 1)$ and $\tilde{\boldsymbol{\mu}}_0 \leftarrow \texttt{QHTME}^+\left(X, \sigma, p, d, \frac{1}{c^{1/p}}\epsilon\right)$
Randomly sample $j \sim \mathrm{Geom}(1/2) \in \mathbb{N}$
$\tilde{\boldsymbol{\mu}}_j \leftarrow \texttt{QHTME}^+\left(\mathbf{X}, \sigma, p, d, \frac{2^{-3(p-1)j/(2p)}}{c^{1/p}}\epsilon\right)$
$\tilde{\boldsymbol{\mu}}_{j-1} \leftarrow \texttt{QHTME}^+\left(\mathbf{X}, \sigma, p, d, \frac{2^{-3(p-1)(j-1)/(2p)}}{c^{1/p}}\epsilon\right)$
$\tilde{\boldsymbol{\mu}} \leftarrow \tilde{\boldsymbol{\mu}}_0 + 2^j\left(\tilde{\boldsymbol{\mu}}_j - \tilde{\boldsymbol{\mu}}_{j-1}\right)$
Output: $\tilde{\boldsymbol{\mu}}$

3.2 Quantum Unbiased Heavy-Tailed Mean Estimator

The quantum heavy-tailed mean estimator QHTME (Algorithm 1) is, in general, biased. This limitation prevents its direct use in many stochastic optimization frameworks [13, 15, 35, 46], and motivates our study of Problem 3.2. To address this problem, we construct an unbiased quantum estimator by extending the multilevel Monte Carlo (MLMC) variance-reduction technique [20, 46] to the heavy-tailed setting with bounded p-th central moment for $p \in (1, 2]$, and combining this generalized MLMC framework with our estimator QHTME. Our approach proceeds in two steps.

First, we design a modified quantum heavy-tailed mean estimator, denoted by $\texttt{QHTME}^+(\mathbf{X}, \sigma, p, d, \epsilon)$, whose estimation error satisfies the moment bound unconditionally, while retaining the same query complexity as QHTME. The construction combines a single quantum estimate produced by QHTME with a classical estimate, and generates an additional independent classical estimate when the two differ significantly. Details of $\texttt{QHTME}^+$ are given in Appendix C.3.

Second, we embed $\texttt{QHTME}^+$ into a generalized MLMC framework to obtain an unbiased quantum estimator, denoted by $\texttt{QUHTME}(\mathbf{X}, \sigma, p, d, \epsilon)$ and presented in Algorithm 2. The procedure invokes $\texttt{QHTME}^+$ at three randomly selected p-dependent accuracy levels and combines the resulting estimates to produce an unbiased output. Although the accuracy parameters range over an infinite set, we show that the quantum estimator QUHTME is able to solve Problem 3.2 while maintaining the same expected query complexity as $\texttt{QHTME}^+$.

Theorem 3.7. *Algorithm 2 (*QUHTME*) solves Problem 3.2 using* $\tilde{\mathcal{O}}\left((d^{\frac{1}{2}}\sigma/\epsilon)^{\frac{p}{2(p-1)}}\right)$ *queries to the quantum sampling oracle* $O_{\mathbf{X}}$ *in expectation.*

Remark 3.8. $\texttt{QHTME}^+$ and QUHTME generalize the quantum mean estimators of [46, Algorithms 1 and 3] from the bounded-variance regime ($p = 2$) to the heavy-tailed regime ($p \in (1, 2]$).

3.3 Quantum Lower Bounds on Heavy-Tailed Mean Estimation Problems

Here we establish quantum lower bounds for solving Problems 3.1 and 3.2. Proofs are deferred to Appendix C.6.

Theorem 3.9. *There exists a constant* $\alpha > 1$ *such that the following lower bounds hold.*

- *Any quantum algorithm that solves Problem 3.1 with success probability at least* $2/3$ *must make, in the worst case,* $\Omega((\sigma/\epsilon)^{\frac{p}{p-1}})$ *queries to the quantum sampling oracle if* $d \geq (\sigma/\epsilon)^{\frac{2}{p-1}}$ *and* $(\sigma/\epsilon)^{\frac{2-p}{p-1}} \leq \alpha$*, and* $\Omega((\sigma/\epsilon)^{\frac{p}{2(p-1)}})$ *queries otherwise.*
- *Any quantum algorithm that solves Problem 3.2 must make, in worst-case expectation,* $\Omega((\sigma/\epsilon)^{\frac{p}{p-1}})$ *queries to the quantum sampling oracle if* $d \geq (\sigma/\epsilon)^{\frac{2}{p-1}}$ *and* $(\sigma/\epsilon)^{\frac{2-p}{p-1}} \leq \alpha$*, and* $\Omega((\sigma/\epsilon)^{\frac{p}{2(p-1)}})$ *queries otherwise.*

Remark 3.10. For $p \in (1, 2)$, the precision condition required for the high-dimensional lower bound $\Omega((\sigma/\epsilon)^{\frac{p}{p-1}})$ is equivalent to $\epsilon \geq \sigma\alpha^{-\frac{p-1}{2-p}}$, i.e., a low-precision regime. The condition is automatic when $p = 2$. Theorem 3.9 shows that our algorithms QHTME and QUHTME are optimal in their dependence on σ and ϵ. Moreover, when the dimension d is sufficiently large, our quantum lower bounds match the classical lower bounds under the low-precision condition. This implies that no quantum algorithm can outperform classical methods in the high-dimensional and low-precision regime. Consequently, the optimal estimators in this regime are those classical estimators,

e.g., Theorem 2.7 and [54]. For $p \in (1,2)$, whether the same quantum lower bound holds in the high-dimensional, high-precision regime remains open.

4 Quantum Stochastic Nonconvex Optimization with Heavy-Tailed Noise

In this section, we present quantum algorithms for finding stationary points of the objective (1), where $\mathcal{X} = \mathbb{R}^d$ and $f(\cdot)$ is non-convex. We begin by defining the notion of an ϵ-stationary point.

Definition 4.1. A random point $\hat{\mathbf{x}} \in \mathbb{R}^d$ is called an ϵ-stationary point of $f(\cdot)$ if $\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}})\|\right] \leq \epsilon$.

We impose the following standard regularity assumptions on the objective function.

Assumption 4.2. *We assume that the gradient of $f(\cdot)$ is Lipschitz continuous, such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it holds that $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$. We also assume that $f(\cdot)$ is lower bounded such that $f^* = \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$.*

Unlike most existing approaches that use gradient clipping to handle heavy-tailed noise [45, 58], we consider the following normalized stochastic gradient descent (NSGD) update:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|}, \tag{4}$$

where $\mathbf{g}_t$ is an estimator of $\nabla f(\mathbf{x}_t)$. When $\mathbf{g}_t = \mathbf{0}$, we set $\mathbf{x}_{t+1} = \mathbf{x}_t$. The following lemma establishes the descent property of the NSGD update.

Lemma 4.3 (Lemma 2, [11])**.** *Let $\boldsymbol{\epsilon}_t := \mathbf{g}_t - \nabla f(\mathbf{x}_t)$. If $f(\cdot)$ satisfies Assumption 4.2, then the* NSGD *update* (4) *satisfies*

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{\eta}{3}\|\nabla f(\mathbf{x}_t)\| + \frac{8\eta}{3}\|\boldsymbol{\epsilon}_t\| + \frac{L\eta^2}{2}.$$

Lemma 4.3 shows that the descent behavior of NSGD depends explicitly on the magnitude of the gradient estimation error $\|\boldsymbol{\epsilon}_t\|$. By applying QHTME to estimate the stochastic gradient via the quantum stochastic gradient oracle, we can ensure that $\|\boldsymbol{\epsilon}_t\|$ is sufficiently small with high probability. Lastly, with an appropriate choice of the step size η, the objective value $f(\cdot)$ decreases sufficiently along the NSGD update. Motivated by this observation, we propose a quantum normalized stochastic gradient descent method (QNSGD), presented in Algorithm 3, and state its query complexity in the following theorem.

Algorithm 3 Quantum Normalized Stochastic Gradient Descent (QNSGD)

1: **Input:** η, n, δ, and T.
2: **for** $t = 0, \cdots, T-1$ **do**
3: $\quad \mathbf{g}_t = \texttt{QHTME}(\nabla F(\mathbf{x}_t; \xi), n, p, d, \delta/T)$
4: $\quad \mathbf{x}_{t+1} = \mathbf{x}_t - \eta \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|}$
5: **end for**
6: **Output:** choose $\hat{\mathbf{x}}$ uniformly at random from $\{\mathbf{x}_0, \cdots, \mathbf{x}_{T-1}\}$

Theorem 4.4. *Suppose that $f(\cdot)$ satisfies Assumptions 1.1 and 4.2. Run* QNSGD *with the following hyper-parameters:*

$$\eta = L^{-1}\epsilon, \;\; T = \lceil L(f(\mathbf{x}_0) - f^*)\epsilon^{-2}\rceil, \;\; \delta = \frac{1}{2T}, \;\; \text{and} \;\; n = \max\left\{\tilde{\mathcal{O}}\left(d^{\frac{p}{4(p-1)}}\sigma^{\frac{p}{2(p-1)}}\epsilon^{-\frac{p}{2(p-1)}}\right), 1\right\}.$$

Then the output $\hat{\mathbf{x}}$ satisfies

$$\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}})\|\right] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|\right] = \mathcal{O}(\epsilon).$$

Consequently, QNSGD *finds an ϵ-stationary point of $f(\cdot)$ using $\tilde{\mathcal{O}}\left(d^{\frac{p}{4(p-1)}}\sigma^{\frac{p}{2(p-1)}}\epsilon^{-\frac{5p-4}{2p-2}}\right)$ queries to the quantum stochastic gradient oracle.*

Remark 4.5. Theorem 4.4 recovers the $\tilde{\mathcal{O}}\left(d^{1/2}\epsilon^{-3}\right)$ query complexity established in [46] for stochastic non-convex optimization under the bounded variance assumption. Moreover, the theorem indicates that quantum speedups extend to a broader class of stochastic optimization problems with heavy-tailed gradient noise, namely for all $p \in (1,2]$. In particular, for any fixed $\kappa \in (0,2)$, when $d = \mathcal{O}(\epsilon^{-2+\kappa})$, our result improves upon the classical lower bound $\Omega(\epsilon^{-\frac{3p-2}{p-1}})$.

Comparison with classical NSGD**.** [28] showed that NSGD with the gradient estimator $\mathbf{g}_t = \frac{1}{|\mathcal{B}|}\sum_{\xi\in\mathcal{B}}\nabla f(\mathbf{x}_t;\xi)$ in Algorithm 3 can find an ϵ-stationary point using $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$ queries to the classical stochastic gradient oracle, matching the corresponding classical lower bound. To ensure that the gradient estimation error is small enough such that $\|\boldsymbol{\epsilon}_t\| \le \epsilon$, their approach requires a mini-batch size $|\mathcal{B}| = \mathcal{O}(\epsilon^{-\frac{p}{p-1}})$. In contrast, our approach QNSGD applies the quantum heavy-tailed mean estimator QHTME to obtain $\mathbf{g}_t$ and ensure $\|\boldsymbol{\epsilon}_t\| = \mathcal{O}(\epsilon)$ using $\tilde{\mathcal{O}}(d^{\frac{p}{4(p-1)}}\epsilon^{-\frac{p}{2(p-1)}})$ queries to the quantum stochastic gradient oracle at each iteration. As a result, QNSGD achieves a quantum speedup over the classical NSGD in the low-dimensional regime.

5 Quantum Stochastic Convex Optimization with Heavy-Tailed Noise

In this section, we consider the case where the objective $f(\cdot)$ in (1) is convex but possibly non-smooth. We impose the following assumptions on $f(\cdot)$ and the constraint set $\mathcal{X}$.

Assumption 5.1. *The objective function $f(\cdot)$ and the constraint set $\mathcal{X}$ satisfy the following conditions: (i) $\mathcal{X}$ is a nonempty, closed, and convex set with bounded diameter $D > 0$; (ii) The function $f(\cdot)$ is convex on $\mathcal{X}$; (iii) $\|\nabla f(\mathbf{x})\| \le G$.*

In addition, we denote by $\mathbf{x}^*$ an optimal solution of (1), chosen from the optimal solution set $\mathcal{X}^* := \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ so that $\mathbf{x}^* \in \mathcal{X}^*$. Unlike existing approaches based on gradient clipping [58] or the online-to-convex framework [35] for handling heavy-tailed noise, we consider the projected stochastic gradient descent (PSGD):

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}(\mathbf{x}_t - \eta\mathbf{g}_t), \tag{5}$$

where $\mathbf{g}_t$ is a stochastic estimator of $\nabla f(\mathbf{x})$. The following lemma bounds the suboptimality of the averaged iterate produced by PSGD.

Lemma 5.2. *Run* PSGD (5) *from $\mathbf{x}_0 \in \mathcal{X}$ for T iterations, and define the averaged iterate $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{x}_t$ as well as the gradient estimation error $\boldsymbol{\epsilon}_t = \mathbf{g}_t - \nabla f(\mathbf{x}_t)$. If $\mathbb{E}[\mathbf{g}_t \mid \mathbf{x}_t] = \nabla f(\mathbf{x}_t)$ and $f(\cdot)$ satisfies Assumptions 1.1 and 5.1, then for any $p \in (1, 2]$, we have*

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*)\right] \le \frac{D^2}{2\eta T} + \eta G^2 + C_p\eta^{p-1}D^{2-p}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^p\right],$$

where $C_p := (4p-4)^{p-1}/p^p$.

By properly tuning the step size η and controlling the error term $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{\epsilon}_t\|^p]$, one can guarantee the convergence of PSGD. To this end, we apply QUHTME introduced in Section 3.2 to construct an **unbiased estimator** $\mathbf{g}_t$ of $\nabla f(\mathbf{x}_t)$ with controlled heavy-tailed noise. This leads to the quantum projected stochastic gradient descent algorithm (QPSGD), presented in Algorithm 4. The following theorem presents the query complexity of QPSGD for convex stochastic optimization under heavy-tailed noise.

Algorithm 4 Quantum Projected Stochastic Gradient Descent (QPSGD)

1: **Input:** $\mathbf{x}_0 \in \mathcal{X}$, η, T, and $\tilde{\epsilon}$.
2: **for** $t = 0, \cdots, T-1$ **do**
3: $\quad\mathbf{g}_t = \texttt{QUHTME}(\nabla F(\mathbf{x}_t;\xi), \sigma, p, d, \tilde{\epsilon})$
4: $\quad\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}(\mathbf{x}_t - \eta\mathbf{g}_t)$
5: **end for**
6: **Output:** $\hat{\mathbf{x}} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{x}_t$

Theorem 5.3. *Suppose that $f(\cdot)$ satisfies Assumptions 1.1 and 5.1. Run* QPSGD *(Algorithm 4) with the hyper-parameters*

$$\eta = \frac{\epsilon}{3G^2}, T = \left\lceil 3G^2D^2\epsilon^{-2}\right\rceil, \text{ and } \tilde{\epsilon} = \left(\frac{3^{p-1}}{6C_p}\right)^{\frac{1}{p}}\epsilon^{\frac{2-p}{p}}G^{\frac{2(p-1)}{p}}D^{-\frac{2-p}{p}}.$$

Then the output $\hat{\mathbf{x}}$ satisfies $\mathbb{E}[f(\hat{\mathbf{x}}) - f(\mathbf{x}^*)] \le \epsilon$. Consequently, QPSGD *requires at most*

$$\tilde{\mathcal{O}}\left(d^{\frac{p}{4(p-1)}}GD^{\frac{3p-2}{2(p-1)}}\sigma^{\frac{p}{2(p-1)}}\epsilon^{-\frac{3p-2}{2(p-1)}} + G^2D^2\epsilon^{-2}\right)$$

queries to the quantum stochastic gradient oracle in expectation.

