QUANTUM HEAVY-TAILED BANDITS: LOWER BOUNDS AND IMPROVED DIMENSION DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study stochastic multi-armed and linear bandits with heavy-tailed rewards under quantum access. The learner chooses one action at a time and accesses a quantum procedure representing that action's reward distribution, which it can run forwards and backwards to estimate the expected reward. Assume that the reward Y_a associated with each action a satisfies $\mathbb{E}|Y_a|^{1+\alpha} \le 1$, where $\alpha \in (0, 1]$. For K arms and a budget of T oracle queries, with T at least a constant multiple of K, we achieve regret $\widetilde{O}\big(K^{2\alpha/(1+\alpha)}T^{(1-\alpha)/(1+\alpha)}\big)$ and prove a lower bound matching up to logarithmic factors. This improves the linear dependence on K in Wu et al. (2023) whenever $\alpha < 1$. For d-dimensional linear bandits with $T \ge d$, we obtain the same exponent in T, with dimension dependence $d^{1/2+\alpha/(1+\alpha)}$ for finite action sets and $d^{1/2+2\alpha/(1+\alpha)}$ for general compact action sets. The finite-action bound depends only logarithmically on the number of actions; the compact-action result assumes access to routines for experimental design and linear optimization. These results improve the quadratic dimension dependence of Wu et al. (2023) over their established range $\alpha \in [1/3, 1]$ and extend regret guarantees to the full range $\alpha \in (0, 1]$. The improvements come from jointly allocating estimation queries across actions and reconstructing linear rewards from a small set of representative actions. We also establish worst-case lower bounds for linear bandits and extend the analysis to bounded central moments. Numerical simulations illustrate the theoretical trade-offs.

1 INTRODUCTION

Stochastic bandits provide a basic framework for learning to make decisions from uncertain rewards. In many applications, however, rare observations can be much larger than typical rewards, making bounded or sub-Gaussian assumptions restrictive. Classical heavy-tailed bandit algorithms address this difficulty through robust estimation under bounded low-order moments (Bubeck et al., 2013). For linear bandits, the learner must additionally exploit shared structure across actions, and recent work shows that the dimension and the size of the action set play distinct roles in determining regret (Tajdini et al., 2025). Understanding these dependencies is essential for identifying the statistical cost of learning from heavy-tailed rewards.

Quantum access offers a different way to estimate expected rewards. In the quantum reward-oracle model, the learner selects an action classically and accesses its reward distribution through a quantum procedure that can be applied forwards and backwards. Each oracle invocation counts as one bandit round and incurs the regret associated with that action. Quantum mean estimation can therefore reduce the number of costly interactions needed to distinguish competing actions. This approach yields logarithmic regret upper bounds for bounded-reward bandits (Wan et al., 2023). Under a bounded raw moment of order $1 + \alpha$, Wu et al. (2023) established quantum heavy-tailed regret upper bounds for multi-armed bandits and, when $\alpha \in [1/3, 1]$, linear bandits. Their bounds improve the horizon dependence over classical guarantees but retain linear dependence on the number of arms and quadratic dependence on the linear dimension.

These results leave two questions unresolved: whether the arm dependence is minimax optimal, and how effectively quantum estimation can exploit linear structure. Answering them requires more than a scalar estimation guarantee. For upper bounds, estimation costs must be allocated across actions under a shared query budget, with each query charged according to its action's suboptimality.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Classical and quantum heavy-tailed regret bounds under $\mathbb{E}|Y_a|^{1+\alpha} \le u$. Every leading moment-dependent term should be understood as multiplied by $u^{1/(1+\alpha)}$; additive logarithmic terms are displayed explicitly. The bounds of Tajdini et al. (2025) are stated under the central-moment condition $\mathbb{E}|Y_a - \mu(a)|^{1+\alpha} \le v$. Their upper bounds apply to our setting with $v = 2^{1+\alpha}u$ by equation 40, and their lower bounds are reported as stated in the original work.

Setting		Classical	Quantum
MAB	UB	$\widetilde{O}\left(K^{\frac{\alpha}{1+\alpha}} T^{\frac{1}{1+\alpha}}\right)$ Bubeck et al. (2013); Lee et al. (2020)	*Prior:* $\widetilde{O}\left(KT^{\frac{1-\alpha}{1+\alpha}}\right)$ Wu et al. (2023) *Ours:* $\widetilde{O}\left(K^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}} + K\log T\right)$
	LB	$\widetilde{\Omega}\left(K^{\frac{\alpha}{1+\alpha}} T^{\frac{1}{1+\alpha}}\right)$ Bubeck et al. (2013); Lee et al. (2020)	*Ours:* $\Omega\left(K^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}} + K\log(T/K)\right)$
Finite-action linear bandits $\lvert\mathcal{A}\rvert = n$	UB	$\widetilde{O}\left(\sqrt{d}\,(\log n)^{\frac{\alpha}{1+\alpha}} T^{\frac{1}{1+\alpha}}\right)$ Tajdini et al. (2025)	*Ours:* $\widetilde{O}\left(d^{\frac{1+3\alpha}{2(1+\alpha)}} (\log(nT))^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}} + d\log T\log(nT)\right)$
	LB	$\widetilde{\Omega}\left(d^{\frac{\alpha}{1+\alpha}} (\log n)^{\frac{\alpha}{1+\alpha}} T^{\frac{1}{1+\alpha}}\right)$ Tajdini et al. (2025)	*Ours:* $\Omega\left(d^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}} + d\log(T/d)\right)$
General-action linear bandits	UB	$\widetilde{O}\left(d^{\frac{1+3\alpha}{2(1+\alpha)}} T^{\frac{1}{1+\alpha}}\right)$ Tajdini et al. (2025)	*Prior:* $\widetilde{O}\left(d^2 T^{\frac{1-\alpha}{1+\alpha}}\right)$ Wu et al. (2023) *Ours:* $\widetilde{O}\left(d^{\frac{1+5\alpha}{2(1+\alpha)}} T^{\frac{1-\alpha}{1+\alpha}} + d\log T\right)$
	LB	$\Omega\left(d^{\frac{2\alpha}{1+\alpha}} T^{\frac{1}{1+\alpha}}\right)$ Tajdini et al. (2025)	*Ours:* $\Omega\left(d^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}} + d\log(T/d)\right)$

In linear bandits, estimates obtained from a small collection of actions must also provide reliable predictions over the entire active set. For lower bounds, classical sample-based testing arguments must be replaced by arguments that accommodate coherent queries, intermediate measurements, and data-dependent stopping.

We address these questions through regret lower bounds and improved analyses of design-based phased elimination. Throughout the main paper, rewards satisfy the bounded raw-moment condition $\mathbb{E}|Y_a|^{1+\alpha} \le u$ for $\alpha \in (0, 1]$. For clarity, the following overview normalizes $u = 1$ and reports the leading regret terms; our theorems retain the moment-scale dependence and additive logarithmic terms.

- **Minimax guarantees for quantum heavy-tailed multi-armed bandits.** For K actions, we obtain regret $\widetilde{O}\left(K^{2\alpha/(1+\alpha)} T^{(1-\alpha)/(1+\alpha)}\right)$ and a matching lower bound up to logarithmic factors. This characterizes the optimal arm and horizon dependence and improves the previous linear dependence on K whenever $\alpha < 1$. The upper-bound improvement follows from retaining the shared budget constraint across active actions in the regret analysis.
- **Improved dimension dependence for quantum heavy-tailed linear bandits.** For a finite action set of size n, we establish the leading bound $\widetilde{O}\left(d^{1/2+\alpha/(1+\alpha)} (\log(nT))^{2\alpha/(1+\alpha)} T^{(1-\alpha)/(1+\alpha)}\right)$. For general compact action sets, we obtain $\widetilde{O}\left(d^{1/2+2\alpha/(1+\alpha)} T^{(1-\alpha)/(1+\alpha)}\right)$, independently of cardinality and covering numbers, assuming access to experimental-design and linear-optimization oracles. These results improve the prior quadratic dimension dependence over its established parameter range $\alpha \in [1/3, 1]$ and extend guarantees to every $\alpha \in (0, 1]$. The algorithms estimate a small design support and reconstruct active-action means, using distinct aggregation arguments for finite and general action sets.
- **Lower bounds under adaptive quantum access.** We combine rare-event reward constructions with a stopped-query hybrid argument that lower-bounds the expected number of queries needed for reliable testing. A bandit-to-testing reduction yields the matching multi-armed lower bound and the linear-bandit lower bound $\Omega\left(d^{2\alpha/(1+\alpha)} T^{(1-\alpha)/(1+\alpha)}\right)$. The linear constructions cover both finite action sets and compact infinite action sets.

