# Quantum Speedups for Stochastic Optimization with Heavy-Tailed Noise

NeurIPS 2026

Bin Luo[1], Chengchang Liu[1,*], Jonathan Allcock[2], Shengyu Zhang[2], John C.S. Lui[1]

[1]The Chinese University of Hong Kong [2]Tencent Quantum Laboratory

## 01 Introduction

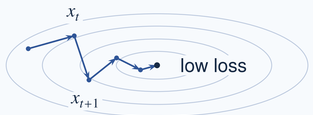
**Stochastic optimization problem:**

$$\min_{x\in\mathscr{X}\subseteq\mathbb{R}^d} f(x) = \mathbb{E}_\xi[F(x;\xi)].$$

- $f$ may be nonconvex
- $\xi$: random seed
- $F(x;\xi)$: random loss function

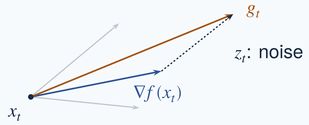
### Stochastic gradient descent

Loss contours (schematic)




### Noisy gradient

$g_t = \nabla f(x_t) + z_t$




### Fact 1: The noise $z_t$ can be a heavy-tailed random vector.

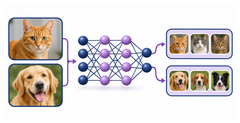
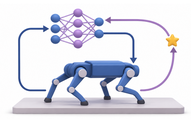


Image classification [1]



LLM training (BERT) [2]



Policy-gradient RL (PPO) [3]

### Definition 1: Heavy-tailed gradient noise

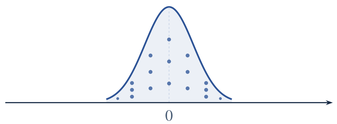
For every $x \in \mathscr{X}$, let $g = \nabla F(x;\xi)$. The gradient noise is **heavy-tailed** if

(i). $z = g - \nabla f(x)$; (ii). $\mathbb{E}_\xi[z] = 0$; (iii). $\mathbb{E}_\xi[\|z\|^p] \le \sigma^p$,

where $1 < p \le 2$, $\sigma > 0$.

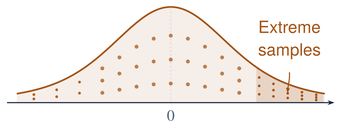
- $p = 2$: **light-tailed noise** with bounded variance.
- $p < 2$: gradient noise **need not** have finite variance.

**Light-tailed noise**




(i) Each coordinate $z_j$ follows Gaussian distribution [5].

**Heavy-tailed noise**




(ii) Each coordinate $z_j$ follows symmetric Lévy stable distribution [1].

Prior work [4] establishes **quantum speedups** for low-dimensional stochastic optimization under light-tailed gradient noise ($p = 2$).

### Open question

*Can we achieve quantum speedups for stochastic optimization under heavy-tailed gradient noise?*

## 02 Quantum Heavy-tailed Mean Estimator (QHTME)

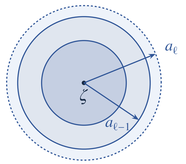
### Problem 1: Multivariate heavy-tailed mean estimation

Let $X$ have mean $\mu = \mathbb{E}[X] \in \mathbb{R}^d$ and $\mathbb{E}[\|X-\mu\|^p] \le \sigma^p$, where $1 < p \le 2$ and $\sigma > 0$. Given access to the **quantum sampling oracle**

$$O_X|0\rangle|0\rangle = \sum_{x\in\mathscr{E}} \sqrt{p_X(x)}\,|x\rangle|\psi_x\rangle, \qquad p_X(x) = \mathbb{P}[X = x],$$

output $\widetilde{\mu}$ with $\|\widetilde{\mu} - \mu\| \le \epsilon$, with probability at least $1 - \delta$.

### Algorithmic design




**Step 1. Rough mean estimation and recentering**

- Obtain a rough mean estimate $\zeta$ via a classical mean estimator
- Recenter $Y = X - \zeta$

**Step 2. Adaptive quantile shells**

- Obtain approximate $2^{-\ell}$-tail quantiles of $\|Y\|$: $0 = a_0 \le \cdots \le a_k$.