Comparison with [16, 35]. Our approach is partially motivated by [35], who established regret bounds for online gradient descent under heavy-tailed noise and then applied the online-to-convex conversion framework to obtain a query complexity of $\mathcal{O}(\epsilon^{-\frac{p}{p-1}})$ for stochastic convex optimization. [16] also applied a similar projected stochastic gradient descent method to handle heavy-tailed noise. Both approaches use a single sample per iteration, so their iteration complexities are also $\mathcal{O}(\epsilon^{-\frac{p}{p-1}})$. In contrast, Theorem 5.3 shows that the iteration complexity of `QPSGD` is $T = \mathcal{O}(\epsilon^{-2})$, which is better than that of the methods in [16, 35] for $p \in (1, 2)$. Moreover, although `QPSGD` uses mini-batch sampling to construct quantum gradient estimators at each iteration, the quantum speedups provided by our newly developed quantum mean estimators `QUHTME` ensure that the total query complexity of `QPSGD` still improves upon the classical methods in the low-dimension regime.

6 Conclusion

In this paper, we studied quantum stochastic optimization under heavy-tailed noise. We developed a novel quantum mean estimator for random vectors with heavy-tailed distributions, called `QHTME`, and demonstrated that it achieves a quantum speedup over optimal classical mean estimators. We further refined `QHTME` into an unbiased quantum mean estimator, namely `QUHTME`. We established quantum lower bounds, showing that both `QHTME` and `QUHTME` are optimal up to logarithmic factors when the dimension d is a constant, and no provable quantum speedup is possible for the high-dimensional and low-precision regime. Building on these quantum tools, we proposed `QNSGD` and `QPSGD` for non-convex and convex stochastic optimization, respectively, and established quantum speedups in both settings. In the low-dimensional regime, we proved that these methods require fewer stochastic gradient oracle queries than the optimal classical algorithms.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their insightful advice. The work of Chengchang Liu was supported by the National Natural Science Foundation of China (624B2125). The work of John C.S. Lui was supported in part by the RGC SRFS2122-4S02 and GRF-14202925.

References

[1] Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth International Conference on Learning Representations*, 2024.

[2] Charles H. Bennett and John Gill. Relative to a random oracle A, $\mathbf{P}^A \neq \mathbf{NP}^A \neq$ co-$\mathbf{NP}^A$ with probability 1. *SIAM Journal on Computing*, 10(1):96–113, 1981. doi: 10.1137/0210008.

[3] Charles H. Bennett, Ethan Bernstein, Gilles Brassard, and Umesh Vazirani. Strengths and weaknesses of quantum computing. *SIAM Journal on Computing*, 26(5):1510–1523, October 1997. ISSN 1095-7111. doi: 10.1137/s0097539796300933.

[4] Jose Blanchet, Mario Szegedy, and Guanyang Wang. Quadratic speed-up in infinite variance quantum Monte Carlo. arXiv preprint arXiv:2401.07497, 2024.

[5] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, August 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00348-9.

[6] Shouvanik Chakrabarti, Andrew M Childs, Tongyang Li, and Xiaodi Wu. Quantum algorithms and lower bounds for convex optimization. *Quantum*, 4:221, 2020.

[7] Iris Cong, Soonwon Choi, and Mikhail D Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.

[8] Arjan Cornelissen and Yassine Hamoudi. A sublinear-time quantum algorithm for approximating partition functions. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete*

Algorithms (SODA), pages 1245–1264. Society for Industrial and Applied Mathematics, 2023. doi: 10.1137/1.9781611977554.ch46.

[9] Arjan Cornelissen and Sofiene Jerbi. Quantum algorithms for multivariate Monte Carlo estimation. arXiv preprint arXiv:2107.03410, 2021.

[10] Arjan Cornelissen, Yassine Hamoudi, and Sofiene Jerbi. Near-optimal quantum algorithms for multivariate mean estimation. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC '22, page 33–43. ACM, June 2022.

[11] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020.

[12] Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021.

[13] Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pages 6643–6670. PMLR, 2023.

[14] Ronald de Wolf. Quantum computing: Lecture notes. arXiv preprint arXiv:1907.09415, 2026.

[15] 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.

[16] Ilyas Fatkhullin, Florian Hübler, and Guanghui Lan. Can SGD handle heavy-tailed noise? *arXiv preprint arXiv:2508.04860*, 2025.

[17] 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.

[18] Saurabh Garg, Joshua Zhanston, Emilio Parisotto, Adarsh Prasad, Zico Kolter, Zachary Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3610–3619. PMLR, 2021.

[19] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

[20] Michael B Giles. Multilevel Monte Carlo methods. *Acta numerica*, 24:259–328, 2015.

[21] Weiyuan Gong, Chenyi Zhang, and Tongyang Li. Robustness of quantum algorithms for non-convex optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

[22] Robert M. Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, January 2015. ISSN 1095-7162. doi: 10.1137/15m1025487.

[23] Yassine Hamoudi. Quantum sub-gaussian mean estimator. In *29th Annual European Symposium on Algorithms (ESA 2021)*, volume 204 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 50:1–50:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. doi: 10.4230/LIPIcs.ESA.2021.50.

[24] Yassine Hamoudi and Frédéric Magniez. Quantum Chebyshev's inequality and applications. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 69:1–69:16. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi: 10.4230/LIPIcs.ICALP.2019.69.

[25] Stefan Heinrich. Quantum summation with an application to integration. *Journal of Complexity*, 18(1):1–50, 2002.

[26] Maxwell Henderson, Samriddhi Shakya, Shashindra Pradhan, and Tristan Cook. Quanvolutional neural networks: powering image recognition with quantum circuits. *Quantum Machine Intelligence*, 2(1):2, 2020.

[27] Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. In *Advances in Neural Information Processing Systems*, volume 36, pages 56576–56588, 2023. doi: 10.52202/075280-2469.

[28] Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2413–2421. PMLR, 2025.

[29] Tak Hur, Leeseok Kim, and Daniel K. Park. Quantum convolutional neural network for classical data classification. *Quantum Machine Intelligence*, 4(1), February 2022. ISSN 2524-4914. doi: 10.1007/s42484-021-00061-x.

[30] Junhyung Lyle Kim, Brandon Augustino, Dylan Herman, Enrico Fontana, Jacob Watkins, Marco Pistoia, and Shouvanik Chakrabarti. Fast zeroth-order convex optimization with quantum gradient methods. In *Advances in Neural Information Processing Systems*, volume 38, pages 94243–94291, 2025. doi: 10.52202/085713-2834.

[31] Jiaqi Leng, Yufan Zheng, Zhiyuan Jia, Lei Fan, Chaoyue Zhao, Yuxiang Peng, and Xiaodi Wu. Quantum Hamiltonian descent for non-smooth optimization. *arXiv preprint arXiv:2503.15878*, 2025.

[32] Tongyang Li, Chunhao Wang, Shouvanik Chakrabarti, and Xiaodi Wu. Sublinear classical and quantum algorithms for general matrix games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8465–8473, 2021.

[33] Chengchang Liu, Chaowen Guan, Jianhao He, and John C.S. Lui. Quantum algorithms for non-smooth non-convex optimization. *Advances in Neural Information Processing Systems*, 37: 35288–35312, 2024.

[34] Chengchang Liu, Zongqi Wan, Jialin Zhang, Xiaoming Sun, and John C.S. Lui. Quantum speedups for minimax optimization and beyond. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[35] Zijian Liu. Online convex optimization with heavy tails: Old algorithms, new regrets, and applications. In *Proceedings of The 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pages 1–47. PMLR, 2026.

[36] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019. doi: 10.1007/s10208-019-09427-x.

[37] George Marsaglia and Wai Wan Tsang. The ziggurat method for generating random variables. *Journal of statistical software*, 5:1–7, 2000.

[38] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1), November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07090-4.

[39] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4), November 2015. ISSN 1350-7265. doi: 10.3150/14-bej645.

[40] Ashley Montanaro. Quantum speedup of Monte Carlo methods. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2181):20150301, September 2015.

[41] Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. *Advances in Neural Information Processing Systems*, 36:24191–24222, 2023.

[42] Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2010.

[43] Guneykan Ozgul, Xiantao Li, Mehrdad Mahdavi, and Chunhao Wang. Quantum speedups for sampling and non-convex optimization with stochastic oracles. In *21st Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2026)*, volume 389, 2026.

[44] Maxim Raginsky and Alexander Rakhlin. Information complexity of black-box convex optimization: A new look via feedback information theory. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 803–810. IEEE, 2009. doi: 10.1109/ALLERTON.2009.5394945.

[45] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pages 29563–29648. PMLR, 2023.

[46] Aaron Sidford and Chenyi Zhang. Quantum speedups for stochastic optimization. *Advances in Neural Information Processing Systems*, 36:35300–35330, 2023.

[47] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[48] James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992. doi: 10.1109/9.119632.

[49] Letian Tang. More-efficient quantum multivariate mean value estimator from generalized Grover operator. arXiv preprint arXiv:2504.06940, 2025.

[50] Joran van Apeldoorn. Quantum probability oracles and multidimensional amplitude estimation. In *Proceedings of the 16th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2021)*, volume 197 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 9:1–9:11, Dagstuhl, Germany, 2021. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[51] Joran van Apeldoorn, András Gilyén, Sander Gribling, and Ronald de Wolf. Convex optimization using quantum oracles. *Quantum*, 4:220, 2020.

[52] Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303, 1965. doi: 10.1214/aoms/1177700291.

[53] Hao Wang, Chenyi Zhang, and Tongyang Li. Near-optimal quantum algorithm for minimizing the maximal loss. In *The Twelfth International Conference on Learning Representations*, 2024.

[54] Justin Whitehouse, Ben Chugg, Diego Martinez-Taboada, and Aaditya Ramdas. Mean estimation in Banach spaces under infinite variance and martingale dependence. arXiv preprint arXiv:2411.11271v2, 2025.

[55] Mark M Wilde. *Quantum information theory*. Cambridge University Press, 2013.

[56] Yulian Wu, Chaowen Guan, Vaneet Aggarwal, and Di Wang. Quantum heavy-tailed bandits. *arXiv preprint arXiv:2301.09680*, 2023.

[57] Chenyi Zhang and Tongyang Li. Quantum lower bounds for finding stationary points of nonconvex functions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41268–41299. PMLR, 2023.

[58] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

[59] Yexin Zhang, Chenyi Zhang, Cong Fang, Liwei Wang, and Tongyang Li. Quantum algorithms and lower bounds for finite-sum optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60244–60270. PMLR, 2024.

[60] Vincent Zhuang and Yanan Sui. No-regret reinforcement learning with heavy-tailed rewards. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3385–3393. PMLR, 2021.

A Discussion on the Quantum Oracles

A.1 Discussion on the Quantum Access Model for Random Variables

Comparisons with existing quantum sampling oracles. Unlike in the classical setting, existing work [10] assumes that quantum access to the distribution $\mathbb{P}$ and to the random variable $\mathbf{X}$ is typically provided separately, via a probability oracle and a binary oracle, respectively.

Definition A.1 (Probability oracle)**.** Let $\mathbf{X} : \Omega \to E$ be a random variable on a probability space $(\Omega, 2^{\Omega}, \mathbb{P})$, and let $\mathcal{H}_{\Omega}$ be a Hilbert space with basis states $\{|\omega\rangle\}_{\omega\in\Omega}$. A probability oracle for the probability distribution $\mathbb{P}$ is a unitary operator acting on $\mathcal{H}_{\Omega}$ such that

$$U_{\mathbb{P}} : |0\rangle \mapsto \sum_{\omega\in\Omega} \sqrt{\mathbb{P}(\omega)}\,|\omega\rangle,$$

assuming $0 \in \Omega$.

Definition A.2 (Binary oracle)**.** Let $\mathbf{X} : \Omega \to E$ be a finite-valued random variable on a probability space $(\Omega, 2^{\Omega}, \mathbb{P})$. Let $\mathcal{H}_{\Omega}$ and $\mathcal{H}_E$ be two Hilbert spaces with basis states $\{|\omega\rangle\}_{\omega\in\Omega}$ and $\{|x\rangle\}_{x\in E}$, respectively. A binary oracle for $\mathbf{X}$ is a unitary $\mathcal{B}_{\mathbf{X}}$ acting on $\mathcal{H}_{\Omega} \otimes \mathcal{H}_E$ such that

$$\mathcal{B}_{\mathbf{X}} : |\omega\rangle|\vec{0}\rangle \mapsto |\omega\rangle|\mathbf{X}(\omega)\rangle$$

for all $\omega \in \Omega$, assuming $\vec{0} \in E$.

However, this quantum access model is stronger than the classical sampling oracle defined in Definition 2.1. In particular, it provides coherent access to the underlying random seed ω, as the probability oracle $U_{\mathbb{P}}$ prepares a superposition over all $\omega \in \Omega$, which can then be queried via the binary oracle B_X. This allows quantum algorithms to process all samples in superposition.

In contrast, the classical sampling procedure only returns independent samples of the form $\mathbf{X}(\omega)$, and does not provide access to the underlying seed ω. We therefore choose to base our results on the quantum sampling oracle in Definition 2.2, which can be viewed as a coherent analogue of the classical sampling oracle. Although our quantum access model is weaker than the one considered in [10], this difference does not affect the upper bound in Lemma 3.3. In particular, the algorithm `QBounded` in [10] only relies on coherent access to the values of the random variable, rather than the underlying random seed.

We also note that [46, Definition 1] adopts a closely related definition of a quantum sampling oracle, where the quantum state is expressed as a superposition over a continuous probability distribution using an integral representation.

Definition A.3 (Quantum sampling oracle in [46])**.** For a d-dimensional random variable $\mathbf{X}$, its quantum sampling oracle $O_{\mathbf{X}}$ is defined as

$$O_{\mathbf{X}}\,|0\rangle \;\mapsto\; \int_{\mathbf{x}\in\mathbb{R}^d} \sqrt{p_{\mathbf{X}}(\mathbf{x})}\,dx\;\,|\mathbf{x}\rangle \otimes |\mathrm{garbage}(\mathbf{x})\rangle\,, \tag{6}$$

where $p_{\mathbf{X}}(\cdot)$ represents the probability density function of $\mathbf{X}$.

This formulation should be interpreted as an "*idealized model*", since physical implementations necessarily operate on finite-precision representations. In practice, continuous random variables must be discretized to be encoded into quantum states. Under a suitable discretization scheme, the integral representation in [46, Definition 1] can be approximated by a finite superposition over a discrete support. In this sense, their definition is consistent with ours up to discretization, and both models capture the same underlying notion of coherent sampling from a distribution.