Table 1 summarizes the most closely related regret bounds for classical and quantum heavy-tailed bandits, highlighting our improvements in arm and dimension dependence. We defer a detailed discussion of related work on classical heavy-tailed bandits, quantum bandits, and the underlying mean-estimation techniques to Appendix A. All proofs are deferred to Appendix B and C. We further extend the analysis to bounded central moments in Appendix D. More details of numerical experiments are in Appendix E.

2 PROBLEM SETTING

In this section, we first introduce the heavy-tailed multi-armed and linear bandit models, and then formalize the quantum reward-oracle access model.

2.1 HEAVY-TAILED BANDITS

Fix a horizon $T \geq 1$, a tail parameter $\alpha \in (0, 1]$, and a moment bound $u > 0$, all known to the learner. Every action a is associated with a reward distribution ν_a on $\mathbb{R}$. We write $Y_a \sim \nu_a, \quad \mu(a) := \mathbb{E}_{\nu_a}[Y_a]$, where Y_a denotes a generic random reward associated with action a. Throughout the main paper, we impose the bounded raw-moment condition

$$\mathbb{E}_{\nu_a}\left[|Y_a|^{1+\alpha}\right] \leq u \qquad \text{for every action } a. \tag{1}$$

For a generic random variable $Y \sim \nu$, define the corresponding raw-moment class by

$$\mathcal{H}_\alpha^{\mathrm{raw}}(u) := \left\{\nu : |\mathbb{E}_\nu[Y]| \leq 1,\ \mathbb{E}_\nu[|Y|^{1+\alpha}] \leq u\right\}. \tag{2}$$

Thus, throughout the main paper, $\nu_a \in \mathcal{H}_\alpha^{\mathrm{raw}}(u)$ for every action a. The restriction $|\mathbb{E}_\nu[Y]| \leq 1$ normalizes the range of expected rewards, whereas the second condition controls the heaviness of the reward tails. Uppercase symbols such as Y_a and Y_t denote random variables, while lowercase symbols such as y and y_t denote their realized values. The bounded central-moment model and its different estimation interface are deferred to Appendix D.

Heavy-tailed multi-armed bandits. The action set is $\mathcal{A} = [K]$. At each round $t \in [T]$, the learner selects an action $a_t \in \mathcal{A}$ using the information available before round t.

Under the classical sampling model, the environment then generates a fresh random reward Y_t according to $Y_t \mid (\mathcal{F}_{t-1}, a_t = a) \sim \nu_a, \mathbb{E}[Y_t \mid \mathcal{F}_{t-1}, a_t = a] = \mu(a)$, where $\mathcal{F}_{t-1}$ denotes the interaction history before round t. The reward Y_t is conditionally independent of the past given a_t, and the learner observes its realized value $y_t \in \mathbb{R}$.

The quantum model studied in this paper replaces this classical sampling step with coherent access to the action-specific quantum reward oracle defined below. The learner need not measure a classical reward after each query; however, measuring the oracle's reward register produces a realization $y \sim \nu_a$. Each invocation of the oracle associated with action a, or of its inverse, counts as one bandit round labeled by a.

Let $a^* \in \arg\max_{a \in \mathcal{A}} \mu(a), \mu^* := \mu(a^*), \Delta(a) := \mu^* - \mu(a)$. Since $|\mu(a)| \leq 1$, every suboptimal gap satisfies $\Delta(a) \in [0, 2]$. We evaluate a policy by its expected pseudo-regret

$$\mathsf{Reg}_T := \mathbb{E}\left[\sum_{t=1}^{T} \Delta(a_t)\right], \tag{3}$$

where the expectation is taken over all classical and quantum randomness generated by the policy and its interaction with the reward oracles.

Heavy-tailed linear bandits. Let $\mathcal{A} \subseteq \mathbb{R}^d$ be a known finite or compact action set satisfying $\|a\|_2 \leq 1$ for every $a \in \mathcal{A}$. We assume without loss of generality that $\mathrm{span}(\mathcal{A}) = \mathbb{R}^d$; otherwise, the problem can be restricted to $\mathrm{span}(\mathcal{A})$ and d replaced by its dimension. When $\mathcal{A}$ is finite, we write $n := |\mathcal{A}|$.

An unknown true parameter $\theta^* \in \mathbb{R}^d$ with $\|\theta^*\|_2 \leq 1$ determines the expected reward of each action through the linear relation

$$\mu(a) := \mathbb{E}_{\nu_a}[Y_a] = a^\top \theta^*, \qquad a \in \mathcal{A}, \tag{4}$$

where each action-specific reward distribution satisfies $\nu_a \in \mathcal{H}_\alpha^{\mathrm{raw}}(u)$. Apart from the linearity of their means, the reward distributions may be heterogeneous across actions; in particular, we do not assume an action-independent additive-noise model.

At each round $t \in [T]$, the learner selects an action $a_t \in \mathcal{A}$ using the information available before round t and coherently accesses the corresponding reward distribution through the quantum reward oracle defined below. Let $a^* \in \arg\max_{a \in \mathcal{A}} a^\top \theta^*, \mu^* := \mu(a^*) = (a^*)^\top \theta^*$. The suboptimal gap of

selecting action a is $\Delta(a) := \mu^* - \mu(a) = (a^* - a)^\top \theta^*$. Consequently, the expected pseudo-regret is

$$\mathsf{Reg}_T := \mathbb{E}\left[\sum_{t=1}^{T} \Delta(a_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} (a^* - a_t)^\top \theta^*\right]. \tag{5}$$

This is the same performance criterion as in equation 3, specialized to linearly parameterized mean rewards.

2.2 Quantum reward-oracle model

For each action $a \in \mathcal{A}$, let $\Omega_a \subset \mathbb{R}$ denote the finite support of its encoded reward distribution. Following Wu et al. (2023) and Luo et al. (2026), we formulate quantum access using finite-valued random variables. This is the standard finite-precision quantum representation of a real-valued reward; for an unbounded distribution, it can be interpreted as a sufficiently accurate truncated and discretized representation.

For every $y \in \Omega_a$, define $p_a(y) \coloneqq \Pr_{\nu_a}(Y_a = y)$. Quantum access to ν_a is provided through a unitary reward oracle $\mathcal{O}_a$ satisfying

$$\mathcal{O}_a|0\rangle|0\rangle = \sum_{y \in \Omega_a} \sqrt{p_a(y)}\,|y\rangle|\psi_{a,y}\rangle, \tag{6}$$

where the two occurrences of $|0\rangle$ are fixed reference states, $\{|y\rangle : y \in \Omega_a\}$ is an orthonormal computational basis for the reward register, and each auxiliary state is normalized: $\langle \psi_{a,y} \mid \psi_{a,y} \rangle = 1$. The auxiliary states need not be mutually orthogonal. Measuring the reward register of the state in equation 6 returns a realized reward value y with probability $p_a(y)$ and therefore produces an ordinary sample from ν_a.

Equation 6 specifies the action of $\mathcal{O}_a$ only on its reference input. A fixed unitary extension to the orthogonal complement is taken as part of the oracle instance. The learner is given black-box access to both $\mathcal{O}_a$ and its inverse $\mathcal{O}_a^\dagger$. At each query round, the learner selects an action a classically and applies either $\mathcal{O}_a$ or $\mathcal{O}_a^\dagger$.

3 Quantum Heavy-tailed Multi-Armed Bandits

We first derive a minimax regret lower bound via rare-event quantum testing, and then show that a phased-elimination policy matches its leading dependence on K and T up to logarithmic factors.