**Step 3. Estimate bounded pieces via quantum mean estimator and recombine**

- Obtain mean estimates $\widetilde{\mu}_\ell$ of pieces $Y_\ell = \frac{Y\,\mathbf{1}\{a_{\ell-1}<\|Y\|\le a_\ell\}}{a_\ell}$
- Compute final estimate $\widetilde{\mu} = \zeta + \sum_{\ell=1}^k a_\ell \widetilde{\mu}_\ell$

### Theorem 1: Quantum upper bound for Problem 1

QHTME solves Problem 1 using $\widetilde{O}\left(\left(\sqrt{d}\sigma/\epsilon\right)^{\frac{p}{2(p-1)}}\right)$ queries to $O_X$, with probability at least $1 - \delta$.

### Theorem 2: Quantum lower bound for Problem 1

Any quantum algorithm solving Problem 1 requires at least $\Omega\left((\sigma/\epsilon)^{\frac{p}{2(p-1)}}\right)$ queries to $O_X$.

## 03 Quantum speedup for nonconvex optimization

**Setup:** Access to a unitary quantum stochastic gradient oracle $O_{\nabla F,x}$ at query $x$:

$$O_{\nabla F,x}|0\rangle|0\rangle = \sum_{v\in\mathscr{E}} \sqrt{p_x(v)}\,|v\rangle|\psi_{x,v}\rangle, \qquad p_x(v) = \Pr_{\xi\sim P}[\nabla F(x;\xi) = v].$$

**Assumption:** $f$ is $L$-smooth and lower bounded; gradient noise satisfies Definition 1.

**Algorithm: Quantum Normalized SGD (QNSGD)**

**Input:** $x_0, \eta, n, p, d, \delta, T$.
**For** $t = 0, \dots, T-1$ **do**
  $g_t \leftarrow \text{QHTME}(\nabla F(x_t;\xi), n, p, d, \delta/T)$ → $\|g_t - \nabla f(x_t)\|$ is small
  $x_{t+1} \leftarrow x_t - \eta \frac{g_t}{\|g_t\|}$ → Normalized update gives sufficient descent
**EndFor**
**Output:** $\hat{x} \sim \text{Unif}\{x_0, \dots, x_{T-1}\}$.

**Table 1.** Quantum upper bounds and classical lower bounds for finding an $\epsilon$-stationary point in expectation, i.e., $\hat{x}$ with $\mathbb{E}[\|\nabla f(\hat{x})\|] \le \epsilon$.

| Setting | Quantum upper bound | Classical lower bound |
|---|---|---|
| Bounded variance ($p = 2$) | $\widetilde{O}(\sqrt{d}\,\epsilon^{-3})$ [4] | $\Omega(\epsilon^{-4})$ [2] |
| General case $p \in (1,2]$ | $\widetilde{O}\left(d^{\frac{p}{4(p-1)}}\epsilon^{-\frac{5p-4}{2(p-1)}}\right)$ **QNSGD** | $\Omega\left(\epsilon^{-\frac{3p-2}{p-1}}\right)$ [2] |

**Remark:**

1. QNSGD recovers the result in [4] when $p = 2$.
2. QNSGD breaks the classical lower bound when $d \le \epsilon^{-2+\kappa}$, for fixed $0 < \kappa < 2$.

## 04 Quantum Unbiased Heavy-tailed Mean Estimator (QUHTME)

### Problem 2: Unbiased multivariate heavy-tailed mean estimation

Let $X$ have mean $\mu = \mathbb{E}[X] \in \mathbb{R}^d$ and $\mathbb{E}[\|X-\mu\|^p] \le \sigma^p$, where $1 < p \le 2$. Given access to $O_X$ and the bound $\sigma$, output $\widetilde{\mu}$ satisfying

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

**Why unbiasedness?**

Our convex stochastic optimization analysis requires unbiased gradient estimates.

### Algorithmic design

| QHTME | → | QHTME$^+$ | → | QUHTME |
|---|---|---|---|---|
| High-probability estimator $\mathbb{P}[\|\widetilde{\mu}-\mu\| \le \epsilon] \ge 1-\delta$ | | moment control $\mathbb{E}[\|\widetilde{\mu}-\mu\|^p] \le \epsilon^p$ | | unbiased $\mathbb{E}[\widetilde{\mu}] = \mu, \mathbb{E}[\|\widetilde{\mu}-\mu\|^p] \le \epsilon^p$ |

**How to remove bias?**

**Add one randomized correction on a base estimate.**

$$\widetilde{\mu} = \underbrace{\widetilde{\mu}_0}_{\text{base estimate}} + \underbrace{2^J(\widetilde{\mu}_J - \widetilde{\mu}_{J-1})}_{\text{randomized correction}}.$$

- $J \sim \text{Geom}(1/2)$, $\mathbb{P}[J = j] = 2^{-j}, j = 1, 2, \dots$.
- Each level estimate $\widetilde{\mu}_j \leftarrow \text{QHTME}^+(X, \epsilon_j)$ satisfies $\mathbb{E}[\|\widetilde{\mu}_j - \mu\|^p] \le \epsilon_j^p$ and $\epsilon_j \downarrow 0$.