Compatibility with heavy-tailed assumptions. Our oracle model assumes that the random variable $\mathbf{X}$ takes values in a finite set $E \subset \mathbb{R}^d$, which is necessary for representing quantum states using a finite number of qubits. At first glance, this may appear at odds with the heavy-tailed setting considered in this paper, where the underlying distribution may have unbounded support and possibly infinite variance.

This discrepancy can be resolved by interpreting our oracle as operating on a truncated and discretized approximation of an underlying heavy-tailed distribution. Concretely, let $\mathbf{X}$ be a random variable

satisfying a finite p-th moment condition for some $p \in (1, 2]$, while its variance may be infinite when $p < 2$. We consider a truncated version

$$\mathbf{X}^{(R)} := \mathbf{X} \cdot \mathbf{1}\{\|\mathbf{X}\| \le R\}$$

for a sufficiently large truncation radius R, and discretize its support using a sufficiently fine grid. The resulting finite-valued random variable can then be encoded into a quantum state as in Definition 2.2.

Under this interpretation, the finite support E should be viewed as a finite-precision approximation of a continuous distribution, where both the truncation radius and discretization granularity are chosen so that the induced bias is negligible at the target accuracy. In particular, for sufficiently large R, the bias introduced by truncation can be made arbitrarily small, while the p-th central moment of $\mathbf{X}^{(R)}$ remains finite and comparable to that of $\mathbf{X}$.

Although $\mathbf{X}^{(R)}$ has finite support and hence finite variance, this variance may grow with R and can be arbitrarily large as the approximation becomes more faithful to the original heavy-tailed distribution. As a result, the finite-support implementation does not reduce the problem to a bounded-variance regime in any meaningful uniform sense. Our analysis instead relies on p-th moment bounds, which are stable under truncation and discretization up to controlled constants.

Finally, we note that such truncation and discretization are not specific to the quantum setting. In classical computing, sampling from continuous or unbounded distributions also necessarily involves finite-precision representations and approximation schemes. For example, the Ziggurat method [37] can be viewed as a structured instance of a truncate-and-discretize approach, combined with efficient rejection sampling for generating random variables from unimodal distributions such as the Gaussian or exponential.

A.2 Discussion on the Quantum Stochastic Gradient Oracle

We note that the oracle model in [46, Definitions 3 and 4] is based on a quantum stochastic gradient oracle that requires coherent access to the query point $\mathbf{x}$. By contrast, our model (Definition 2.4) only assumes quantum access to the stochastic gradient distribution at each fixed query point. In optimization problems, the query point $\mathbf{x}$ represents the current parameter vector (e.g., the weights of a classical or quantum neural network), which is updated adaptively by the algorithm and typically ranges over a continuous or very high-precision domain. Requiring coherent access to $|\mathbf{x}\rangle$ therefore amounts to the (strong) assumption of a quantum interface that can reversibly encode and query the stochastic gradient distribution for all possible parameter values in superposition.

B Discussion on Practicality and Potential Applications

We briefly discuss the practical scope of our algorithms, focusing on the quantum access model, the role of the dimension dependence, and potential application domains.

Quantum oracle implementation. Our algorithms are formulated in the quantum oracle models for stochastic optimization and mean estimation: a quantum stochastic gradient oracle (Definition 2.4) coherently prepares a stochastic gradient sample, while a quantum sampling oracle (Definition 2.2) coherently prepares a random vector whose mean is to be estimated. These models should be interpreted as the quantum analogue of the classical stochastic first-order oracle and sampling oracle, respectively. In particular, if the corresponding classical sampling or stochastic-gradient routine is implemented by an explicit reversible circuit, then the standard reversible-computation construction yields a quantum circuit implementing the associated unitary oracle and its inverse, up to the constant overhead for reversible simulation [42, Section 1.4.1]. Thus, the relevant comparison in this paper is a black-box query-complexity comparison between classical and quantum algorithms under analogous oracle access assumptions. Within this framework, our results demonstrate that quantum algorithms can achieve a provable advantage over their classical counterparts in the low-dimensional regime.

Hardware implementation. The quantum (unbiased) heavy-tailed mean estimators used in our quantum stochastic optimization algorithms require coherent oracle access, inverse oracle invocation, and sufficiently deep controlled quantum operations. The quantum circuits required to implement our algorithms are too deep to run on existing devices (which have limited qubit counts and high error

rates), and require large-scale fault-tolerant quantum computers to implement. This is also the case for the related literature on quantum univariate and multivariate mean estimation [4, 10, 40, 49, 56]. Until the technology becomes available to implement these algorithms on real devices, demonstrating theoretical, provable quantum speedups for different optimization problems remains an important task in the machine learning and optimization communities. In this sense, our paper is aligned with a broader body of work on quantum algorithms for learning and optimization [33, 34, 46, 53, 57, 59].

Dimension dependence and the low-dimensional regime. The quantum speedups obtained by `QNSGD` and `QPSGD` are in the low-dimensional regime. For nonconvex optimization, `QNSGD` uses

$$\widetilde{O}\Big(d^{\frac{p}{4(p-1)}}\epsilon^{-\frac{5p-4}{2p-2}}\Big)$$

queries to the quantum stochastic gradient oracle, compared with the classical lower bound $\Omega\Big(\epsilon^{-\frac{3p-2}{p-1}}\Big)$. Hence `QNSGD` improves the dependence on ϵ when $d \lesssim \epsilon^{-2}$, up to logarithmic and problem-dependent factors. For convex optimization, `QPSGD` uses

$$\widetilde{O}\Big(d^{\frac{p}{4(p-1)}}\epsilon^{-\frac{3p-2}{2p-2}}\Big)$$

queries, compared with the classical lower bound $\Omega\Big(\epsilon^{-\frac{p}{p-1}}\Big)$, and gives a speedup when $d \lesssim \epsilon^{-\frac{2(2-p)}{p}}$, up to logarithmic factors.

Although we demonstrate quantum speedup only in the low-dimensional regime, it remains meaningful in some settings.

- First, while in classical machine learning (ML), the large d setting is typically of interest, in quantum ML, it is desirable to have much parameter-sparse neural networks. This stems from the fact that (i) backpropagation cannot be performed in quantum neural networks, and thus gradients are typically estimated via either the parameter-shift [5, Section C.1] or finite difference methods [48], both of which require $\mathcal{O}(d)$ network evaluations to compute a single gradient; and (ii) increasing the number of parameters can lead to "barren plateaus" [38], the quantum equivalent of vanishing gradients. To keep the quantum neural networks efficiently trainable, highly structured networks with a sparse number of parameters are desirable. Our algorithms require access to mini-batches of data in quantum superposition (which is implicit in Definition 2.2) and therefore can be used for estimating the gradients and training the parameters in quantum neural networks. In such low-dimensional or parameter-sparse settings, our algorithms for finding (locally) optimal parameters are potentially more efficient than their classical counterparts.
- Second, our quantum algorithms may also be useful for accelerating optimization problems encountered in the classical domain, as many high-dimensional optimization methods reduce each update to a low-dimensional randomized subproblem. Examples include sketch-and-project methods for linear systems [22] and an algorithm for the approximately best point problem [46, Problem 5]. These methods suggest that the effective dimension of an update can be much smaller than the problem dimension, even when the original problem is high-dimensional. In a stochastic setting with heavy-tailed gradient noise, if the low-dimensional projected gradient or sketched stochastic quantity satisfies a bounded p-th moment condition, then our quantum heavy-tailed mean estimators can be potentially used as subroutines for estimating these low-dimensional quantities and provide a quantum speedup.

Natural application scenarios. A natural application domain of our quantum stochastic optimization algorithms is training parameterized quantum models under heavy-tailed stochastic gradients. The stochasticity may arise from mini-batching, randomized measurements, coordinate sampling, or stochastic circuit components. The data may be classical but available through coherent access mechanisms such as QRAM, or it may be intrinsically quantum, for example, when the training sample is produced by another quantum circuit. In these settings, the goal is to optimize quantum [7, 29] or hybrid quantum-classical [26] models whose parameter dimension is moderate and whose stochastic gradients may fail to have bounded variance.

More broadly, the quantum heavy-tailed mean estimators developed in this paper are independent algorithmic primitives. Heavy-tailed expectation estimation appears in learning and decision-making

problems beyond stochastic optimization, including reinforcement learning with heavy-tailed rewards [27, 60] and policy-gradient methods with heavy-tailed gradient noise [18]. If the relevant environment, transition, reward, or gradient oracle can be accessed coherently, then QHTME or QUHTME may serve as a quantum subroutine for reducing the query complexity of these problems.

Scope and limitations. Our results are theoretical query-complexity results, and we do not make claims about quantum advantage on near-term quantum hardware devices. The main contribution is to show that, under the standard coherent oracle model, stochastic optimization with heavy-tailed noise admits provable quantum query complexity speedups in regimes where the effective dimension is sufficiently small. This positions the algorithms as fault-tolerant quantum subroutines for robust stochastic optimization, rather than immediate replacements for classical optimizers on current hardware.

C Proof in Section 3

C.1 Classical Heavy-Tailed Mean Estimator

Whitehouse et al. [54] was the first work to explicitly formalize a classical heavy-tailed mean estimator. Prior to this, the geometric median-of-means approach of Minsker [39] already implied such an estimator, although they did not state this result explicitly. In Appendix B, Whitehouse et al. [54] further provide a formal treatment of the geometric median-of-means construction. However, both their original estimator and this reformulation require the heavy-tailed parameter σ as an input. In contrast, we present below an alternative formalization of the geometric median-of-means approach that eliminates the need for prior knowledge of σ, but achieves the same error bound as the previous two classical estimators in [54].

Algorithm 5 Classical Heavy-tailed Mean Estimator ($\texttt{CHTME}(\mathbf{X}, n, \delta)$)

Require: Samples $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^d$, confidence level $\delta \in (0, 1)$

Set $k \leftarrow \lceil 24 \ln(1/\delta) \rceil$

Set $m \leftarrow \lfloor n/k \rfloor$

Partition $\{\mathbf{X}_i\}_{i=1}^{km}$ into k disjoint blocks $\{B_j\}_{j=1}^{k}$ of size m

for $j = 1$ to k **do**

  Compute block mean

$$\mathbf{Z}_j \leftarrow \frac{1}{m} \sum_{i \in B_j} \mathbf{X}_i$$

end for

Compute the geometric median of $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_k\}$:

$$\tilde{\boldsymbol{\mu}} \leftarrow \arg\min_{x \in \mathbb{R}^d} \sum_{j=1}^{k} \|\mathbf{Z}_j - x\|$$

return $\tilde{\boldsymbol{\mu}}$

To provide the performance guarantee for the classical mean estimator CHTME, we will use the following lemmas.

Lemma C.1 (Proximity lemma for the geometric median)**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_k$ be independent samples of a d-dimensional variable with mean $\boldsymbol{\mu}$, and let $\tilde{\boldsymbol{\mu}}$ be its geometric median. Fix $\alpha \in (0, 1/2]$ and $r > 0$. If at least $(1/2 + \alpha)k$ samples lie within Euclidean distance r of $\boldsymbol{\mu}$, then*

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \left(1 + \frac{1}{2\alpha}\right) r.$$

Proof. Define function $f : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$f(y) = \sum_{i=1}^{k} \|y - \mathbf{X}_i\|.$$

We define $\tilde{\boldsymbol{\mu}} := \arg\min_{y\in\mathbb{R}^d} f(y)$. Then it holds that $f(\tilde{\boldsymbol{\mu}}) \le f(\boldsymbol{\mu})$ for the mean $\boldsymbol{\mu} \in \mathbb{R}^d$.

For arbitrary $y \in \mathbb{R}^d$, we denote $\rho := \|y - \boldsymbol{\mu}\|$. We consider the following two sets:

$$G := \{\mathbf{X}_i \in \{\mathbf{X}_i\}_{i=1}^k : \|\mathbf{X}_i - \boldsymbol{\mu}\| \le r\}, \quad B := \{\mathbf{X}_i \in \{\mathbf{X}_i\}_{i=1}^k : \|\mathbf{X}_i - \boldsymbol{\mu}\| > r\}.$$

Specifically, the set G contains the "good" points among the samples $\{\mathbf{X}_i\}_{i=1}^k$ satisfying $\|\mathbf{X}_i - \boldsymbol{\mu}\| \le r$. We assume the cardinality of G is $|G| = m$. The set B contains the "bad" samples, which are far from the mean $\boldsymbol{\mu}$. There are $k - m$ such bad samples.

For any good point $\mathbf{X}_i$ in G, by triangle inequality, we have

$$\|y - \mathbf{X}_i\| \ge \|y - \boldsymbol{\mu}\| - \|\mathbf{X}_i - \boldsymbol{\mu}\| \ge \rho - r.$$

For any bad point $\mathbf{X}_i$ in B, by triangle inequality,

$$\|y - \mathbf{X}_i\| \ge \|\mathbf{X}_i - \boldsymbol{\mu}\| - \|y - \boldsymbol{\mu}\| = \|\mathbf{X}_i - \boldsymbol{\mu}\| - \rho.$$

Therefore, we have

$$f(y) = \sum_{\mathbf{X}_i\in G} \|y - \mathbf{X}_i\| + \sum_{\mathbf{X}_i\in B} \|y - \mathbf{X}_i\| \ge m(\rho - r) + \sum_{\mathbf{X}_i\in B} \|\mathbf{X}_i - \boldsymbol{\mu}\| - (k - m)\rho.$$

Besides, we know that

$$f(\boldsymbol{\mu}) = \sum_{\mathbf{X}_i\in G} \|\boldsymbol{\mu} - \mathbf{X}_i\| + \sum_{\mathbf{X}_i\in B} \|\boldsymbol{\mu} - \mathbf{X}_i\| \le mr + \sum_{\mathbf{X}_i\in B} \|\boldsymbol{\mu} - \mathbf{X}_i\|.$$

By computing the subtraction, we have

$$f(y) - f(\boldsymbol{\mu}) \ge (m - (k - m))\rho - 2mr = (2m - k)\rho - 2mr.$$

By assumption, $m \ge (1/2 + \alpha)k$, so $2m - k \ge 2\alpha k > 0$. Since $\tilde{\boldsymbol{\mu}}$ is a geometric median, we have $f(\tilde{\boldsymbol{\mu}}) \le f(\boldsymbol{\mu})$. Therefore,

$$0 \ge f(\tilde{\boldsymbol{\mu}}) - f(\boldsymbol{\mu}) \ge (2m - k)\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - 2mr,$$

which implies

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \le \frac{2m}{2m - k} r = \left(1 + \frac{k}{2m - k}\right) r \le \left(1 + \frac{1}{2\alpha}\right) r.$$

□

Lemma C.2 (Vector-valued von Bahr–Esseen inequality [52, 54])**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent d-dimensional real random variables, with $\mathbb{E}[\mathbf{X}_i] = 0$ and $\mathbb{E}[\|\mathbf{X}_i\|^p] < \infty$ for some $p \in [1, 2]$. Then,*

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \mathbf{X}_i\right\|^p\right] \le C_p \sum_{j=1}^n \mathbb{E}[\|\mathbf{X}_j\|^p]$$

where C_p is a (p-dependent) constant that does not depend on d.