3.1 Minimax lower bound

We first establish a minimax regret lower bound for quantum multi-armed bandits (QMAB) under the bounded raw-moment condition.

Theorem 3.1 (Minimax lower bound for raw-moment QMAB)**.** *There exist universal constants $c, C > 0$ such that, for every $\alpha \in (0, 1]$, $u > 0$, $K \geq 2$, and $T \geq CK$,*

$$\inf_{\mathsf{Alg}} \sup_{\boldsymbol{\nu} \in (\mathcal{H}_\alpha^{\mathrm{raw}}(u))^K} \mathsf{Reg}_T(\mathsf{Alg}, \boldsymbol{\nu}) \quad \geq c\left[T \min\left\{1,\, u^{\frac{1}{1+\alpha}} \left(\frac{K}{T}\right)^{\frac{2\alpha}{1+\alpha}}\right\} + \min\left\{1, u^{\frac{1}{1+\alpha}}\right\} K \log \frac{T}{K}\right]. \tag{7}$$

Here the supremum ranges over valid arm-wise quantum reward-oracle instances whose reward distributions belong to $\mathcal{H}_\alpha^{\mathrm{raw}}(u)$. Consequently, in the non-saturated regime, the polynomial heavy-tail term is $\Omega\left(u^{\frac{1}{1+\alpha}} K^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}}\right)$. If u is bounded below by a positive universal constant, the second term in the above result implies $\Omega(K \log(T/K))$.

Proof sketch. The proof combines a rare-event quantum testing lower bound with a bandit-to-testing reduction.

We first consider the two reward distributions $P_- = (1 - p/2)\delta_0 + (p/2)\delta_B$, and $P_+ = (1 - p)\delta_0 + p\delta_B$ where $B = (u/(2p))^{1/(1+\alpha)}$. Both distributions belong to $\mathcal{H}_\alpha^{\mathrm{raw}}(u)$, and their means

are separated by $\Delta = \Theta\big(u^{1/(1+\alpha)}p^{\alpha/(1+\alpha)}\big)$. For the canonical unitary extensions of their reward oracles, the corresponding rotation angles differ by $\Theta(\sqrt{p})$, and hence their operator-norm distance is $\Theta(\sqrt{p})$. Applying the stopped-query hybrid bound in Lemma B.1 shows that every adaptive constant-error test requires $\Omega(p^{-1/2})$ expected oracle queries under P_-. Rewriting p in terms of the mean separation Δ gives precisely the quantum mean-testing lower bound in Theorem B.2, namely $\Omega\big(u^{1/(2\alpha)}\Delta^{-(1+\alpha)/(2\alpha)}\big)$ expected queries.

We then embed this testing problem into a K-armed bandit family. One reference action has deterministic reward $3pB/4$, while the other $K-1$ candidate actions follow P_- under the *null* instance. Under *alternative* a, only candidate action a is changed from P_- to P_+. The reference action is optimal under the null, action a is optimal under alternative a, and the smallest relevant gap is $g = \Delta/2$ as recorded in equation 13.

The one-sided reduction in Lemma B.3 formalizes the regret-to-testing argument. If a policy has worst-case regret $M \le cTg$, then thresholding the number $N_a(T)$ of queries assigned to candidate action a at $T/2$ distinguishes the null instance from alternative a with constant error. Combining this test with Theorem B.2 forces $\mathbb{E}_0 N_a(T) = \Omega\big(u^{1/(2\alpha)}g^{-(1+\alpha)/(2\alpha)}\big)$ for every candidate action. Summing these query requirements under the null instance gives the two-branch regret lower bound of $\Omega\Big(\min\Big\{Tg,\ Ku^{1/(2\alpha)}g^{1-\frac{1+\alpha}{2\alpha}}\Big\}\Big)$ in equation 14. If instead $M > cTg$, its first branch already holds.

Balancing the two branches of equation 14 gives $g \asymp u^{1/(1+\alpha)}(K/T)^{2\alpha/(1+\alpha)}$, as stated in equation 15. Substitution yields the polynomial term in equation 16. When this choice of g exceeds the constant range permitted by the construction, fixing g at a sufficiently small constant gives the saturated lower bound $\Omega(T)$.

The logarithmic term follows from a separate hard family rather than from the preceding rare-event construction. Specifically, scaling the bounded-reward point-versus-interval construction of Liu et al. (2026) by $\min\{1, u^{1/(1+\alpha)}\}$ gives $\Omega\big(\min\{1, u^{1/(1+\alpha)}\}K\log(T/K)\big)$ regret while preserving the raw-moment constraint. Since the maximum of two valid lower bounds is at least half their sum, combining the polynomial and logarithmic constructions proves equation 16. □

3.2 Algorithms and Improved Regret Bounds

Algorithm 1 Quantum heavy-tailed successive elimination for MAB

Input: horizon T, confidence δ, actions $[K]$, tail index α, and raw-moment scale u.

1. Set $\mathcal{V}_1 \leftarrow [K]$, $N \leftarrow 0$, and $\epsilon_\ell \leftarrow 2^{-\ell}$ for $\ell \ge 1$.
2. For phase $\ell = 1, 2, \ldots$, set $\delta_\ell \leftarrow \delta/[2K(\ell+1)^2]$.
3. Compute $q_\ell \leftarrow Q_{\mathrm{raw}}(\epsilon_\ell, \delta_\ell; u)$ from Lemma B.4.
4. If $N + |\mathcal{V}_\ell| q_\ell > T$, stop the phased procedure.
5. For every $a \in \mathcal{V}_\ell$, run the corresponding quantum mean estimator with deterministic cap q_ℓ and obtain $\widehat{\mu}_\ell(a)$; update $N \leftarrow N + |\mathcal{V}_\ell| q_\ell$.
6. Put $\widehat{\mu}_\ell^{\max} \leftarrow \max_{a\in\mathcal{V}_\ell} \widehat{\mu}_\ell(a)$ and retain
$$\mathcal{V}_{\ell+1} \leftarrow \{a \in \mathcal{V}_\ell : \widehat{\mu}_\ell(a) \ge \widehat{\mu}_\ell^{\max} - 2\epsilon_\ell\}.$$

If at least one phase is completed, choose an action maximizing the last available estimate over the current active set; otherwise choose any action in $[K]$. Query that action for every remaining round.

Our upper bound combines the raw-moment quantum truncated mean estimator (QTME) of Wu et al. (2023) with a standard phased-elimination scheme. We defer the estimator statement, and complete proof to Appendix B.2.

Theorem 3.2 (Improved regret for raw-moment QMAB)**.** *For every $\alpha \in (0, 1]$, $u > 0$, $K \ge 2$, and $T \ge K$, Algorithm 1, with failure probability $\delta = T^{-2}$, satisfies*

$$\mathsf{Reg}_T = \widetilde{O}\Big(u^{\frac{1}{1+\alpha}}K^{\frac{2\alpha}{1+\alpha}}T^{\frac{1-\alpha}{1+\alpha}} + K\log T\Big)$$

under the raw-moment condition $\mathbb{E}|Y_a|^{1+\alpha} \le u$ for every action $a \in [K]$. Together with Theorem 3.1, the polynomial term is minimax optimal in its dependence on u, K, and T.

Remark 3.3 (Source of the improvement over Wu et al. (2023).)**.** *Wu et al. (2023) proved the regret bound $\widetilde{O}(u^{1/(1+\alpha)}KT^{(1-\alpha)/(1+\alpha)})$ using the same raw-moment quantum mean estimator, so the improvement does not stem from a sharper estimator. Instead, it comes from retaining the global coupling among the query budgets assigned to different actions. In the analysis of Wu et al. (2023), the contribution of each action is bounded separately using its individual worst-case budget, after which the resulting bounds are summed over all K actions. This produces the linear factor K. Our phase-level analysis instead uses the total-budget constraint across all active actions before bounding regret. Equivalently, if M_a denotes the final doubling level associated with action a, the global constraint $\sum_a 2^{M_a} \lesssim T$ and concavity imply $\sum_a (2^{M_a})^{(1-\alpha)/(1+\alpha)} \le K^{2\alpha/(1+\alpha)}(\sum_a 2^{M_a})^{(1-\alpha)/(1+\alpha)}$. This yields the factor $K^{2\alpha/(1+\alpha)}T^{(1-\alpha)/(1+\alpha)}$ directly. Consequently, for every $\alpha \in (0,1)$, the polynomial term improves the previous bound by a factor $K^{(1-\alpha)/(1+\alpha)}$, without changing its dependence on u or its exponent in T. At $\alpha = 1$, this factor equals one, and both bounds reduce to the same near-linear dependence on K, up to logarithmic factors.*

4 Quantum Heavy-tailed Linear Bandits

We next study quantum linear bandits with bounded raw $(1+\alpha)$-moments. We first establish lower bounds and then present improved regret guarantees for finite and general action sets.