**The randomized correction removes the bias:**

$$\mathbb{E}[\widetilde{\mu}] = \mathbb{E}[\widetilde{\mu}_0] + \sum_{j\ge1} 2^{-j}2^j(\mathbb{E}[\widetilde{\mu}_j] - \mathbb{E}[\widetilde{\mu}_{j-1}]) = \lim_{j\to\infty} \mathbb{E}[\widetilde{\mu}_j] = \mu.$$

### Theorem 3: Quantum upper bound for Problem 2

QUHTME solves Problem 2 using $\widetilde{O}\left(\left(\sqrt{d}\sigma/\epsilon\right)^{\frac{p}{2(p-1)}}\right)$ queries to $O_X$ **in expectation**.

### Theorem 4: Quantum lower bound for Problem 2

Any quantum algorithm that solves Problem 2 requires $\Omega\left((\sigma/\epsilon)^{\frac{p}{2(p-1)}}\right)$ queries to $O_X$ **in worst-case expectation**.

## 05 Quantum speedup for convex optimization

**Assumption:** $f$ is convex on a nonempty, closed convex set $\mathscr{X}$ of diameter $D$; $\|\nabla f(x)\| \le G$; gradient noise satisfies Definition 1.

**Algorithm: Quantum Projected SGD (QPSGD)**

**Input:** $x_0 \in \mathscr{X}, \eta, \sigma, p, d, \epsilon_g, T$.
**For** $t = 0, \dots, T-1$ **do**
  $g_t \leftarrow \text{QUHTME}(\nabla F(x_t;\xi), \sigma, p, d, \epsilon_g)$ → Obtain unbiased gradient estimates
  $x_{t+1} \leftarrow \Pi_{\mathscr{X}}(x_t - \eta g_t)$ → Projected update gives sufficient descent
**EndFor**
**Output:** $\bar{x} \leftarrow \frac{1}{T}\sum_{t=0}^{T-1} x_t$.

$\Pi_{\mathscr{X}}(y) = \arg\min_{u\in\mathscr{X}}\|u - y\|^2$: Euclidean projection onto $\mathscr{X}$.

**Table 2.** Quantum upper bound (expected queries) and classical lower bound for finding an $\epsilon$-suboptimal point in expectation, i.e., $\hat{x}$ with $\mathbb{E}[f(\hat{x}) - f(x^*)] \le \epsilon$; fixed $G, D, \sigma$.

| Setting | Quantum upper bound | Classical lower bound |
|---|---|---|
| Heavy-tailed case $p \in (1,2]$ | $\widetilde{O}\left(d^{\frac{p}{4(p-1)}}\epsilon^{-\frac{3p-2}{2(p-1)}} + \epsilon^{-2}\right)$ **QPSGD** | $\Omega\left(\epsilon^{-\frac{p}{p-1}}\right)$ [6] |

**Remark:** QPSGD breaks the classical lower bound when $d \le \epsilon^{-\frac{2(2-p)}{p}+\kappa}$, where $1 < p < 2$ and fixed $0 < \kappa < \frac{2(2-p)}{p}$.

**Contact**

**Bin Luo** binluo@link.cuhk.edu.hk
The Chinese University of Hong Kong
*Corresponding author: Chengchang Liu
7liuchengchang@gmail.com

**Notation**

$d$: dimension $p \in (1,2]$: moment index
$\sigma$: central-moment scale $\epsilon$: accuracy
$\delta$: failure probability
$\widetilde{O}$: polylogarithmic factors suppressed

**Selected references**

[1] Simsekli et al. *A tail-index analysis of stochastic gradient noise in deep neural networks.* ICML 2019.
[2] Zhang et al. *Why are adaptive methods good for attention models?* NeurIPS 2020.
[3] Garg et al. *On proximal policy optimization's heavy-tailed gradients.* ICML 2021.
[4] Sidford & Zhang. *Quantum speedups for stochastic optimization.* NeurIPS 2023.
[5] Panigrahi et al. *Non-Gaussianity of Stochastic Gradient Noise.* arXiv:1910.09626, 2019.
[6] Raginsky & Rakhlin. *Information complexity of black-box convex optimization: A new look via feedback information theory.* Allerton 2009.