Theorem C.3. *Let $\mathbf{X}$ be a d-dimensional random variable with mean $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}]$ and finite p-th central moment $\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]$ for some $p \in (1, 2]$. Let $\delta \in (0, 1)$ and $k = \lceil 24 \ln(1/\delta) \rceil$. For integer n such that $n \ge 2k$, the classical heavy-tailed mean estimator $\texttt{CHTME}(\mathbf{X}, n, \delta)$ outputs a mean estimate $\tilde{\boldsymbol{\mu}}$ such that*

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \le 3(8C_p)^{1/p} \left(\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]\right)^{1/p} \left(\frac{2\lceil 24 \ln(1/\delta)\rceil}{n}\right)^{\frac{p-1}{p}},$$

by using n queries to the classical sampling oracle in Definition 2.1 with probability at least $1 - \delta$.

Proof. Let $\mathbf{Y}_i = \mathbf{X}_i - \boldsymbol{\mu}$. Then,

$$\begin{aligned}
\mathbb{E}[\|\mathbf{Z}_j - \boldsymbol{\mu}\|^p] &= \mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m}(\mathbf{X}_i - \boldsymbol{\mu})\right\|^p = \mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{Y}_i\right\|^p \\
&= \frac{1}{m^p}\mathbb{E}\left\|\sum_{i=1}^{m}\mathbf{Y}_i\right\|^p \\
&\le \frac{C_p}{m^p}\sum_{i=1}^{m}\mathbb{E}[\|\mathbf{Y}_i\|^p] && \text{Lemma } C.2 \\
&\le \frac{C_p\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]}{m^{p-1}} && \mathbf{X}_i \text{ are independent.} \\
&\le C_p\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]\left(\frac{k}{n-k}\right)^{p-1} && \left(m = \left\lfloor\frac{n}{k}\right\rfloor \ge \frac{n}{k} - 1 = \frac{n-k}{k}\right) \\
&\le C_p\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]\left(\frac{2k}{n}\right)^{p-1}. && \left(n \ge 2k \Rightarrow n - k \ge \frac{n}{2}\right)
\end{aligned}$$

By Markov's inequality, we have

$$P(\|\mathbf{Z}_j - \boldsymbol{\mu}\| > r) = P(\|\mathbf{Z}_j - \boldsymbol{\mu}\|^p > r^p) \le \frac{\mathbb{E}[\|\mathbf{Z}_j - \boldsymbol{\mu}\|^p]}{r^p} \le C_p\frac{\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]}{r^p}\left(\frac{2k}{n}\right)^{p-1} := p_{\text{fail}}.$$

By setting

$$r = (8C_p)^{1/p}\left(\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]\right)^{1/p}\left(\frac{2k}{n}\right)^{\frac{p-1}{p}},$$

we have $p_{\text{fail}} = \frac{1}{8}$.

Define the random variables $B_j = \mathbb{I}[\|\mathbf{Z}_j - \boldsymbol{\mu}\| > r]$, i.e., with probability $p < p_{\text{fail}}$ we have $B_j = 1$. We also define the random variables $G_j = \mathbb{I}[\|\mathbf{Z}_j - \boldsymbol{\mu}\| \le r]$. Note that $\sum_{j=1}^{k}(G_j + B_j) = k$. Let Q_j be a Bernoulli random variable with $p(Q_j = 1) = p_{\text{fail}}$. Then, since $p < p_{\text{fail}}$ we have

$$P(\sum_{j=1}^{k}B_j \ge M) \le P(\sum_{j=1}^{k}Q_j \ge M)$$

Take $M = (1+\Delta)k \cdot p_{\text{fail}}$. By the multiplicative Chernoff bound for $\Delta \in [0, 1]$,

$$P\left(\sum_{j=1}^{k}B_j \ge (1+\Delta)k \cdot p_{\text{fail}}\right) \le P\left(\sum_{j=1}^{k}Q_j \ge (1+\Delta)k \cdot p_{\text{fail}}\right) \le e^{-k\cdot p_{\text{fail}}\Delta^2/3}.$$

In order to have $(1+\Delta)p_{\text{fail}} = 1 - (1/2 + \alpha)$, we choose $\Delta = 8(1/2 - \alpha) - 1 = 3 - 8\alpha$. Therefore, we know that

$$P\left(\sum_{j=1}^{k}G_j \ge (1/2+\alpha)k\right) \ge 1 - e^{-k\cdot(3-8\alpha)^2/24}.$$

It indicates that, with probability at least $1 - e^{-k\cdot(3-8\alpha)^2/24}$, there are $(1/2+\alpha)k$ good samples among $\{\mathbf{Z}_j\}_{j=1}^{k}$ satisfying $\|\mathbf{Z}_j - \boldsymbol{\mu}\| \le r$. By Lemma C.1, we know that the geometric median $\tilde{\boldsymbol{\mu}}$ satisfies

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \le \left(1 + \frac{1}{2\alpha}\right)r,$$

with probability at least $1 - e^{-k\cdot(3-8\alpha)^2/24}$.

By choosing $\alpha = \frac{1}{4}$ so that $\Delta = 1$, and $k = \lceil 24 \ln(\frac{1}{\delta}) \rceil$, we have

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \left(1 + \frac{1}{2\alpha}\right) r = 3(8C_p)^{1/p} \left(\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]\right)^{1/p} \left(\frac{2\lceil 24 \ln(1/\delta)\rceil}{n}\right)^{\frac{p-1}{p}},$$

with probability at least $1 - e^{-k/24} \geq 1 - \delta$, by using $n \geq 2k \geq 2\lceil 24 \ln(1/\delta)\rceil$ classical samples obtained by the classical sampling oracle. □

Corollary C.4. `CHTME` *solves Problem 3.1 using* $n(\epsilon) := \tilde{\mathcal{O}}\left(\sigma^{\frac{p}{p-1}} \epsilon^{-\frac{p}{p-1}}\right)$ *queries to the classical sampling oracle with probability at least* $1 - \delta$.

C.2 Proof of Theorem 3.4

The proof of Theorem 3.4 is a direct result of the following theorem.

Theorem C.5. *Let* $\mathbf{X}$ *be a* d*-dimensional random variable with mean* $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}]$ *and finite* p*-th central moment* $\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]$ *for some* $p \in (1, 2]$*. Then, for any* $\delta \in (0, 1)$ *and* $n \geq 2\lceil 24 \log(2/\delta)\rceil$*, Algorithm 1, denoted by* $\texttt{QHTME}(\mathbf{X}, n, p, d, \delta)$*, outputs an estimate* $\tilde{\boldsymbol{\mu}}$ *such that*

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \frac{4d^{\frac{1}{2}} \left(\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p]\right)^{\frac{1}{p}} \log(2d/\delta)}{n^{\frac{2(p-1)}{p}}},$$

with probability at least $1 - \delta$*, by using* $\tilde{\mathcal{O}}(n)$ *queries to the quantum sampling oracle* $O_{\mathbf{X}}$.

Proof. Our proof idea mainly follows the idea in the proof of Theorem 3.4 in Cornelissen et al. [10]. We first assume that all the quantum subroutines, e.g., `Quantile` and $\texttt{QBounded}_d$, are successful. There are k calls to `Quantile` and at most k calls to $\texttt{QBounded}_d$. Applying the subroutine guarantees conditionally on the preceding history, all the required quantum-subroutine guarantees hold with probability at least $(1 - \delta/(10k))^{(2k)} \geq 1 - (2k)\delta/(10k) \geq 1 - \delta/2$ by the union bound and Bernoulli inequality. Then the whole algorithm succeeds with probability at least $1 - \delta$ by including the failure probability $\delta/2$ of `CHTME`. Our subsequent analysis assumes all the classical and quantum subroutines are successfully implemented. In the following analysis, expectations involving $\mathbf{Y}$ and its truncations are taken over a fresh draw of $\mathbf{X}$ while keeping the rough center $\boldsymbol{\zeta}$ and the computed thresholds fixed.

Note that $a_0 = 0$, while the corrected thresholds $\{a_\ell\}_{\ell=1}^k$ satisfy $Q(2^{-\ell}) \leq a_\ell \leq Q(c2^{-\ell})$ for all $\ell \in \{1, \ldots, k\}$, where c is the constant mentioned in Lemma 2.6. Specifically, let b_ℓ denote the raw output of the quantile estimator at level ℓ, and set $a_\ell = \max\{a_{\ell-1}, b_\ell\}$. On the event that all quantile estimates are successful, we have $Q(2^{-\ell}) \leq b_\ell \leq Q(c2^{-\ell})$. Since Q is nonincreasing in its argument, for every $1 \leq j \leq \ell$, we have $b_j \leq Q(c2^{-j}) \leq Q(c2^{-\ell})$. Consequently, for $\ell = 1, \ldots, k$,

$$Q(2^{-\ell}) \leq b_\ell \leq a_\ell \leq Q(c2^{-\ell}), \qquad a_{\ell-1} \leq a_\ell.$$

For $\ell \in \{1, \ldots, k\}$ with $a_\ell > 0$, on one hand, by Markov's inequality for the p-th moment,

$$\Pr(\|\mathbf{Y}\| \geq a_\ell) \leq \frac{\mathbb{E}[\|\mathbf{Y}\|^p]}{a_\ell^p}.$$

On the other hand, by definition of the quantile function in Definition 2.5, it holds that

$$\Pr(\|\mathbf{Y}\| \geq a_\ell) \geq \Pr\left(\|\mathbf{Y}\| \geq Q(c2^{-\ell})\right) \geq c2^{-\ell}.$$

Therefore, since $p \in (1, 2]$, we have

$$a_\ell \leq c^{-1/p} 2^{\ell/p} \left(\mathbb{E}[\|\mathbf{Y}\|^p]\right)^{1/p}.$$

This bound also holds trivially when $a_\ell = 0$.

For $\ell = 1, \ldots, k$, define the normalized truncated random variable

$$\mathbf{Y}_\ell := \begin{cases} \dfrac{\mathbf{Y} \cdot \mathbf{1}\{a_{\ell-1} < \|\mathbf{Y}\| \leq a_\ell\}}{a_\ell}, & a_\ell > 0, \\ 0, & a_\ell = 0. \end{cases}$$

Then $\|\mathbf{Y}_\ell\| \le 1$. For $\ell = 1$, we directly have $\mathbb{E}[\|\mathbf{Y}_1\|] \le 1$. For $\ell = 2, \dots, k$, we have

$$\Pr(\|\mathbf{Y}\| > a_{\ell-1}) \le \Pr\big(\|\mathbf{Y}\| > Q(2^{-(\ell-1)})\big) < 2^{-(\ell-1)}.$$

Therefore, for every $\ell = 1, \dots, k$,

$$\mathbb{E}[\|\mathbf{Y}_\ell\|] \le 2^{-(\ell-1)}.$$

Hence, $2^{-(\ell-1)} \in (0, 1]$ is a valid upper bound on $\mathbb{E}[\|\mathbf{Y}_\ell\|]$.

We denote $\tilde{\boldsymbol{\mu}}_{\mathbf{Y}} = \sum_{\ell=1}^{k} a_\ell \tilde{\boldsymbol{\mu}}_\ell$ as the mean estimate of $\boldsymbol{\mu}_{\mathbf{Y}} := \mathbb{E}[\mathbf{Y}] = \mathbb{E}[\mathbf{X} - \boldsymbol{\zeta}] = \mathbb{E}[\mathbf{X}] - \boldsymbol{\zeta}$. Then, we know that

$$\|\tilde{\boldsymbol{\mu}}_{\mathbf{Y}} - \boldsymbol{\mu}_{\mathbf{Y}}\|_\infty = \left\|\boldsymbol{\zeta} + \sum_{\ell=1}^{k} a_\ell \tilde{\boldsymbol{\mu}}_\ell - \boldsymbol{\mu}\right\|_\infty = \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty,$$

which indicates that we only need to focus on bounding the error of the random variable $\mathbf{Y}$.

For $\ell = 1, \dots, k$, Lemma 3.3 gives

$$\|\tilde{\boldsymbol{\mu}}_\ell - \mathbb{E}[\mathbf{Y}_\ell]\|_\infty \le \frac{2^{-(\ell-1)/2}\log(10kd/\delta)}{n'}.$$

When $a_{\ell-1} = a_\ell$, this inequality holds trivially, since both $\mathbf{Y}_\ell$ and $\tilde{\boldsymbol{\mu}}_\ell$ are zero. Moreover, for the random variable $\mathbf{Z} := \mathbf{Y} \cdot \mathbf{1}\{a_k < \|\mathbf{Y}\|\}$, by Hölder's inequality and Jensen's inequality, it holds that

$$\|\mathbb{E}[\mathbf{Z}]\|_\infty \le \|\mathbb{E}[\mathbf{Z}]\| \le \mathbb{E}[\|\mathbf{Z}\|] \le (\mathbb{E}[\|\mathbf{Y}\|^p])^{1/p} (\Pr(\|\mathbf{Y}\| > a_k))^{(p-1)/p} \le \frac{(\mathbb{E}[\|\mathbf{Y}\|^p])^{1/p}}{2^{k(p-1)/p}}.$$

When $p \in (1, 2)$, we choose $n' := \left\lceil \frac{2^{1/p} n \log(10kd/\delta)}{\left(1 - 2^{-(\frac{1}{p} - \frac{1}{2})}\right) c^{1/p} \log(2d/\delta)} \right\rceil$ and $k := \lceil 2\log(n) \rceil$. Then, by Lemma 3.3, we can derive the error is

$$\begin{aligned}
\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty &\le \sum_{\ell=1}^{k} a_\ell \frac{2^{-(\ell-1)/2}\log(10kd/\delta)}{n'} + \|\mathbb{E}[\mathbf{Y} \cdot \mathbf{1}\{a_k < \|\mathbf{Y}\|\}]\| \\
&\le \frac{\sum_{\ell=1}^{k} 2^{\ell(\frac{1}{p}-\frac{1}{2})} \sqrt{2} (\mathbb{E}[\|\mathbf{Y}\|^p])^{1/p} \log(10kd/\delta)}{c^{1/p} n'} + \frac{(\mathbb{E}[\|\mathbf{Y}\|^p])^{1/p}}{2^{k(p-1)/p}} \\
&\le \frac{2^{k(\frac{1}{p}-\frac{1}{2})} \sqrt{2} (\mathbb{E}[\|\mathbf{Y}\|^p])^{1/p} \log(10kd/\delta)}{(1 - 2^{-(\frac{1}{p}-\frac{1}{2})}) c^{1/p} n'} + \frac{(\mathbb{E}[\|\mathbf{Y}\|^p])^{1/p}}{2^{k(p-1)/p}} \\
&\le \frac{(\mathbb{E}[\|\mathbf{Y}\|^p])^{1/p} \log(2d/\delta)}{n^{\frac{2(p-1)}{p}}} + \frac{(\mathbb{E}[\|\mathbf{Y}\|^p])^{1/p}}{n^{\frac{2(p-1)}{p}}} \\
&\le 2\frac{(\mathbb{E}[\|\mathbf{Y}\|^p])^{1/p} \log(2d/\delta)}{n^{\frac{2(p-1)}{p}}}.
\end{aligned}$$