4.1 Lower bounds

The rare-event testing construction developed for multi-armed bandits also yields hard linear-bandit instances. Importantly, the reduction requires only d linearly independent actions and therefore does not obtain its dimension dependence from a large action-set cardinality.

Theorem 4.1 (Quantum linear-bandit lower bounds)**.** *There exist universal constants $c, C > 0$ such that, for every $d \ge 2$ and $T \ge Cd$, there exists a finite action set $\mathcal{A}_{\text{fin}} = \{a_1, \dots, a_d\} \subset \mathbb{R}^d$ for which every quantum linear-bandit policy satisfies*

$$\mathsf{Reg}_T \ge c\left[T\min\left\{1,\, u^{\frac{1}{1+\alpha}}\left(\frac{d}{T}\right)^{\frac{2\alpha}{1+\alpha}}\right\} + \min\left\{1, u^{\frac{1}{1+\alpha}}\right\} d\log\left(\frac{T}{d}\right)\right].$$

The same bound holds for the specific compact infinite action set $\mathcal{A}_{\text{ray}} := \bigcup_{i=1}^{d}\{ta_i : t \in [0,1]\}$.

Proof sketch. Let P_- and P_+ be the rare-event distributions defined in equation 9. Their means differ by $\Delta = \Theta\big(u^{1/(1+\alpha)}p^{\alpha/(1+\alpha)}\big)$, while their canonical reward oracles have operator-norm separation $\Theta(\sqrt{p})$. The stopped-query hybrid argument in Lemma B.1 therefore implies the constant-confidence mean-testing lower bound of Theorem B.2: distinguishing the two distributions requires $\Omega\big(u^{1/(2\alpha)}\Delta^{-(1+\alpha)/(2\alpha)}\big)$ expected queries. To embed the resulting bandit family, let $a_1 = e_1$ and $a_i = \frac{2}{3}e_1 + \frac{\sqrt{5}}{3}e_i$ for $i = 2, \dots, d$. Under the null parameter, a_1 is the unique optimal action and every other action has distribution P_-. Under alternative i, only the distribution of a_i changes to P_+. These mean vectors are realized by the linear parameters $\theta^{(0)} = \frac{3pB}{4}e_1$ and $\theta^{(i)} = \theta^{(0)} + \frac{3pB}{2\sqrt{5}}e_i$. Consequently, Lemma B.3 applies with $K = d$ and gives

$$\mathsf{Reg}_T = \Omega\Big(\min\Big\{Tg,\, d\,u^{\frac{1}{2\alpha}}g^{1-\frac{1+\alpha}{2\alpha}}\Big\}\Big),$$

where g is the gap between the reference action and each candidate. Balancing the two terms gives $g \asymp u^{1/(1+\alpha)}(d/T)^{2\alpha/(1+\alpha)}$ and proves the polynomial term. The bounded-reward construction of Liu et al. (2026), which belongs to our raw-moment class after constant rescaling, supplies the additive $\Omega(d\log(T/d))$ term.

For the infinite-action statement, we radially extend the construction to $\mathcal{A}_{\text{ray}} = \bigcup_{i=1}^{d}\{ta_i : t \in [0,1]\}$. A query at ta_i has oracle separation at most $O(\sqrt{tp}) \le O(\sqrt{p})$, so selecting an interior point of a ray cannot reveal more information than selecting its endpoint. The same stopped-query and bandit-to-testing arguments therefore apply after counting all queries made on each ray. Complete constructions and moment checks are deferred to Appendix C.1. □

Algorithm 2 Finite-action heavy-tailed quantum linear elimination

Input: finite $\mathcal{A} \subseteq \mathbb{R}^d$, $n = |\mathcal{A}|$, horizon T, confidence δ, α, and raw-moment scale u.

1. Set $\mathcal{V}_1 \leftarrow \mathcal{A}$, $N \leftarrow 0$, and $L_T \leftarrow \lceil \log_2 T \rceil$.
2. For $\ell = 1, \dots, L_T$, set $\epsilon_\ell = 2^{-\ell}$ and $H_\ell \leftarrow \lceil C \log(2n(\ell+1)^2/\delta) \rceil$.
3. Compute $(S_\ell, \pi_\ell, M_\ell)$ from Lemma C.4; form equation 28.
4. Let C_ℓ be H_ℓ times the deterministic timeout cap of one replicate. If $N + C_\ell > T$, stop the phased procedure.
5. Run H_ℓ independent, timeout-protected replicates. For every $a \in \mathcal{V}_\ell$, let $\widehat{\mu}_\ell(a)$ be the median of its reconstructed predictions. Update $N \leftarrow N + C_\ell$.
6. Put $\widehat{\mu}_\ell^{\max} = \max_{a \in \mathcal{V}_\ell} \widehat{\mu}_\ell(a)$ and set
$$\mathcal{V}_{\ell+1} = \{a \in \mathcal{V}_\ell : \widehat{\mu}_\ell(a) \geq \widehat{\mu}_\ell^{\max} - 2\epsilon_\ell\}.$$

Use an action maximizing the last completed-phase estimate for all remaining rounds; if no phase completed, use any action in $\mathcal{A}$.

4.2 Algorithms and improved regret bounds

This subsection develops algorithms and upper bounds for quantum heavy-tailed linear bandits with finite and general action sets.

Our algorithms follow a common design-based elimination template. In phase ℓ, with target accuracy $\epsilon_\ell = 2^{-\ell}$ and active set $\mathcal{V}_\ell$, we compute an approximate G-optimal design π_ℓ supported on $O(d_\ell)$ actions, where $d_\ell = \dim(\mathrm{span}(\mathcal{V}_\ell))$. Writing $M_\ell = \sum_{a_i \in S_\ell} \pi_\ell(a_i) a_i a_i^\top$, $\gamma_{\ell,i}(a) = \pi_\ell(a_i) a_i^\top M_\ell^+ a$, the design guarantees $a = \sum_{a_i \in S_\ell} \gamma_{\ell,i}(a) a_i$, $\sum_{a_i \in S_\ell} \frac{\gamma_{\ell,i}(a)^2}{\pi_\ell(a_i)} \lesssim d_\ell$. The algorithm estimates only the rewards of the support actions, reconstructs the mean of every active action, and eliminates actions whose predicted means are more than $2\epsilon_\ell$ below the predicted maximum. The two action-set regimes require different reconstruction arguments.

Finite action sets. For finite $\mathcal{A}$, we use independent, unbiased raw-moment quantum mean estimates on the design support. The query budget assigned to support action a_i is proportional to $\sqrt{\pi_\ell(a_i)}$. This allocation, together with the von Bahr–Esseen inequality, controls the $(1+\alpha)$-moment of the reconstruction error $\widetilde{\mu}_\ell(a) - \mu(a) = \sum_{a_i \in S_\ell} \gamma_{\ell,i}(a) \big(\widehat{\mu}_\ell(a_i) - \mu(a_i)\big)$. Each reconstruction is initially accurate with constant probability. Independent timeout-protected repetitions followed by action-wise medians provide simultaneous accuracy over all $n = |\mathcal{A}|$ actions. The complete estimator interface, query allocation, and phased algorithm appear in Appendix C.2.

General action sets. We assume access to two computational oracles on every active set $\mathcal{V} \subseteq \mathcal{A}$ from Algorithm 3, each of which is $\mathcal{A}$ intersected with finitely many halfspaces. The design oracle returns a support $S \subseteq \mathcal{V}$ with $|S| \leq c_0 \dim \mathrm{span}(\mathcal{V})$ and weights π satisfying equation 25. The linear-optimization oracle returns $\arg\max_{a \in V} a^\top \theta$ for any $\theta \in \mathbb{R}^d$. Neither oracle makes reward queries. When $\mathcal{A}$ is convex, each active set is convex and linear optimization reduces to a convex program; the design oracle, which involves maximizing a quadratic form over $\mathcal{V}$, is treated as a primitive.

For a compact, possibly infinite action set, an action-wise median and union bound are unavailable. Instead, we estimate each of the $O(d_\ell)$ support means with high probability to accuracy $\epsilon_\ell / \sqrt{d_\ell}$. On this support-level event, the design identity and Cauchy–Schwarz give $\sup_{a \in \mathcal{V}_\ell} |\widetilde{\mu}_\ell(a) - \mu(a)| \lesssim \sqrt{d_\ell} \max_{a_i \in S_\ell} |\widehat{\mu}_\ell(a_i) - \mu(a_i)| \leq \epsilon_\ell$. This is a deterministic uniform reconstruction bound once the support estimates are accurate. It controls the entire compact active set without a covering argument or a union bound over actions. The complete algorithm is given in Appendix C.3.

Theorem 4.2 (Improved quantum linear-bandit regret)**.** *Under the bounded raw-moment condition, the following guarantees hold.*

1. If $\mathcal{A} \subset \mathbb{R}^d$ is finite with $n = |\mathcal{A}|$, the finite-action elimination algorithm (Algorithm 2) satisfies

$$\mathsf{Reg}_T = \widetilde{O}\Big(u^{\frac{1}{1+\alpha}} d^{\frac{1+3\alpha}{2(1+\alpha)}} (\log(nT))^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}} + d \log T \log(nT)\Big).$$

Algorithm 3 General-action heavy-tailed quantum linear elimination

Input: compact $\mathcal{A} \subseteq \mathbb{R}^d$, horizon T, confidence δ, α, raw-moment scale u, and design and linear-optimization oracles.

1. Set $\mathcal{V}_1 \leftarrow \mathcal{A}$, $N \leftarrow 0$, and $L_T \leftarrow \lceil \log_2 T \rceil$.
2. For $\ell = 1, \ldots, L_T$, set $\epsilon_\ell = 2^{-\ell}$ and compute $(S_\ell, \pi_\ell, M_\ell)$ from Lemma C.4.
3. Estimate each support mean to accuracy equation 34 and failure probability $\delta/[2|S_\ell|(\ell+1)^2]$. Before starting, stop the phased procedure if the deterministic query cap exceeds the remaining horizon.
4. Form the linear predictor
$$\widetilde{\mu}_\ell(a) = \sum_{a_i \in S_\ell} \pi_\ell(a_i) a_i^\top M_\ell^+ a \, \widehat{\mu}_\ell(a_i).$$
Use the optimization oracle to compute $\widetilde{\mu}_\ell^{\max} = \max_{a \in \mathcal{V}_\ell} \widetilde{\mu}_\ell(a)$ and retain
$$\mathcal{V}_{\ell+1} = \{a \in \mathcal{V}_\ell : \widetilde{\mu}_\ell(a) \geq \widetilde{\mu}_\ell^{\max} - 2\epsilon_\ell\}.$$

Use an empirical maximizer from the last completed phase for the remaining rounds; if no phase completed, use any action in $\mathcal{A}$.

The displayed dependence on n is not absorbed into the $\widetilde{O}$ notation.

2. *If $\mathcal{A} \subset \mathbb{R}^d$ is compact, then, given access to a small-support approximate G-design oracle and a linear-optimization oracle over each active set, the general-action elimination algorithm (Algorithm 3) satisfies*
$$\mathsf{Reg}_T = \widetilde{O}\Big(u^{\frac{1}{1+\alpha}} d^{\frac{1+5\alpha}{2(1+\alpha)}} T^{\frac{1-\alpha}{1+\alpha}} + d \log T\Big).$$
This guarantee is independent of the cardinality and covering numbers of $\mathcal{A}$.

Remark 4.3 (Improvement and source of the gains)**.** *The outer phased-elimination framework and G-optimal design are standard; our contribution lies in adapting their estimation and reconstruction steps to quantum heavy-tailed rewards. For finite action sets, design-matched query allocation and $(1+\alpha)$-moment aggregation avoid the worst-case absolute-error accumulation incurred by applying a high-probability scalar estimator separately to every support action. For general action sets, estimating only $O(d)$ support actions and using deterministic uniform reconstruction avoids discretizing the action set. The general-action guarantee improves the $\widetilde{O}(d^2 T^{(1-\alpha)/(1+\alpha)})$ bound of Wu et al. (2023) throughout $\alpha \in (0, 1]$. At $\alpha = 1$, the leading term becomes $\widetilde{O}(d^{3/2})$. Since bounded rewards in $[-1, 1]$ satisfy $E|Y_a|^2 \leq 1$, this setting contains the bounded-reward model of Wan et al. (2023). Our guarantee therefore improves their $O(d^2 \operatorname{polylog} T)$ general-action regret by a factor of $\sqrt{d}$ given the design and optimization oracles. In the finite-action case, the leading dimension dependence reduces to d at $\alpha = 1$, up to the displayed logarithmic dependence on n and T.*

5 Experiments

We complement our theoretical analysis with synthetic experiments across multi-armed and linear bandit settings. Our evaluation examines how tail heaviness, problem scale, and the feedback model affect finite-horizon performance, with further details and additional results deferred to Appendix E.

Synthetic bandit instances. We consider stochastic MABs, finite-action linear bandits, and general-action linear bandits with rewards $Y_t = \mu(a_t) + \xi_t$, where $\mu(a) = \mu_a$ for MABs and $\mu(a) = a^\top \theta^\star$ for linear bandits. The noise follows a symmetric Pareto distribution with shape $1.5 + \alpha$, scaled so that $\mathbb{E}|\xi_t|^{1+\alpha} = 0.25$, for $\alpha \in \{0.25, 0.5, 0.75, 1\}$. Figure 1 presents representative MAB instances with $K = 5$, a finite-action instance with $d = 8$ and $n = 128$, and a general-action instance on $\mathcal{A} = [-1/2, 1/2]^4$. To cover a range of dimensions, a complementary finite-action instance with $d = 2$ is reported for all $\alpha \in \{0.25, 0.5, 0.75, 1\}$ in Figure 3 of Appendix E. Complete instance specifications and additional parameter sweeps are provided in Appendix E.

Experimental settings and metric. We compare Classical Raw (CR), Classical Central (CC), Quantum Raw (QR), and Quantum Central (QC). All variants share phased elimina-

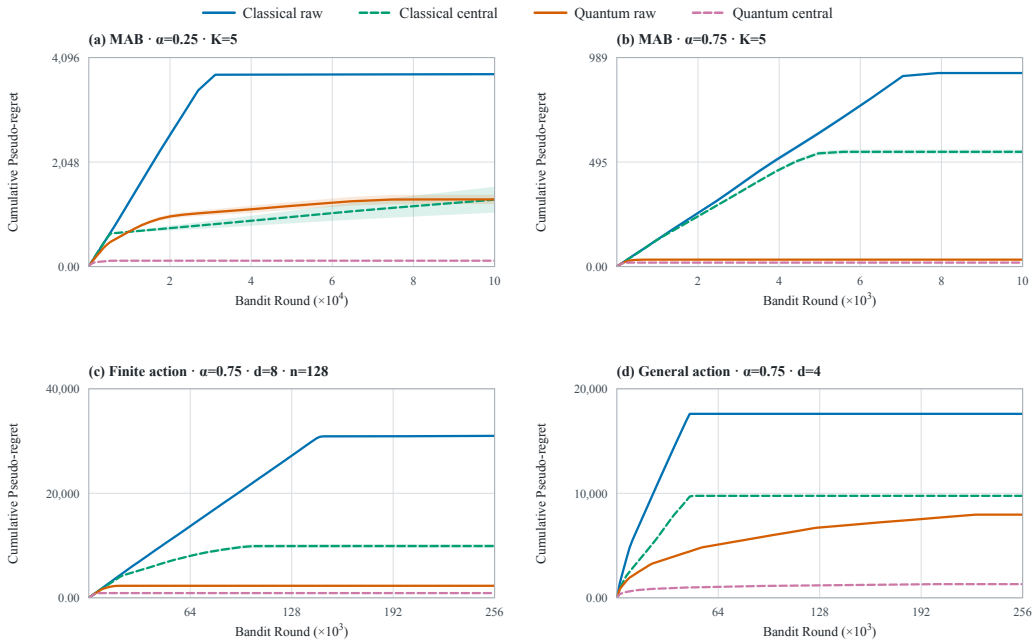


Figure 1: Cumulative pseudo-regret of classical and quantum phased elimination algorithms under Pareto distribution.