When $p = 2$, we choose $n' = \left\lceil n \cdot \frac{\sqrt{2} k \log(10kd/\delta)}{c^{1/2} \log(2d/\delta)} \right\rceil$ and $k := \lceil 2\log(n) \rceil$. Then, by Lemma 3.3, we can derive the error is

$$\begin{aligned}
\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty &\le \sum_{\ell=1}^{k} a_\ell \frac{2^{-(\ell-1)/2}\log(10kd/\delta)}{n'} + \|\mathbb{E}[\mathbf{Y} \cdot \mathbf{1}\{a_k < \|\mathbf{Y}\|\}]\| \\
&\le \frac{\sum_{\ell=1}^{k} 2^{\ell(\frac{1}{p}-\frac{1}{2})} \sqrt{2} (\mathbb{E}[\|\mathbf{Y}\|^p])^{1/p} \log(10kd/\delta)}{c^{1/p} n'} + \frac{(\mathbb{E}[\|\mathbf{Y}\|^p])^{1/p}}{2^{k(p-1)/p}} \\
&= \frac{k\sqrt{2} (\mathbb{E}\left[\|\mathbf{Y}\|^2\right])^{1/2} \log(10kd/\delta)}{c^{1/2} n'} + \frac{(\mathbb{E}\left[\|\mathbf{Y}\|^2\right])^{1/2}}{2^{k/2}} \\
&\le \frac{(\mathbb{E}\left[\|\mathbf{Y}\|^2\right])^{1/2} \log(2d/\delta)}{n} + \frac{(\mathbb{E}\left[\|\mathbf{Y}\|^2\right])^{1/2}}{n}
\end{aligned}$$

$$\leq 2\frac{(\mathbb{E}\left[\|\mathbf{Y}\|^2\right])^{1/2}\log(2d/\delta)}{n}.$$

Therefore, for $p \in (1,2]$, we have that

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \leq 2\frac{(\mathbb{E}\left[\|\mathbf{Y}\|^p\right])^{1/p}\log(2d/\delta)}{n^{\frac{2(p-1)}{p}}},$$

and hence

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \frac{2\sqrt{d}(\mathbb{E}\left[\|\mathbf{Y}\|^p\right])^{1/p}\log(2d/\delta)}{n^{\frac{2(p-1)}{p}}}.$$

Note that $\mathbf{Y} := \mathbf{X} - \boldsymbol{\zeta}$. Then we know that

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{Y}\|^p\right] &= \mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\zeta}\|^p\right] \\ &\leq 2^{p-1}\left(\mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\mu}\|^p\right] + \|\boldsymbol{\mu} - \boldsymbol{\zeta}\|^p\right)\end{aligned}$$

By the definition of $n'' := \left\lceil 2\left\lceil 24\ln(2/\delta)\right\rceil 3^{\frac{p}{p-1}}(8C_p)^{\frac{1}{p-1}}\right\rceil$ and the performance guarantee by CHTME in Theorem 2.7, we can know that

$$\|\boldsymbol{\mu} - \boldsymbol{\zeta}\| \leq (\mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\mu}\|^p\right])^{1/p}.$$

Therefore, it holds that

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{Y}\|^p\right] &= \mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\zeta}\|^p\right] \\ &\leq 2^{p-1}\left(\mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\mu}\|^p\right] + \|\boldsymbol{\mu} - \boldsymbol{\zeta}\|^p\right) \\ &\leq 2^p\mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\mu}\|^p\right]\end{aligned}$$

and

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \leq 4\frac{(\mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\mu}\|^p\right])^{1/p}\log(2d/\delta)}{n^{\frac{2(p-1)}{p}}}.$$

and hence

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \frac{4\sqrt{d}(\mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\mu}\|^p\right])^{1/p}\log(2d/\delta)}{n^{\frac{2(p-1)}{p}}}.$$

Last but not least, the total number of queries to the quantum sampling oracle $O_{\mathbf{X}}$ for $p \in (1,2)$ and $p = 2$ only differ by at most a logarithmic factor in n. Note that when calling the classical mean estimator CHTME, we need n'' classical samples. In fact, these n'' classical samples can be obtained via querying the quantum sampling oracle $O_{\mathbf{X}}$ n'' times and performing the measurements in the computational basis. In both cases, the query complexity satisfies

$$\tilde{\mathcal{O}}(\underbrace{k2^{k/2}\log(k/\delta)}_{\texttt{Quantile}} + \underbrace{kn'}_{\texttt{QBounded}} + \underbrace{n''}_{\texttt{CHTME}}) = \tilde{\mathcal{O}}(n).$$

□

C.3 A Variant of Quantum Heavy-Tailed Mean Estimator QHTME$^+$

We first derive a variant of the quantum mean estimator QHTME, denoted by QHTME$^+$. This estimator has no failure event and outputs an estimate $\tilde{\boldsymbol{\mu}}$ satisfying $\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^p\right] \leq \epsilon^p$ deterministically, where the expectation is taken over all randomness of the algorithm. The key technical ingredient is a new theoretical insight formalized in Lemma C.6. This lemma shows that, by using a carefully designed acceptance-rejection rule, two random variables $\mathbf{X}_2$ and $\mathbf{X}_3$ with bounded p-th central moments can be used to control the p-th raw moment of the estimation error of a third random variable $\mathbf{X}_1$, provided that $\mathbf{X}_1$ is sufficiently close to the common expectation of $\mathbf{X}_2$ and $\mathbf{X}_3$. A related idea appeared in [46], where an analogous result was established for the special case $p = 2$. In contrast, we generalize their result to all $p \in (1,2]$ in Lemma C.6.

Lemma C.6. *Let $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 \in \mathbb{R}^d$ be independent random variables such that $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}_2] = \mathbb{E}[\mathbf{X}_3]$, $\mathbb{E}[\|\mathbf{X}_2 - \boldsymbol{\mu}\|^p] \leq \sigma^p$, and $\mathbb{E}[\|\mathbf{X}_3 - \boldsymbol{\mu}\|^p] \leq \sigma^p$, for some $p \in (1, 2]$ and $\sigma \geq 0$. Assume further that $\|\mathbf{X}_1 - \boldsymbol{\mu}\| \leq \epsilon$ with probability at least $1 - \delta$, for some $\epsilon > 0$ and $\delta \in (0, 1)$. Define a random variable $\mathbf{Z}$ by $\mathbf{Z} = \mathbf{X}_1$ if $\|\mathbf{X}_1 - \mathbf{X}_2\| \leq D$ and $\mathbf{Z} = \mathbf{X}_3$ otherwise, where $D > \epsilon$. Then,*

$$\mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p] \leq \epsilon^p + \frac{\sigma^{p+1}}{D - \epsilon} + \delta(2^{p-1}D^p + (2^{p-1} + 1)\sigma^p).$$

Therefore, if $D := \epsilon + \frac{\sigma^{p+1}}{\epsilon^p}$, $\epsilon < \sigma$, and $\delta \leq (\epsilon/\sigma)^{(p^2+p)}$, then $\mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p] \leq 13 \cdot \epsilon^p$.

Proof. Our proof idea is inspired by Sidford and Zhang [46] but is more general than their result. Let S denote the event that $\|\mathbf{X}_1 - \boldsymbol{\mu}\| \leq \epsilon$ and let T denote the event that $\|\mathbf{X}_1 - \mathbf{X}_2\| \leq D$. Then, by the law of total expectation, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p] &= \Pr(S)\mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid S] + \Pr(\overline{S})\mathbb{E}\left[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid \overline{S}\right] \\ &\leq \mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid S] + \delta\mathbb{E}\left[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid \overline{S}\right],\end{aligned}$$

where we used the fact that $1 - \delta \leq \Pr(S) \leq 1$ and $0 \leq \Pr(\overline{S}) = 1 - \Pr(S) \leq \delta$. We next prove the lemma by further upper bounding both $\mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid S]$ and $\mathbb{E}\left[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid \overline{S}\right]$.

First, we upper bound $\mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid S]$. By the law of total expectation and the definition of $\mathbf{Z}$, we have that

$$\begin{aligned}\mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid S] &= \Pr(T \mid S)\mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid T \cap S] + \Pr(\overline{T} \mid S)\mathbb{E}\left[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid \overline{T} \cap S\right] \\ &\leq \Pr(T \mid S)\epsilon^p + \Pr(\overline{T} \mid S)\sigma^p.\end{aligned}$$

The second line follows from the facts that

- if T and S both hold, then $\mathbf{Z} = \mathbf{X}_1$ and $\|\mathbf{X}_1 - \boldsymbol{\mu}\| \leq \epsilon$, indicating that $\mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid T \cap S] \leq \epsilon^p$;
- if T does not hold but S holds, then $\mathbf{Z} = \mathbf{X}_3$ and $\mathbf{X}_3$ is independent of S and T, indicating that $\mathbb{E}\left[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid \overline{T} \cap S\right] = \mathbb{E}[\|\mathbf{X}_3 - \boldsymbol{\mu}\|^p] \leq \sigma^p$.

Now, we proceed to upper bound $\Pr(\overline{T} \mid S)$. Note that, if $\|\mathbf{X}_2 - \boldsymbol{\mu}\| \leq D - \epsilon$, then when S holds ($\|\mathbf{X}_1 - \boldsymbol{\mu}\| \leq \epsilon$), we have that

$$\|\mathbf{X}_1 - \mathbf{X}_2\| \leq \|\mathbf{X}_1 - \boldsymbol{\mu}\| + \|\mathbf{X}_2 - \boldsymbol{\mu}\| \leq D,$$

by the triangle inequality. Therefore, by applying Markov's inequality to $\mathbf{X}_2$ and using that $\mathbf{X}_2$ is independent of S, we have that

$$\Pr(T \mid S) \geq \Pr(\|\mathbf{X}_2 - \boldsymbol{\mu}\| \leq D - \epsilon \mid S) = \Pr(\|\mathbf{X}_2 - \boldsymbol{\mu}\| \leq D - \epsilon) \geq 1 - \frac{\mathbb{E}[\|\mathbf{X}_2 - \boldsymbol{\mu}\|]}{D - \epsilon}.$$

Note that $\mathbb{E}[\|\mathbf{X}_2 - \boldsymbol{\mu}\|] \leq (\mathbb{E}[\|\mathbf{X}_2 - \boldsymbol{\mu}\|^p])^{\frac{1}{p}} \leq \sigma$ by Hölder's inequality and the upper bound of $\mathbb{E}[\|\mathbf{X}_2 - \boldsymbol{\mu}\|^p]$. This implies that

$$\mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid S] \leq \epsilon^p + \frac{\sigma^{p+1}}{D - \epsilon}.$$

Next, we upper bound $\mathbb{E}\left[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid \overline{S}\right]$. Note that if T holds ($\|\mathbf{X}_1 - \mathbf{X}_2\| \leq D$), then

$$\begin{aligned}\|\mathbf{Z} - \boldsymbol{\mu}\|^p &\leq 2^{p-1}(\|\mathbf{Z} - \mathbf{X}_2\|^p + \|\mathbf{X}_2 - \boldsymbol{\mu}\|^p) \\ &= 2^{p-1}(\|\mathbf{X}_1 - \mathbf{X}_2\|^p + \|\mathbf{X}_2 - \boldsymbol{\mu}\|^p) \\ &\leq 2^{p-1}(D^p + \|\mathbf{X}_2 - \boldsymbol{\mu}\|^p).\end{aligned}$$

If T does not hold, then $\mathbf{Z} = \mathbf{X}_3$. Therefore, in either case, it holds that

$$\mathbb{E}\left[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid \overline{S}\right] \leq \mathbb{E}\left[2^{p-1}D^p + 2^{p-1}\|\mathbf{X}_2 - \boldsymbol{\mu}\|^p + \|\mathbf{X}_3 - \boldsymbol{\mu}\|^p\right] = 2^{p-1}D^p + (2^{p-1} + 1)\sigma^p.$$

Therefore, we have

$$\mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p] \leq \mathbb{E}[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid S] + \delta\mathbb{E}\left[\|\mathbf{Z} - \boldsymbol{\mu}\|^p \mid \overline{S}\right]$$

$$\leq \epsilon^p + \frac{\sigma^{p+1}}{D-\epsilon} + \delta(2^{p-1}D^p + (2^{p-1}+1)\sigma^p).$$

Suppose $\epsilon < \sigma$, by choosing $D = \epsilon + \frac{\sigma^{p+1}}{\epsilon^p}$ and $\delta = (\epsilon/\sigma)^{p^2+p}$, we can compute that

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{Z}-\boldsymbol{\mu}\|^p\right] &\leq \epsilon^p + \frac{\sigma^{p+1}}{D-\epsilon} + \delta(2^{p-1}D^p + (2^{p-1}+1)\sigma^p) \\
&\leq \epsilon^p + \epsilon^p + \delta(2^{p-1}2^{p-1}(\epsilon^p + \frac{\sigma^{p(p+1)}}{\epsilon^{p^2}}) + (2^{p-1}+1)\sigma^p) \\
&\leq \epsilon^p + \epsilon^p + 2^{2(p-1)}\frac{\epsilon^{p^2+2p}}{\sigma^{p(p+1)}} + 2^{2(p-1)}\epsilon^p + (2^{p-1}+1)\frac{\epsilon^{p^2+p}}{\sigma^{p^2}} \\
&\leq (3 + 2^{2(p-1)} + 2^{2(p-1)} + 2^{p-1})\epsilon^p \\
&\leq 13\epsilon^p.
\end{aligned}$$

The second line uses the definition of $D = \epsilon + \frac{\sigma^{p+1}}{\epsilon^p}$ and Jensen's inequality. The third line uses the definition of $\delta = \frac{\epsilon^{p^2+p}}{\sigma^{(p+1)p}}$. The fourth line uses the fact that $\epsilon < \sigma$. The fifth line uses the fact that $p \in (1, 2]$. □

Algorithm 6 $\texttt{QHTME}^+(\mathbf{X}, \sigma, p, d, \epsilon)$

Input: random variable $\mathbf{X}$, heavy-tailed parameter σ, tail index p, problem dimension d, error $\epsilon \in (0, \sigma)$.
$\delta \leftarrow (\epsilon/(13\sigma))^{p(p+1)}$, $D \leftarrow \epsilon/13 + 13^p\sigma^{p+1}/\epsilon^p$.
Set $\mathbf{X}_1 \leftarrow \texttt{QHTME}(\mathbf{X}, n(\epsilon/13), p, d, \delta)$
Randomly draw one classical sample $\mathbf{X}_2$ of $\mathbf{X}$
if $\|\mathbf{X}_1 - \mathbf{X}_2\| \leq D$ **then**
    $\tilde{\boldsymbol{\mu}} \leftarrow \mathbf{X}_1$
else
    Randomly draw one classical sample $\mathbf{X}_3$ of $\mathbf{X}$
    $\tilde{\boldsymbol{\mu}} \leftarrow \mathbf{X}_3$
end if
Output: $\tilde{\boldsymbol{\mu}}$

We are ready to present a refined quantum mean estimator, denoted $\texttt{QHTME}^+$, given in Algorithm 6. This algorithm outputs an estimate for a heavy-tailed random variable whose estimation error has a bounded p-th raw moment with probability 1. In $\texttt{QHTME}^+$, $\mathbf{X}_1$ is generated by the quantum mean estimator QHTME (Algorithm 1), while $\mathbf{X}_2$ and $\mathbf{X}_3$ are obtained by querying the quantum sampling oracle $O_{\mathbf{X}}$ and measuring the resulting quantum states in the computational basis. Since generating $\mathbf{X}_2$ and $\mathbf{X}_3$ requires only two additional queries to $O_{\mathbf{X}}$, the total query complexity of $\texttt{QHTME}^+$ differs from that of QHTME by at most a constant factor. We summarize the performance guarantee of $\texttt{QHTME}^+$ in Lemma C.7.

Lemma C.7. *Let $\mathbf{X}$ be a d-dimensional random variable with mean $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}]$ satisfying $\mathbb{E}[\|\mathbf{X}-\boldsymbol{\mu}\|^p] \leq \sigma^p$, for some $p \in (1, 2]$ and $\sigma \geq 0$. Then, for any $\epsilon \in (0, \sigma)$, Algorithm 6, $\texttt{QHTME}^+(\mathbf{X}, \sigma, p, d, \epsilon)$, uses*

$$\tilde{\mathcal{O}}\left(d^{\frac{p}{4(p-1)}}\sigma^{\frac{p}{2(p-1)}}\epsilon^{-\frac{p}{2(p-1)}}\right)$$

queries to the quantum sampling oracle $O_{\mathbf{X}}$ and outputs an estimate $\tilde{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ satisfying $\mathbb{E}[\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^p] \leq \epsilon^p$.

Proof. Note that the random variable $\mathbf{X}$ satisfies $\mathbb{E}[\|\mathbf{X}-\boldsymbol{\mu}\|^p] \leq \sigma^p$. By using Lemma C.6, and replacing ϵ, σ with $\frac{\epsilon}{13}$ and σ, respectively, the output $\tilde{\boldsymbol{\mu}}$ of $\texttt{QHTME}^+$ satisfies

$$\mathbb{E}[\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^p] \leq \epsilon^p,$$

where $\epsilon \in (0, \sigma)$.

As for the query complexity, Algorithm 6 only calls $\texttt{QHTME}(\mathbf{X}, n(\epsilon/13), p, d, \delta)$ once and one can prepare the classical samples $\mathbf{X}_2$ and $\mathbf{X}_3$ by querying the quantum sampling oracle $O_{\mathbf{X}}$ once for each

and performing the measurements on the computational basis. Since we have that $\mathbb{E}\left[\|\mathbf{X}-\boldsymbol{\mu}\|^p\right] \leq \sigma^p$, by Theorem C.5, the overall query complexity equals

$$\tilde{O}\left(\frac{d^{\frac{p}{4(p-1)}}\sigma^{\frac{p}{2(p-1)}}}{\epsilon^{\frac{p}{2(p-1)}}}\right).$$

□

C.4 Proof of Theorem 3.7

Proof. Our proof mainly follows Sidford and Zhang [46]. Observe that the output $\tilde{\boldsymbol{\mu}}$ of Algorithm 2 can be expressed as

$$\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}_0 + 2^J(\tilde{\boldsymbol{\mu}}_J - \tilde{\boldsymbol{\mu}}_{J-1}), \quad J \sim \mathrm{Geom}\left(\frac{1}{2}\right) \in \mathbb{N}.$$

Then, we have an unbiased output satisfying

$$\mathbb{E}\left[\tilde{\boldsymbol{\mu}}\right] = \mathbb{E}\left[\tilde{\boldsymbol{\mu}}_0\right] + \sum_{j=1}^{\infty}\Pr(J=j)2^j(\mathbb{E}\left[\tilde{\boldsymbol{\mu}}_j\right] - \mathbb{E}\left[\tilde{\boldsymbol{\mu}}_{j-1}\right]) = \mathbb{E}\left[\tilde{\boldsymbol{\mu}}_\infty\right] = \boldsymbol{\mu},$$

where we used the fact that $\Pr(J=j) = 2^{-j}$ in the second equality. By the triangle inequality and Jensen's inequality ($(a+b)^p \leq 2^{p-1}(a^p+b^p), \forall p \in (1,2]$), we have

$$\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^p\right] \leq 2^{p-1}\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_0\|^p\right] + 2^{p-1}\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}\|^p\right],$$

where

$$\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_0\|^p\right] = \sum_{j=1}^{\infty}\Pr(J=j)2^{pj}\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}}_j - \tilde{\boldsymbol{\mu}}_{j-1}\|^p\right] = \sum_{j=1}^{\infty}2^{(p-1)j}\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}}_j - \tilde{\boldsymbol{\mu}}_{j-1}\|^p\right].$$

For each j, we have

$$\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}}_j - \tilde{\boldsymbol{\mu}}_{j-1}\|^p\right] \leq 2^{p-1}\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}}_j - \boldsymbol{\mu}\|^p\right] + 2^{p-1}\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}}_{j-1} - \boldsymbol{\mu}\|^p\right].$$

By Lemma C.7 and line 4 in Algorithm 2, we have

$$\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}}_j - \boldsymbol{\mu}\|^p\right] \leq \frac{\epsilon^p}{c\cdot 2^{\frac{3(p-1)j}{2}}}, \forall j \geq 1,$$

which indicates that

$$\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}}_j - \tilde{\boldsymbol{\mu}}_{j-1}\|^p\right] \leq \frac{2^{p-1}\epsilon^p}{c\cdot 2^{3(p-1)(j-1)/2}} + \frac{2^{p-1}\epsilon^p}{c\cdot 2^{3(p-1)j/2}} \leq \frac{8\epsilon^p}{c\cdot 2^{3(p-1)j/2}}.$$

The second inequality above follows from the facts that $2^{p-1} \leq 2$ for all $p \in (1,2]$ and

$$\frac{2^{p-1}}{c\cdot 2^{3(p-1)(j-1)/2}} \leq \frac{2}{c\cdot 2^{3(p-1)(j-1)/2}} \leq \frac{2\cdot 2^{\frac{3(p-1)}{2}}}{c\cdot 2^{3(p-1)j/2}} \leq \frac{4\sqrt{2}}{c\cdot 2^{3(p-1)j/2}} \leq \frac{6}{c\cdot 2^{3(p-1)j/2}}.$$

Therefore, since we let $c := 24/(2^{\frac{p-1}{2}}-1)$ in line 2 of Algorithm 2 and the series $\sum_{j=1}^{\infty}\frac{1}{2^{(p-1)j/2}}$ converges to $\frac{1}{2^{(p-1)/2}-1}$ for any given $p \in (1,2]$, then it holds that

$$\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_0\|^p\right] \leq \sum_{j=1}^{\infty}2^{(p-1)j}\frac{8\epsilon^p}{c\cdot 2^{3(p-1)j/2}} \leq \sum_{j=1}^{\infty}\frac{8\epsilon^p}{c\cdot 2^{(p-1)j/2}} \leq \frac{1}{3}\epsilon^p.$$

Since it holds that $\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}\|^p\right] \leq \frac{\epsilon^p}{c} = \frac{(2^{\frac{p-1}{2}}-1)\epsilon^p}{24} \leq \frac{\epsilon^p}{10}$ by Lemma C.7 and line 2 in Algorithm 2, then we must have

$$\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^p\right] \leq 2^{p-1}\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_0\|^p\right] + 2^{p-1}\mathbb{E}\left[\|\tilde{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}\|^p\right] \leq \frac{2^{p-1}}{3}\epsilon^p + \frac{2^{p-1}}{10}\epsilon^p \leq \epsilon^p.$$

Moreover, since it holds that, for a given $p \in (1, 2]$

$$\sum_{j=1}^{\infty} \Pr(J = j) 2^{\frac{3j}{4}} = \sum_{j=1}^{\infty} 2^{-j} 2^{\frac{3j}{4}} = \sum_{j=1}^{\infty} 2^{\frac{-j}{4}} = \frac{1}{2^{\frac{1}{4}} - 1}$$

and

$$\sum_{j=1}^{\infty} \Pr(J = j) 2^{\frac{3(j-1)}{4}} = \sum_{j=1}^{\infty} 2^{-j} 2^{\frac{3(j-1)}{4}} = \frac{1}{2 - 2^{3/4}}$$

are finite constants. The expected number of queries is

$$\tilde{O}\left(\frac{d^{\frac{p}{4(p-1)}} \sigma^{\frac{p}{2(p-1)}}}{\epsilon^{\frac{p}{2(p-1)}}}\right) \cdot \left(1 + \sum_{j=1}^{\infty} \Pr(J = j) \cdot \left(2^{\frac{3j}{4}} + 2^{\frac{3(j-1)}{4}}\right)\right) = \tilde{O}\left(\frac{d^{\frac{p}{4(p-1)}} \sigma^{\frac{p}{2(p-1)}}}{\epsilon^{\frac{p}{2(p-1)}}}\right).$$

□

C.5 Lower Bounds on Mean Estimation of Heavy-Tailed Random Variables

Theorem C.8 (Lower bounds for heavy-tailed mean estimation)**.** *Fix $p \in (1, 2]$ and $\sigma > 0$. Consider the problem of estimating the mean $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ of a d-dimensional real-valued random variable $\mathbf{X}$ satisfying $\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p] \leq \sigma^p$. For any estimator which, with probability at least $2/3$, outputs a mean estimate $\tilde{\boldsymbol{\mu}}$ satisfying $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \epsilon$, the query complexity is*

$$\Omega\left(\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}\right)$$

by using the classical sampling oracle, and

$$\Omega\left(\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{2(p-1)}}\right)$$

by using the quantum sampling oracle.

Proof. It suffices to prove the result for the case $d = 1$, since the general d-dimensional statement follows by considering random vectors supported entirely on the first coordinate. In the one-dimensional setting, our argument is a direct adaptation of the proof technique in [4, Section 6], replacing the bounded p-th raw moment assumption with a bounded p-th central moment assumption. For completeness, we include the full proof here.

We consider reducing our mean estimation problem to the unstructured search problem. Let $N \geq 2$ be an integer to be chosen later, and consider Boolean functions $f : [N] := \{1, \ldots, N\} \to \{0, 1\}$ under the promise that either

$$f(i) = 0 \quad \text{for all } i \in [N],$$

or there exists a unique $j \in [N]$ such that

$$f(j) = 1, \qquad f(i) = 0 \text{ for all } i \neq j.$$

As established in [4, Theorem 5], with probability at least $2/3$, distinguishing these two cases requires $\Omega(N)$ classical queries [2] and $\Omega(\sqrt{N})$ quantum queries [3].

Given such an f, define a random variable

$$\mathbf{X}_f := Mf(U),$$

where U is uniformly distributed on $[N]$ and $M > 0$ is a parameter. One query to a classical sampling oracle for $\mathbf{X}_f$ can be simulated by randomly choosing a number in $[N]$ and then using one query to f; similarly, in the quantum setting, one query to the corresponding quantum sampling oracle

$$O_{\mathbf{X}_f} : |0\rangle |0\rangle \to \sum_{i=1}^{N} \frac{1}{\sqrt{N}} |Mf(i)\rangle |i\rangle$$

can be simulated by preparing a uniform superposition by Hadamard gates and one query to the binary oracle $\mathcal{B}_f$ for f satisfying $\mathcal{B}_f : |i\rangle |0\rangle \to |i\rangle |f(i)\rangle$. Hence, any mean-estimation algorithm for

the class under consideration yields a search-distinguishing algorithm with the same order of query complexity.

If $f \equiv 0$, then $\mathbf{X}_f = 0$ almost surely, so

$$\boldsymbol{\mu}_f := \mathbb{E}[\mathbf{X}_f] = 0, \qquad \mathbb{E}[|\mathbf{X}_f - \boldsymbol{\mu}_f|^p] = 0.$$

If f has a unique marked element, then

$$\mathbb{P}(\mathbf{X}_f = M) = \frac{1}{N}, \qquad \mathbb{P}(\mathbf{X}_f = 0) = 1 - \frac{1}{N},$$

and therefore $\boldsymbol{\mu}_f = \mathbb{E}[\mathbf{X}_f] = \frac{M}{N}$. By choosing $M := 3\epsilon N$, we know that, in the marked case, $\boldsymbol{\mu}_f = 3\epsilon$, whereas in the unmarked case $\boldsymbol{\mu}_f = 0$.

We now verify the moment constraint. In the marked case,

$$\begin{aligned}\mathbb{E}[|\mathbf{X}_f - \boldsymbol{\mu}_f|^p] &= \frac{1}{N}\left|M - \frac{M}{N}\right|^p + \left(1 - \frac{1}{N}\right)\left|\frac{M}{N}\right|^p \\ &\leq \frac{M^p}{N} + \frac{M^p}{N^p} \leq \frac{2M^p}{N}.\end{aligned}$$

Substituting $M = 3\epsilon N$ gives

$$\mathbb{E}[|\mathbf{X}_f - \boldsymbol{\mu}_f|^p] \leq 2 \cdot 3^p \epsilon^p N^{p-1}.$$

Thus the condition $\mathbb{E}[|\mathbf{X}_f - \boldsymbol{\mu}_f|^p] \leq \sigma^p$ holds whenever

$$N \leq c_p \left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}},$$

for a constant $c_p > 0$ depending only on p.

Now suppose there exists a mean-estimation algorithm which, for every distribution satisfying the above moment condition, outputs $\tilde{\boldsymbol{\mu}}$ with $|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_f| \leq \epsilon$ with probability at least $2/3$ using T queries. We use it to distinguish the two search cases as follows: run the estimator on $\mathbf{X}_f$ and output "marked" if $|\tilde{\boldsymbol{\mu}}| \geq \frac{3\epsilon}{2}$, and output "unmarked" otherwise.

If $f \equiv 0$, then $\boldsymbol{\mu}_f = 0$, so on the success event, $|\tilde{\boldsymbol{\mu}}| \leq \epsilon < \frac{3\epsilon}{2}$. If f has a unique marked element, then $\boldsymbol{\mu}_f = 3\epsilon$, so on the success event, $|\tilde{\boldsymbol{\mu}}| \geq 3\epsilon - \epsilon = 2\epsilon > \frac{3\epsilon}{2}$. Therefore, the induced algorithm distinguishes the two promised search cases with probability at least $2/3$ using T queries.

By the classical lower bound for unstructured search [2], this implies

$$T = \Omega(N) = \Omega\left(\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}\right)$$

in the classical setting. By Grover's lower bound [3], it implies

$$T = \Omega(\sqrt{N}) = \Omega\left(\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{2(p-1)}}\right)$$

in the quantum query model. This proves the claim. □

Corollary C.9 (Expected-query lower bound for unbiased heavy-tailed mean estimation)**.** *Fix $p \in (1, 2]$ and $\sigma > 0$. Consider the problem of estimating $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ for every random variable $\mathbf{X}$ satisfying $\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^p] \leq \sigma^p$. Suppose an algorithm outputs an estimator $\tilde{\boldsymbol{\mu}}$ satisfying*

$$\mathbb{E}[\tilde{\boldsymbol{\mu}}] = \boldsymbol{\mu}, \qquad \mathbb{E}[\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^p] \leq \epsilon^p.$$

Then its expected query complexity is at least

$$\Omega\left(\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}\right)$$

by using the classical sampling oracle, and at least

$$\Omega\left(\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{2(p-1)}}\right)$$

by using the quantum sampling oracle.