tion with $\epsilon_\ell = 2^{-\ell}$ and $\delta = 0.05$. Linear-bandit experiments use approximate G-optimal designs with tolerance 0.1 and linear reconstruction. Classical variants use truncation or median-of-means; quantum variants emulate the QTME/QHTME query rates using independent uniform errors within prescribed accuracy intervals. Raw and central variants use moment bounds u and v, respectively. Each nominal oracle query is charged as one bandit round. We report cumulative pseudo-regret computed from the true action means and the number of rounds assigned to each action. Curves show averages over 500 repetitions for MABs and 200 repetitions for linear bandits, with shaded regions representing pointwise 95% normal confidence intervals.

Results. Figure 1 illustrates how estimation access and moment assumptions affect finite-horizon regret. For the heavily tailed MAB instance in panel (a), QC achieves the lowest regret, while QR improves over CR but remains close to CC. At $\alpha = 0.75$, both quantum variants outperform both classical baselines in the MAB, finite-action, and general-action instances shown in panels (b)–(d). QC performs best in all four panels. For linear bandits, the results illustrate how lower estimation costs can benefit design-based exploration and reward reconstruction. The piecewise growth and plateaus are consistent with complete-phase budget constraints.

6 Conclusion

We studied stochastic multi-armed and linear bandits with heavy-tailed rewards under quantum reward-oracle access. For multi-armed bandits, we established matching regret upper and lower bounds up to logarithmic factors, characterizing the optimal dependence on the number of arms and the query budget. For linear bandits, we improved the dimension dependence across the full moment range considered in this work. Our finite-action guarantees exploit the size of the action set, while our general-action guarantees remain independent of its cardinality under the stated design and optimization oracle assumptions. We also established complementary lower bounds for finite and compact infinite action sets. Simulations illustrate how moment assumptions and design allocations affect finite-horizon performance.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the draft, assistance with LaTeX formatting, and debugging the simulation code. Additionally, we used generative AI tools to obtain feedback on the presentation of an earlier draft. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily theoretical. It studies the statistical limits of learning from heavy-tailed rewards under quantum oracle access, and it evaluates the proposed algorithms only on synthetic bandit instances generated from known distributions (Section 5 and Appendix E). The study involves no human subjects, personal or sensitive data, or real-world datasets, and it releases no data. We are not aware of direct negative societal impacts specific to this work beyond those generally associated with advances in sequential decision-making and quantum algorithms.

REPRODUCIBILITY STATEMENT

For the theoretical results, the problem setting, moment conditions, and quantum reward-oracle model are stated in Section 2, with the oracle access model formalized in Section 2.2. Complete proofs of the multi-armed bandit lower and upper bounds (Theorems 3.1 and 3.2) appear in Appendix B. Complete proofs of the linear-bandit lower and upper bounds (Theorems 4.1 and 4.2), including the design lemma (Lemma C.4), appear in Appendix C. The central-moment extension is proved in Appendix D. The quantum mean-estimation primitives we use are taken from prior work, and their exact interfaces are stated in Lemmas B.4, C.3, D.1, and D.3. The derivation of the raw-moment unbiased interface is given in Appendix D.2.1. Pseudocode for all algorithms is provided in Algorithms 1, 2, 3, and 4. For the experiments, Appendix E specifies all bandit instances, noise distributions, horizons, budget constants, numbers of repetitions, and the construction of confidence intervals.

REFERENCES

Shubhada Agrawal, Sandeep K. Juneja, and Wouter M. Koolen. Regret minimization in heavy-tailed bandits. In *Proceedings of the Thirty-Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 26–62. PMLR, 2021. URL https://proceedings.mlr.press/v134/agrawal21a.html.

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal discrete optimization for experimental design: A regret minimization approach. *Mathematical Programming*, 186(1): 439–478, 2021.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013. doi: 10.1109/TIT.2013.2277869.

Balthazar Casalé, Giuseppe Di Molfetta, Hachem Kadri, and Liva Ralaivola. Quantum bandits. *Quantum Machine Intelligence*, 2(1):11, 2020. doi: 10.1007/s42484-020-00024-8.

Gianmarco Genalti, Lupo Marsigli, Nicola Gatti, and Alberto Maria Metelli. (ϵ, u)-adaptive regret minimization in heavy-tailed bandits. In *Proceedings of the Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1882–1915. PMLR, 2024. URL https://proceedings.mlr.press/v247/genalti24a.html.

Brassard Gilles, Høyer Peter, Mosca Michele, and Tapp Alain. Quantum amplitude amplification and estimation. *Contemporary Mathematics*, 305:53–74, 2002. ISSN 0271-4132. doi: 10.1090/conm/305/05215. URL https://cir.nii.ac.jp/crid/1362544419940146816.

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

Yasunari Hikima, Kazunori Murao, Sho Takemori, and Yuhei Umeda. Quantum kernelized bandits. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1640–1657. PMLR, 2024. URL https://proceedings.mlr.press/v244/hikima24a.html.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8452–8462, 2020.

Maoli Liu, Zhuohua Li, and John Lui. Quantum multi-armed bandits and linear bandits: Lower bounds and algorithms. *arXiv preprint arXiv:2608.14319*, 2026.

Bin Luo, Chengchang Liu, Jonathan Allcock, Shengyu Zhang, and John Lui. Quantum speedups for stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2607.25492*, 2026.

Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1642–1650. PMLR, 2016. URL https://proceedings.mlr.press/v48/medina16.html.

Ashley Montanaro. Quantum speedup of monte carlo methods. *Proceedings of the Royal Society A*, 471(2181):20150301, 2015. doi: 10.1098/rspa.2015.0301.

Han Shao, Xiaotian Yu, Irwin King, and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8420–8429, 2018.

Artin Tajdini, Jonathan Scarlett, and Kevin Jamieson. Improved regret bounds for linear bandits with heavy-tailed rewards. *arXiv preprint arXiv:2506.04775*, 2025.

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*, pp. 299–303, 1965.

Zongqi Wan, Zhijie Zhang, Tongyang Li, Jialin Zhang, and Xiaoming Sun. Quantum multi-armed bandits and stochastic linear bandits enjoy logarithmic regrets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10087–10094, 2023. doi: 10.1609/aaai.v37i8.26202.

Daochen Wang, Xuchen You, Tongyang Li, and Andrew M. Childs. Quantum exploration algorithms for multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10102–10110, 2021. doi: 10.1609/aaai.v35i11.17212.

Lai Wei and Vaibhav Srivastava. Minimax policy for heavy-tailed bandits. *IEEE Control Systems Letters*, 5(4):1423–1428, 2021. URL https://arxiv.org/abs/2007.10493.

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

Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 2936–2942. International Joint Conferences on Artificial Intelligence Organization, 2020. doi: 10.24963/ijcai.2020/406.

A Related Work

Classical heavy-tailed MAB. Finite-time regret under only a bounded $(1+\alpha)$-moment was systematically studied by Bubeck et al. (2013). Using truncated means, median-of-means, and Catoni estimators, they obtained the gap-independent rate $\widetilde{O}\Big(K^{\frac{\alpha}{1+\alpha}}T^{\frac{1}{1+\alpha}}\Big)$ and a matching polynomial

lower bound, up to logarithmic factors. Robust MOSS subsequently removed the additional worst-case logarithmic loss and attained the minimax rate (Wei & Srivastava, 2021). Later work developed estimators that do not require the numerical moment bound (Lee et al., 2020), first-order asymptotically optimal KL-based indices (Agrawal et al., 2021), and algorithms adaptive to unknown tail parameters, together with limitations on assumption-free adaptation (Genalti et al., 2024). These results form the classical benchmark for our quantum MAB guarantees.

Classical heavy-tailed linear bandits. Medina & Yang (2016) initiated the study of heavy-tailed linear bandits using truncation and median-of-means. Shao et al. (2018) subsequently introduced MENU under a central-moment condition and TOFU under a raw-moment condition, obtaining $\widetilde{O}\Big(dT^{\frac{1}{1+\alpha}}\Big)$ regret for general action sets. Their frequently quoted $\Omega\Big(dT^{\frac{1}{1+\alpha}}\Big)$ lower bound, however, uses a moment budget of order d; normalizing the budget to be dimension-independent yields only $\Omega\Big(d^{\frac{\alpha}{1+\alpha}}T^{\frac{1}{1+\alpha}}\Big)$. For finite-action contextual linear bandits, Xue et al. (2020) obtained $\widetilde{O}\Big(\sqrt{d}\,T^{\frac{1}{1+\alpha}}\Big)$ regret. Under a dimension-independent central-moment bound, Tajdini et al. (2025) recently improved the general-action upper and lower bounds to $\widetilde{O}\Big(d^{\frac{1+3\alpha}{2(1+\alpha)}}T^{\frac{1}{1+\alpha}}\Big)$ and $\Omega\Big(d^{\frac{2\alpha}{1+\alpha}}T^{\frac{1}{1+\alpha}}\Big)$, respectively, and established refined $\log n$-dependent guarantees for finite action sets. Our raw-moment assumption implies a central-moment bound via $\mathbb{E}|Y-\mathbb{E}Y|^{1+\alpha} \le 2^{1+\alpha}\mathbb{E}|Y|^{1+\alpha}$, but the converse need not hold.

Quantum bandits. Quantum amplitude estimation and quantum Monte Carlo provide the mean-estimation primitives underlying quantum bandit algorithms (Gilles et al., 2002; Montanaro, 2015; Hamoudi, 2021). Early quantum bandit work focused on best-arm identification under coherent superposition access to action labels (Casalé et al., 2020; Wang et al., 2021), a stronger oracle than the arm-wise access considered here. Wan et al. (2023) introduced the cumulative-regret model adopted in this paper: the action is selected classically, and each use of $\mathcal{O}_a$ or $\mathcal{O}_a^\dagger$ is charged as one round. For bounded rewards, they obtained $O(K\log T)$ regret for MAB and $O(d^2\,\mathrm{polylog}\,T)$ regret for general-action linear bandits. Hikima et al. (2024) extended this framework to kernelized bandits, while Liu et al. (2026) derived bounded-reward regret lower bounds and sharper finite-action linear-bandit algorithms. Most closely related to our work, Wu et al. (2023) developed raw-moment quantum mean estimation and obtained $\widetilde{O}\Big(KT^{\frac{1-\alpha}{1+\alpha}}\Big)$ MAB regret for every $\alpha\in(0,1]$ and $\widetilde{O}\Big(d^2T^{\frac{1-\alpha}{1+\alpha}}\Big)$ linear-bandit regret for $\alpha\in[1/3,1]$. Luo et al. (2026) developed complementary high-probability and unbiased estimators under central-moment assumptions. We inherit the raw-moment estimator from Wu et al. (2023); our contributions lie in the bandit-level allocation and elimination analyses, improved regret guarantees, and stopped-query lower bounds.

B Omitted Algorithm and Proofs for Quantum Heavy-tailed MAB under Bounded Raw Moments

B.1 Minimax lower bound

B.1.1 Rare-event quantum testing

We prove the lower bound under the raw-moment assumption. Fix $p\in(0,p_0]$ and set

$$B=(u/(2p))^{1/(1+\alpha)}. \tag{8}$$

Define

$$P_-=(1-p/2)\delta_0+(p/2)\delta_B,\qquad P_+=(1-p)\delta_0+p\delta_B. \tag{9}$$

Both distributions have raw $(1+\alpha)$-moment at most u, since $rB^{1+\alpha}\le u/2$ for $r\in\{p/2,p\}$. Their means are $pB/2$ and pB, so the mean gap is

$$\Delta=\frac{pB}{2}=\Theta\Big(u^{1/(1+\alpha)}p^{\alpha/(1+\alpha)}\Big). \tag{10}$$

Lemma B.1 (Stopped-query hybrid bound)**.** *For a linear operator A on a Hilbert space H, write*

$$\|A\|_{\mathrm{op}} := \sup_{|\phi\rangle\in\mathsf{H}:\,\||\phi\rangle\|_2=1} \|A|\phi\rangle\|_2,$$

the operator (or spectral) norm, equivalently the largest singular value of A. Let O_0 and O_1 be two unitary candidate oracles acting on the same query register, and put

$$D := \|O_1 - O_0\|_{\mathrm{op}} = \sup_{\||\phi\rangle\|_2=1} \|(O_1 - O_0)|\phi\rangle\|_2,$$

so D is the largest Euclidean change that one oracle call can produce on a common normalized input state. An unknown hypothesis $J \in \{0,1\}$ is fixed, and the test is given black-box access only to O_J and $O_J^\dagger$ (tensored with the identity on its private workspace). Thus J is not revealed to the test: it indexes which of the two oracle instances is present and does not denote a bandit action. The test may use intermediate measurements and a data-dependent stopping rule. Let N be its number of oracle calls and $\widehat{J}$ its output, and write $\Pr_j$ and $\mathbb{E}_j$ when $J = j$. Suppose

$$\Pr_j(\widehat{J} \neq j) \le \frac{1}{16}, \qquad j \in \{0,1\}, \qquad Q := \mathbb{E}_0 N.$$

There are universal constants $c, D_0 > 0$ such that, whenever $D \le D_0$,

$$Q \ge \frac{c}{D}. \tag{11}$$

Proof. Set $m = \lceil 8Q \rceil$ and form a fixed-budget test by interrupting the original test immediately after its mth query. On timeout, output "alternative." Under O_0, Markov's inequality gives $\Pr_0(N > m) \le Q/m \le 1/8$, so the new null error is at most $1/16 + 1/8 = 3/16$. Under O_1, the forced timeout answer is correct, so the alternative error remains at most $1/16$. Hence the probability of outputting "alternative" differs by at least $3/4$ between the two oracles.

Purify all intermediate measurements and classically controlled operations. Pad branches that stop early by oracle-independent identity operations. The result is an m-query unitary computation followed by one measurement. At each query slot, the history register controls a direct sum of I, O_j, and $O_j^\dagger$; consequently the operator-norm difference between the slot under $J = 1$ and the same slot under $J = 0$ is at most $\max\{\|O_1 - O_0\|_{\mathrm{op}}, \|O_1^\dagger - O_0^\dagger\|_{\mathrm{op}}\}$. Let $|\Phi_t^{(j)}\rangle$ be its state just after the tth query when the oracle is O_j. A telescoping hybrid argument and unitary invariance give

$$\big\| |\Phi_m^{(1)}\rangle - |\Phi_m^{(0)}\rangle \big\|_2 \le \sum_{t=1}^{m} \max\{\|O_1 - O_0\|_{\mathrm{op}}, \|O_1^\dagger - O_0^\dagger\|_{\mathrm{op}}\} = mD.$$

The two operator norms are equal. The variation distance between the outcomes of any final measurement is at most the trace distance of the final states, which is at most a universal constant times mD. The preceding $3/4$ separation therefore implies $mD \ge c_0$. Since $m \le 8Q + 1$, taking D_0 small enough yields equation 11. □