Proof. We prove the quantum case; the classical case follows analogously.

Let $\mathcal{A}$ be a quantum algorithm satisfying the stated guarantee, and let $\mathbb{E}[Q]$ denote the worst-case expected number of oracle queries made by $\mathcal{A}$. Set $T := \mathbb{E}[Q]$.

By Markov's inequality, we have

$$\Pr(Q > 6T) \leq \frac{1}{6}.$$

On the other hand, from $\mathbb{E}[\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^p] \leq \epsilon^p$, Markov's inequality gives

$$\Pr\Big(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| > 6^{1/p}\epsilon\Big) \leq \frac{1}{6}.$$

Therefore, by the union bound, we have

$$\Pr\Big(Q \leq 6T \text{ and } \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq 6^{1/p}\epsilon\Big) \geq 1 - \frac{1}{6} - \frac{1}{6} = \frac{2}{3}.$$

Now define a truncated algorithm $\mathcal{A}'$ as follows: run $\mathcal{A}$, but stop it once it has made more than $6T$ oracle queries. If this happens, output an arbitrary value. Then $\mathcal{A}'$ makes at most $6T$ queries in the worst case. Moreover, by the inequality above, $\mathcal{A}'$ outputs an estimate satisfying

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq 6^{1/p}\epsilon$$

with probability at least $2/3$.

Thus, $\mathcal{A}'$ is a valid high-probability mean estimator with accuracy $6^{1/p}\epsilon$ and worst-case query complexity at most $6T$. By the high-probability quantum lower bound in Theorem C.8, we have

$$T \geq \Omega\left(\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{2(p-1)}}\right).$$

The classical case follows in the same way, replacing the quantum high-probability lower bound by the classical one in Theorem C.8, which yields

$$\mathbb{E}[Q] \geq \Omega\left(\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}\right).$$

□

Furthermore, we establish a stronger quantum lower bound in the high-dimensional regime, demonstrating that quantum algorithms offer no asymptotic advantage over classical methods for solving Problems 3.1 and 3.2 under a low-precision condition. This phenomenon parallels known results in the bounded-variance setting [10, Theorem 3.7 and 3.8]. Our lower bound is obtained via a reduction to the $\texttt{Search}^N \circ \texttt{Parity}^M$ problem.

Problem C.10 ($\texttt{Search}^N \circ \texttt{Parity}^M$)**.** *Let $N, M \geq 1$ be two integers. Let $\mathcal{A}_{N,M}$ denote the set of all matrices $A \in \{0,1\}^{N\times M}$ such that $\lfloor N/2 \rfloor$ rows have Hamming weights $\lfloor M/2 \rfloor$, and the other rows have Hamming weights $\lfloor M/2 \rfloor + 1$. Define the vector $b^{(A)} \in \{0,1\}^N$ such that, for each $i \in [N]$,*

$$b_i^{(A)} = \begin{cases} 0, & \text{if the } i\text{-th row of } A \text{ has Hamming weight } \lfloor M/2 \rfloor, \\ 1, & \text{if the } i\text{-th row of } A \text{ has Hamming weight } \lfloor M/2 \rfloor + 1. \end{cases}$$

Then, the $\texttt{Search}^N \circ \texttt{Parity}^M$ problem consists of finding a vector $\tilde{b} \in \mathbb{R}^N$ that minimizes $\|\tilde{b} - b^{(A)}\|$ given a quantum oracle

$$O_A : |i,j\rangle \mapsto (-1)^{A_{i,j}} |i,j\rangle$$

to access the matrix $A \in \mathcal{A}_{N,M}$.

The quantum lower bound for the $\texttt{Search}^N \circ \texttt{Parity}^M$ problem is already well-established in [50, Lemma 11], [9, Lemma 5.7] and [10, Lemma 3.6], which states that $\Omega(NM)$ queries to the oracle O_A are needed to approximate the vector $b^{(A)}$ with small error.

Lemma C.11. *Let $\alpha > 1$ be a sufficiently large constant. Consider any quantum algorithm for the $\texttt{Search}^N \circ \texttt{Parity}^M$ problem that uses at most NM/α queries on all inputs. Then, there exists an input $A \in \mathcal{A}_{N,M}$ such that this algorithm returns a vector $\tilde{b} \in \mathbb{R}^N$ satisfying*

$$\|\tilde{b} - b^{(A)}\| \geq \Omega(\sqrt{N})$$

with probability at least $2/3$.

Theorem C.12 (Quantum lower bound of heavy-tailed mean estimation in the high-dimensional regime)**.** *There exists a sufficiently large constant $\alpha > 1$ such that the following holds. Let $p \in (1, 2]$, and consider two integers n, d satisfying*

$$n \leq d^{p/2},$$

and, when $p \in (1, 2)$, additionally

$$n \leq \alpha^{p/(2-p)}.$$

Fix $\sigma > 0$ and let $\mathcal{P}_\sigma$ denote the set of all d-dimensional random variables $\mathbf{X}$ that satisfy $\mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\mu}\|^p\right] \leq \sigma^p$ for some $p \in (1, 2]$. Then, for any quantum estimator that uses at most n queries to the quantum sampling oracle $O_{\mathbf{X}}$, there exists $\mathbf{X} \in \mathcal{P}_\sigma$ such that the estimator returns a mean estimate $\tilde{\boldsymbol{\mu}}$ of $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ that satisfies

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \geq \Omega\left(\frac{\sigma}{n^{\frac{p-1}{p}}}\right)$$

with probability at least $2/3$.

Proof. We assume for simplicity that d is a power of two; the general case can be handled by a standard padding argument. Since $n \leq d^{p/2}$ by assumption, we have $n^{\frac{2}{p}} \leq d$.

To account for the fact that $n^{\frac{2}{p}}$ may not be an integer, define $N := 2^{\lfloor \frac{2}{p} \log_2 n \rfloor}$, which satisfies $N \leq n^{\frac{2}{p}} \leq 2N$. Next, define the even integer $M := 2\left\lceil \frac{\alpha n}{2N} \right\rceil$. Since $N \leq n^{\frac{2}{p}}$, we have $\frac{\alpha n}{N} \geq \alpha n^{1-\frac{2}{p}}$. When $p \in (1, 2)$, by assumption $n \leq \alpha^{p/(2-p)}$ and therefore $\alpha n^{1-2/p} \geq 1$. When $p = 2$, $\alpha n^{1-\frac{2}{p}} = \alpha > 1$. Therefore, in all cases, $\frac{\alpha n}{N} \geq 1$. It follows that

$$\alpha n \leq NM \leq \alpha n + 2N \leq 3\alpha n. \tag{7}$$

Consider the partial Hadamard matrix $H \in \mathbb{R}^{N \times d}$ such that

$$H_{i,j} = \frac{1}{\sqrt{d}}(-1)^{\langle i,j \rangle},$$

where $i, j \in \{0,1\}^{\log_2 d}$, i.e., H consists of the first $N \leq n^{\frac{2}{p}} \leq d$ rows of the standard $d \times d$ Hadamard matrix. Note that the matrix H satisfies $HH^\top = I_N$, and has spectral norm $\|H\| = \sqrt{\lambda_{\max}(H^\top H)} = \sqrt{\lambda_{\max}(HH^\top)} = 1$.

Let $(\Omega, 2^\Omega, \mathbb{P})$ be the probability space such that

$$\Omega = [N] \times [M], \qquad \mathbb{P}(\omega) = \frac{1}{NM} \quad \text{for all } \omega \in \Omega.$$

For any input $A \in \mathcal{A}_{N,M}$ to the problem $\mathrm{Search}^N \circ \mathrm{Parity}^M$, define the random variable $\mathbf{X}^{(A)} : \Omega \to \mathbb{R}^d$ by

$$\mathbf{X}^{(A)}(i, j) = c\alpha\sigma\,(-1)^{1+A_{i,j}} H_i, \tag{8}$$

where $H_i \in \mathbb{R}^d$ is the i-th row of H and constant $c := \left(2\left(\alpha^p + 2^{p/2}\right)\right)^{-1/p}$.

The expectation of $\mathbf{X}^{(A)}$ is

$$\begin{aligned}
\mathbb{E}[\mathbf{X}^{(A)}] &= \sum_{\omega \in \Omega} \mathbb{P}(\omega)\mathbf{X}^{(A)}(\omega) \\
&= \frac{1}{NM} \cdot c\alpha\sigma \sum_{i=1}^{N} H_i \sum_{j=1}^{M} (-1)^{1+A_{i,j}} \qquad\qquad \mathbb{P}(\omega) = 1/(NM)
\end{aligned}$$

$$= \frac{2c\alpha\sigma}{NM} \sum_{i=1}^{N} H_i \, b_i^{(A)} \qquad M \text{ is even}$$
$$= \frac{2c\alpha\sigma}{NM} H^\top b^{(A)}. \tag{9}$$

For simplicity, we use $\boldsymbol{\mu}^{(A)}$ to denote $\mathbb{E}\left[\mathbf{X}^{(A)}\right]$ below. We now verify that the p-th central moment of $\mathbf{X}^{(A)}$ satisfies $\mathbb{E}\left[\left\|\mathbf{X}^{(A)} - \boldsymbol{\mu}^{(A)}\right\|^p\right] \leq \sigma^p$. By the triangle inequality and the bound $(a+b)^p \leq 2^{p-1}(a^p + b^p)$ for $p \in (1, 2]$,

$$\mathbb{E}\left[\left\|\mathbf{X}^{(A)} - \boldsymbol{\mu}^{(A)}\right\|^p\right] \leq 2^{p-1}\left(\mathbb{E}\left[\left\|\mathbf{X}^{(A)}\right\|^p\right] + \left\|\boldsymbol{\mu}^{(A)}\right\|^p\right).$$

For the term $\mathbb{E}\left[\left\|\mathbf{X}^{(A)}\right\|^p\right]$, we compute

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{X}^{(A)}\right\|^p\right] \\
&= \sum_{\omega\in\Omega} \mathbb{P}(\omega)\left\|\mathbf{X}^{(A)}(\omega)\right\|^p \\
&\leq \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} c^p\alpha^p\sigma^p \left(\left|(-1)^{(1+A_{i,j})}\right|\right)^p \|H_i\|^p && \text{by (8)} \\
&\leq \frac{1}{NM}\sum_{i=1}^{N} c^p\alpha^p\sigma^p \|H_i\|^p \; M && \left(\left|(-1)^{(1+A_{i,j})}\right|\right)^p \leq 1 \\
&\leq \frac{1}{NM} c^p\alpha^p\sigma^p NM && \|H_i\|^p = (\|H_i\|^2)^{p/2} = (H_iH_i^\top)^{p/2} = 1 \\
&= c^p\alpha^p\sigma^p.
\end{aligned}$$

For the other term $\left\|\boldsymbol{\mu}^{(A)}\right\|^p$, we compute

$$\begin{aligned}
\left\|\boldsymbol{\mu}^{(A)}\right\|^p &= \left\|\frac{2c\alpha\sigma}{NM}H^\top b^{(A)}\right\|^p && \text{by (9)} \\
&= \frac{2^p c^p\alpha^p\sigma^p}{N^pM^p}\left\|H^\top b^{(A)}\right\|^p \\
&= \frac{2^p c^p\alpha^p\sigma^p}{N^pM^p}\left\|b^{(A)}\right\|^p && \left\|H^\top b^{(A)}\right\|^2 = \left\|b^{(A)}\right\|^2 \text{ by } HH^\top = I_N \\
&= \frac{2^p c^p\alpha^p\sigma^p}{N^pM^p}\left\|b^{(A)}\right\|_1^{p/2} && \left\|b^{(A)}\right\|^2 = \sum_{i=1}^{N}(b_i^{(A)})^2 = \sum_{i=1}^{N} b_i^{(A)} = \left\|b^{(A)}\right\|_1 \\
&= \frac{2^p c^p\alpha^p\sigma^p}{N^pM^p}\left(\frac{N}{2}\right)^{p/2} && \left\|b^{(A)}\right\|_1 = \frac{N}{2} \\
&= \frac{2^{p/2} c^p\alpha^p\sigma^p}{N^{p/2}M^p} \\
&\leq \frac{2^{p/2} c^p\sigma^p N^{p/2}}{n^p} && M \geq \frac{\alpha n}{N} \\
&\leq \frac{2^{p/2} c^p\sigma^p}{n^{p-1}} && N \leq n^{\frac{2}{p}} \\
&\leq 2^{p/2} c^p\sigma^p. && n^{p-1} \geq 1
\end{aligned}$$

Therefore, combining the above results gives

$$\mathbb{E}\left[\left\|\mathbf{X}^{(A)} - \boldsymbol{\mu}^{(A)}\right\|^p\right] \leq 2^{p-1}\left(\mathbb{E}\left[\left\|\mathbf{X}^{(A)}\right\|^p\right] + \left\|\boldsymbol{\mu}^{(A)}\right\|^p\right)$$

$$
\begin{aligned}
&\leq 2^{p-1}\left(c^p\alpha^p\sigma^p + 2^{p/2}c^p\sigma^p\right) \\
&\leq 2c^p\sigma^p\left(\alpha^p + 2^{p/2}\right) && 2^{p-1} \leq 2 \text{ for } p \in (1,2] \\
&= \sigma^p && c = \left(2(\alpha^p + 2^{p/2})\right)^{-1/p}.
\end{aligned}
$$

Hence $\mathbf{X}^{(A)} \in \mathcal{P}_\sigma$.

Now consider the associated quantum sampling oracle

$$
O_{\mathbf{X}^{(A)}} : |0\rangle|0\rangle \mapsto \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{1}{\sqrt{NM}}|\mathbf{X}^{(A)}(i,j)\rangle|i,j\rangle = \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{1}{\sqrt{NM}}|c\alpha\sigma(-1)^{1+A_{i,j}}H_i\rangle|i,j\rangle.
$$

Suppose there exists a quantum estimator that uses at most n queries to $O_{\mathbf{X}^{(A)}}$ and outputs an estimate $\tilde{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}^{(A)} := \mathbb{E}[\mathbf{X}^{(A)}]$. We will use this to solve the $\mathrm{Search}^N \circ \mathrm{Parity}^M$ problem using at most n queries to the phase oracle

$$
U_A : |i,j\rangle \mapsto (-1)^{A_{i,j}}|i,j\rangle.
$$

Indeed, one query to $O_{\mathbf{X}^{(A)}}$ can be simulated by preparing the uniform superposition over $[N]\times[M]$, and making one query to the binary oracle

$$
\mathcal{B}_{\mathbf{X}^{(A)}} : |i,j\rangle|0\rangle \mapsto |i,j\rangle|\mathbf{X}^{(A)}(i,j)\rangle = |i,j\rangle|c\alpha\sigma(-1)^{1+A_{i,j}}H_i\rangle,
$$

which can itself be simulated by one query to the controlled version of U_A via standard phase-kickback.

From the mean estimate $\tilde{\boldsymbol{\mu}}$, we define the estimate of vector $b^{(A)}$

$$
\tilde{b} := \frac{NM}{2c\alpha\sigma}H\tilde{\boldsymbol{\mu}}.
$$

Then we know that

$$
\begin{aligned}
\|\tilde{b} - b^{(A)}\| &= \frac{NM}{2c\alpha\sigma}\|H(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(A)})\| \\
&\leq \frac{NM}{2c\alpha\sigma}\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(A)}\|. && \|H\| = 1
\end{aligned}
$$

Since $NM \geq \alpha n$ by (7), the algorithm for the $\mathrm{Search}^N \circ \mathrm{Parity}^M$ problem uses at most $n \leq NM/\alpha$ queries to the phase oracle U_A. By Lemma C.11, with probability at least $2/3$, there exists an input $A \in \mathcal{A}_{N,M}$ such that

$$
\|\tilde{b} - b^{(A)}\| \geq \Omega(\sqrt{N}).
$$

Hence, for that hard instance A, we have

$$
\begin{aligned}
\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(A)}\| &\geq \frac{2c\alpha\sigma}{NM}\|\tilde{b} - b^{(A)}\| \\
&\geq \frac{2c\alpha\sigma}{NM}\Omega(\sqrt{N}) \\
&\geq \Omega\left(\frac{\sigma\sqrt{N}}{n}\right) && NM \leq 3\alpha n.
\end{aligned}
$$

Finally, since $N \leq n^{\frac{2}{p}} \leq 2N$, we have $N \geq n^{\frac{2}{p}}/2$, and therefore

$$
\begin{aligned}
\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(A)}\| &\geq \Omega\left(\frac{\sigma\, n^{\frac{1}{p}}}{n}\right) \\
&= \Omega\Big(\frac{\sigma}{n^{(p-1)/p}}\Big).
\end{aligned}
$$

□

Corollary C.13 (ϵ-error quantum lower bound in the high-dimensional regime)**.** *Let $\alpha > 1$ be the constant from Theorem C.12. Fix $p \in (1, 2]$ and $\sigma > 0$. Let $\mathcal{P}_\sigma$ be the set of all d-dimensional random variables $\mathbf{X}$ with mean $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ satisfying $\mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\mu}\|^p\right] \leq \sigma^p$. Suppose the dimension d satisfies*

$$\left(\frac{\sigma}{\epsilon}\right)^{\frac{2}{p-1}} \leq d,$$

and, when $p \in (1, 2)$, the error ϵ additionally satisfies

$$\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}} \leq \alpha^{\frac{p}{2-p}}.$$

Then any quantum estimator which, for every $\mathbf{X} \in \mathcal{P}_\sigma$, outputs $\widetilde{\boldsymbol{\mu}}$ satisfying $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \epsilon$ with probability at least $2/3$ must use at least

$$\Omega\left(\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}\right)$$

queries to the quantum sampling oracle $O_{\mathbf{X}}$.

Proof. Follows immediately from Theorem C.12. □

Corollary C.14 (Expected-query lower bound for Problem 3.2 in the high-dimensional regime)**.** *Let $\alpha > 1$ be the constant from Theorem C.12. Fix $p \in (1, 2]$ and $\sigma > 0$. Let $\mathcal{P}_\sigma$ be the set of all d-dimensional random variables $\mathbf{X}$ with mean $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ satisfying $\mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\mu}\|^p\right] \leq \sigma^p$. Suppose the dimension d satisfies*

$$\left(\frac{\sigma}{\epsilon}\right)^{\frac{2}{p-1}} \leq d,$$

and, when $p \in (1, 2)$, the error ϵ additionally satisfies

$$\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}} \leq \alpha^{\frac{p}{2-p}}.$$

Then any quantum algorithm which, for every $\mathbf{X} \in \mathcal{P}_\sigma$, outputs an estimator $\widetilde{\boldsymbol{\mu}}$ satisfying

$$\mathbb{E}[\widetilde{\boldsymbol{\mu}}] = \boldsymbol{\mu}, \qquad \mathbb{E}\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^p \leq \epsilon^p$$

must make at least

$$\Omega\left(\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}\right)$$

queries in expectation to the quantum sampling oracle $O_{\mathbf{X}}$.

Proof. Let $\mathcal{A}$ be a quantum algorithm satisfying the stated guarantee, and let $\mathbb{E}[Q]$ denote the worst-case expected number of queries made by $\mathcal{A}$. We define $T := \mathbb{E}[Q]$. By Markov's inequality, $\Pr(Q > 6T) \leq \frac{1}{6}$. Moreover, since it holds that $\mathbb{E}\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^p \leq \epsilon^p$, Markov's inequality also gives

$$\Pr\left(\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| > 6^{1/p}\epsilon\right) \leq \frac{1}{6}.$$

Therefore, by the union bound, we know that

$$\Pr\left(Q \leq 6T \text{ and } \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq 6^{1/p}\epsilon\right) \geq 1 - \frac{1}{6} - \frac{1}{6} = \frac{2}{3}.$$

Now define a truncated algorithm $\mathcal{A}'$ as follows. Run $\mathcal{A}$, but stop it once it has made more than $6T$ queries. If this happens, output an arbitrary vector. Then $\mathcal{A}'$ makes at most $6T$ queries in the worst case. By the preceding inequality, $\mathcal{A}'$ outputs an estimate satisfying

$$\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq 6^{1/p}\epsilon$$

with probability at least $2/3$.

We set this as a new error $\epsilon' := 6^{1/p}\epsilon$. For $p \in (1, 2]$, since it holds that

$$\left(\frac{\sigma}{\epsilon'}\right)^{\frac{2}{p-1}} \leq \left(\frac{\sigma}{\epsilon}\right)^{\frac{2}{p-1}} \leq d,$$

the high-dimensional condition also holds for ϵ'. Besides, for $p \in (1,2)$, since

$$\left(\frac{\sigma}{\epsilon'}\right)^{\frac{p}{p-1}} \le \left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}} \le \alpha^{\frac{p}{2-p}},$$

the low-precision condition also holds for ϵ'. Then we can apply Corollary C.13 with accuracy ϵ', which gives

$$6T \ge c_0 \left(\frac{\sigma}{\epsilon'}\right)^{\frac{p}{p-1}},$$

where c_0 is a constant. Substituting $\epsilon' = 6^{1/p}\epsilon$ yields $T \ge \frac{c_0}{6^{\frac{p}{p-1}}} \left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}$. Hence, the expected number of queries must satisfy

$$\mathbb{E}[Q] = \Omega\left(\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}\right).$$

□

C.6 Proof of Theorem 3.9

Proof. The first claim of Theorem 3.9 follows immediately from Theorem C.8 and Corollary C.13, while the second claim of Theorem 3.9 follows immediately from Corollary C.9 and Corollary C.14. □

D Proof in Section 4

D.1 Proof of Theorem 4.4

Proof. Lemma 4.3, summed over $t = 0, \dots, T-1$, gives

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\| \le \frac{3(f(\mathbf{x}_0) - f^*)}{\eta T} + \frac{8}{T}\sum_{t=0}^{T-1} \|\boldsymbol{\epsilon}_t\| + \frac{3L\eta}{2}. \tag{10}$$

Conditioned on $\mathbf{x}_t$, Theorem 3.4 ensures that $\|\boldsymbol{\epsilon}_t\| \le \epsilon$ with probability at least $1 - \delta/T$ for the stated choice of n. The union bound therefore gives an event $\mathcal{E}$ with $\Pr(\mathcal{E}) \ge 1 - \delta$ on which this bound holds for every t. On $\mathcal{E}$, the choices of η and T yield

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\| \overset{(10)}{\le} 3\epsilon + 8\epsilon + \frac{3\epsilon}{2} = \frac{25\epsilon}{2}. \tag{11}$$

To control the contribution of $\mathcal{E}^c$, note that smoothness and the lower bound on $f(\cdot)$ imply $\|\nabla f(\mathbf{x}_0)\| \le \sqrt{2L(f(\mathbf{x}_0) - f^*)}$. Since the normalized update satisfies $\|\mathbf{x}_{t+1} - \mathbf{x}_t\| \le \eta$ for every outcome, we have

$$\begin{aligned} \|\nabla f(\mathbf{x}_t)\| &\le \|\nabla f(\mathbf{x}_0)\| + L\|\mathbf{x}_t - \mathbf{x}_0\| \\ &\le \sqrt{2L(f(\mathbf{x}_0) - f^*)} + L\eta T \le \epsilon\big(\sqrt{2T} + T\big). \end{aligned} \tag{12}$$

Taking expectation, splitting over $\mathcal{E}$ and $\mathcal{E}^c$, and using $\delta = 1/(2T)$, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|\right] \overset{(11),(12)}{\le} \frac{25\epsilon}{2} + \delta\epsilon\big(\sqrt{2T} + T\big) = \left(13 + \frac{1}{\sqrt{2T}}\right)\epsilon \le 14\epsilon.$$

The uniform choice of the output gives

$$\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}})\|\right] = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|\right] \le 14\epsilon.$$

Each mean-estimation call has failure probability $\delta/T = 1/(2T^2)$, which contributes only logarithmic factors to the query complexity. The total number of queries to the quantum stochastic gradient oracle is bounded by

$$\tilde{\mathcal{O}}(Tn) = \tilde{\mathcal{O}}\big((1 + L(f(\mathbf{x}_0) - f^*)\epsilon^{-2})(1 + d^{\frac{p}{4(p-1)}}\sigma^{\frac{p}{2(p-1)}}\epsilon^{-\frac{p}{2(p-1)}})\big).$$

Replacing ϵ by $\epsilon/14$ gives an ϵ-stationary point with the stated query-complexity order. □

E Proof in Section 5

E.1 Proof of Lemma 5.2

Proof. Our proof follows Theorem 1 and Corollary 1 in Liu [35]. The optimality condition for the projection step gives

$$\langle \mathbf{x}_t - \eta \mathbf{g}_t - \mathbf{x}_{t+1}, \mathbf{x}^* - \mathbf{x}_{t+1} \rangle \le 0.$$

Rearranging this inequality yields

$$\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^* \rangle \le \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{2\eta} + \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle - \frac{\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2}{2\eta}. \tag{13}$$

By the definition of $\boldsymbol{\epsilon}_t$ and the gradient bound in Assumption 5.1, we have

$$\begin{aligned}
&\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \le G\|\mathbf{x}_t - \mathbf{x}_{t+1}\| + \|\boldsymbol{\epsilon}_t\|\|\mathbf{x}_t - \mathbf{x}_{t+1}\|, \\
&G\|\mathbf{x}_t - \mathbf{x}_{t+1}\| \le \eta G^2 + \frac{\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2}{4\eta}.
\end{aligned}$$

Young's inequality and $\|\mathbf{x}_t - \mathbf{x}_{t+1}\| \le D$ give

$$\begin{aligned}
\|\boldsymbol{\epsilon}_t\|\|\mathbf{x}_t - \mathbf{x}_{t+1}\| &\le C_p \eta^{p-1} D^{2-p} \|\boldsymbol{\epsilon}_t\|^p + \frac{\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^{p/(p-1)}}{4\eta D^{(2-p)/(p-1)}} \\
&\le C_p \eta^{p-1} D^{2-p} \|\boldsymbol{\epsilon}_t\|^p + \frac{\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2}{4\eta},
\end{aligned}$$

where $C_p = (4p-4)^{p-1}/p^p$. Substituting these bounds into (13) yields

$$\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^* \rangle \le \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{2\eta} + \eta G^2 + C_p \eta^{p-1} D^{2-p} \|\boldsymbol{\epsilon}_t\|^p. \tag{14}$$

Since $\mathbb{E}[\mathbf{g}_t \mid \mathbf{x}_t] = \nabla f(\mathbf{x}_t)$, taking expectation and using convexity gives

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] &\le \mathbb{E}[\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle] = \mathbb{E}[\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^* \rangle] \\
&\overset{(14)}{\le} \frac{\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\right]}{2\eta} + \eta G^2 + C_p \eta^{p-1} D^{2-p} \mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^p\right].
\end{aligned}$$

Summing over $t = 0, \ldots, T-1$ and using convexity once more, we obtain

$$\begin{aligned}
\mathbb{E}[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*)] &\le \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] \\
&\le \frac{D^2}{2\eta T} + \eta G^2 + C_p \eta^{p-1} D^{2-p} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^p\right],
\end{aligned}$$

where the last inequality uses $\|\mathbf{x}_0 - \mathbf{x}^*\| \le D$. □

E.2 Proof of Theorem 5.3

Proof. Conditioned on $\mathbf{x}_t$, Theorem 3.7 ensures that the estimate produced by `QUHTME` satisfies

$$\mathbb{E}[\mathbf{g}_t \mid \mathbf{x}_t] = \nabla f(\mathbf{x}_t) \qquad \text{and} \qquad \mathbb{E}[\|\boldsymbol{\epsilon}_t\|^p \mid \mathbf{x}_t] \le \tilde{\epsilon}^p.$$

Taking expectation gives $\mathbb{E}[\|\boldsymbol{\epsilon}_t\|^p] \le \tilde{\epsilon}^p$ for every t. With $\eta = \epsilon/(3G^2)$ and $T = \left\lceil 3G^2 D^2 \epsilon^{-2} \right\rceil$, we have

$$\frac{D^2}{2\eta T} + \eta G^2 \le \frac{\epsilon}{2} + \frac{\epsilon}{3} = \frac{5\epsilon}{6}.$$

The stated choice of $\tilde{\epsilon}$ gives

$$C_p \eta^{p-1} D^{2-p} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^p\right] \le C_p \eta^{p-1} D^{2-p} \tilde{\epsilon}^p = \frac{\epsilon}{6}.$$

Since $\hat{\mathbf{x}} = \bar{\mathbf{x}}_T$, Lemma 5.2 yields

$$\mathbb{E}\left[f(\hat{\mathbf{x}}) - f(\mathbf{x}^*)\right] \leq \frac{D^2}{2\eta T} + \eta G^2 + C_p \eta^{p-1} D^{2-p} \tilde{\epsilon}^p \leq \frac{5\epsilon}{6} + \frac{\epsilon}{6} = \epsilon.$$

At each iteration, Theorem 3.7 bounds the expected number of queries to the quantum stochastic gradient oracle by

$$\begin{aligned} n &= \tilde{\mathcal{O}}\big(\max\{d^{\frac{p}{4(p-1)}} \sigma^{\frac{p}{2(p-1)}} \tilde{\epsilon}^{-\frac{p}{2(p-1)}}, 1\}\big) \\ &= \tilde{\mathcal{O}}\big(d^{\frac{p}{4(p-1)}} G^{-1} \sigma^{\frac{p}{2(p-1)}} D^{\frac{2-p}{2(p-1)}} \epsilon^{-\frac{2-p}{2(p-1)}} + 1\big). \end{aligned}$$

Thus, the expected query complexity of Algorithm 4 is bounded by

$$T \cdot n = \tilde{\mathcal{O}}\big(d^{\frac{p}{4(p-1)}} G D^{\frac{3p-2}{2(p-1)}} \sigma^{\frac{p}{2(p-1)}} \epsilon^{-\frac{3p-2}{2(p-1)}} + G^2 D^2 \epsilon^{-2}\big).$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the scope and limitations of this work in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper only focuses on the theory of quantum complexities for multivariate heavy-tailed mean estimation and stochastic optimization with heavy-tailed noise.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.