Theorem B.2 (Constant-confidence quantum mean testing)**.** *There are constants $c, c_0 > 0$ such that, for every $\alpha \in (0,1]$, $u > 0$, and sufficiently small $\Delta > 0$, any quantum algorithm that distinguishes P_- from P_+ with error at most c_0 satisfies*

$$\mathbb{E}_{P_-} N \ge c\, u^{1/(2\alpha)} \Delta^{-\frac{1+\alpha}{2\alpha}}. \tag{12}$$

Proof. Choose the normalized auxiliary state to be fixed and extend the state preparation in equation 6 to the canonical planar rotation

$$O_r = \begin{pmatrix} \sqrt{1-r} & -\sqrt{r} \\ \sqrt{r} & \sqrt{1-r} \end{pmatrix}$$

for $r \in [0,1]$, on the span of the reward basis states $|0\rangle$ and $|B\rangle$. Here r is the probability of the rare reward B, so the two hypothesis oracles are $O_- := O_{p/2}$ and $O_+ := O_p$. Their rotation angles are $\vartheta_- = \arcsin\sqrt{p/2}$ and $\vartheta_+ = \arcsin\sqrt{p}$, hence

$$\|O_+ - O_-\|_{\mathrm{op}} = 2\sin\left(\frac{|\vartheta_+ - \vartheta_-|}{2}\right) = \Theta(\sqrt{p}).$$

Applying Lemma B.1 gives $\mathbb{E}_{P_-} N = \Omega(p^{-1/2})$. Solving equation 10 for p gives

$$p = \Theta\Big(\Delta^{(1+\alpha)/\alpha} u^{-1/\alpha}\Big),$$

which proves equation 12. □

B.1.2 BANDIT REDUCTION

Use one reference arm with deterministic reward $3pB/4$ and $K-1$ candidate arms. In the null instance, every candidate has distribution P_-; in alternative i, candidate i has distribution P_+ and the other candidates remain P_-. The reference is optimal under the null, candidate i is optimal under alternative i, and every relevant gap equals

$$g = \frac{pB}{4} = \Theta\Big(u^{1/(1+\alpha)} p^{\alpha/(1+\alpha)}\Big). \tag{13}$$

The deterministic arm respects the moment constraint when p is below a universal constant.

Lemma B.3 (One-sided bandit-to-testing reduction)**.** *Let a policy have worst-case expected regret at most M on the preceding hard family. If $M \le cTg$, then for every candidate i its expected number of null-instance queries satisfies*

$$\mathbb{E}_0 N_i(T) = \Omega\Big(u^{1/(2\alpha)} g^{-\frac{1+\alpha}{2\alpha}}\Big).$$

Consequently,

$$M = \Omega\Big(\min\Big\{Tg,\ K u^{1/(2\alpha)} g^{1-\frac{1+\alpha}{2\alpha}}\Big\}\Big). \tag{14}$$

Proof. Let $N_i(T)$ count rounds in which the policy selects candidate action i. Run the policy as a tester and output "alternative i" when $N_i(T) > T/2$. Under the null, action i has gap g, and hence

$$\mathbb{P}_0(N_i(T) > T/2) \le \frac{2M}{Tg}.$$

Under alternative i, every other action has gap at least g. Therefore

$$\mathbb{P}_i(N_i(T) \le T/2) \le \frac{2M}{Tg}.$$

All oracles except that of action i are fixed and can be simulated without querying the unknown oracle. When the policy queries action i, the tester uses its unknown oracle; a round in which the policy makes no oracle call costs the tester nothing. Thus the test's null expected query count is at most $\mathbb{E}_0 N_i(T)$. If $M \le cTg$, both testing errors are at most $1/16$, and Theorem B.2 yields the displayed lower bound on $\mathbb{E}_0 N_i(T)$. Summing over the $K-1$ candidates and multiplying by their null gap g gives the second branch of equation 14. If $M > cTg$, the first branch already holds. □

Balancing the two branches in equation 14 means $T \asymp K u^{1/(2\alpha)} g^{-\frac{1+\alpha}{2\alpha}}$, or

$$g \asymp u^{1/(1+\alpha)} (K/T)^{\frac{2\alpha}{1+\alpha}}. \tag{15}$$

This yields the polynomial lower bound. A second lower bound follows because bounded Bernoulli rewards are contained in the raw-moment class after constant rescaling: the point-versus-interval lower-bound construction of Liu et al. (2026) gives $\Omega(K \log(T/K))$. Taking the larger of the two bounds is, up to a factor two, their sum.

Theorem 3.1 (QMAB minimax lower bounds, restated)**.** *There exist universal constants $c, C > 0$ such that, for every $\alpha \in (0,1]$, $u > 0$, $K \ge 2$, and $T \ge CK$,*

$$\inf_{\mathsf{Alg}} \sup_{\boldsymbol{\nu} \in (\mathcal{H}_\alpha^{\mathrm{raw}}(u))^K} \mathsf{Reg}_T(\mathsf{Alg}, \boldsymbol{\nu}) \ \ge c \left[T \min\left\{1,\ u^{\frac{1}{1+\alpha}} \left(\frac{K}{T}\right)^{\frac{2\alpha}{1+\alpha}} \right\} + \min\Big\{1, u^{\frac{1}{1+\alpha}}\Big\} K \log \frac{T}{K} \right]. \tag{16}$$

Here the supremum ranges over valid arm-wise quantum reward-oracle instances whose reward distributions belong to $\mathcal{H}_\alpha^{\mathrm{raw}}(u)$.

Consequently, in the non-saturated regime, the polynomial heavy-tail term is $\Omega\Big(u^{\frac{1}{1+\alpha}} K^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}}\Big)$. If u is bounded below by a positive universal constant, the second term in equation 16 simplifies to $\Omega(K \log(T/K))$.

Proof. Substitute equation 15 into equation 14 when the resulting gap is below the universal constant allowed by the construction. Otherwise choose a fixed sufficiently small gap; the first branch gives $\Omega(T)$. The bounded-reward construction supplies the logarithmic term, and the maximum of the two valid lower bounds dominates half their sum. □

B.2 QMAB Upper-Bound Details

The raw-moment algorithm uses the estimator of Wu et al. (2023).

Lemma B.4 (Raw-moment QTME Wu et al. (2023))**.** *If $\mathbb{E}|Y|^{1+\alpha} \le u$, then for every accuracy $\rho \in (0, 1]$ and failure probability $\delta \in (0, 1/3)$, the QTME of Wu et al. (2023) returns $\widehat{\mu}$ such that*

$$\Pr(|\widehat{\mu} - \mu| > \rho) \le \delta$$

using at most

$$O_\alpha\left(u^{\frac{1}{2\alpha}}\rho^{-\frac{1+\alpha}{2\alpha}}\left(\log\frac{1}{\delta}\right)^{5/2}\right) \tag{17}$$

oracle queries, up to rounding and lower-order logarithms. In particular, the confidence dependence is not suppressed in equation 17.

Source and parameter conversion. Wu et al. (2023)'s Theorem 3 uses a tuning parameter N, makes $O(N\log^{3/2}(1/\delta))$ queries, and gives error

$$O_\alpha\left(u^{\frac{1}{1+\alpha}}\frac{(\log(1/\delta))^{\frac{2\alpha}{1+\alpha}}}{N^{\frac{2\alpha}{1+\alpha}}}\right).$$

Solving this expression for N at target error ρ contributes one further factor $\log(1/\delta)$; multiplying by the $\log^{3/2}(1/\delta)$ implementation cost gives equation 17. No new mean-estimation result is claimed here. □

For later use, let $Q_{\mathrm{raw}}(\rho, \delta; u)$ be a deterministic upper bound from Lemma B.4. The following phased policy differs from the batch-UCB algorithm of Wu et al. (2023): every surviving action is estimated at the same accuracy in a phase, and actions are eliminated before the next phase.

Theorem 3.2 (Heavy-tail quantum arm elimination, restated)**.** *For every $T \ge K$, Algorithm 1 with $\delta = T^{-2}$ satisfies, under the raw-moment condition,*

$$\mathsf{Reg}_T = \widetilde{O}\Big(u^{1/(1+\alpha)}K^{\frac{2\alpha}{1+\alpha}}T^{\frac{1-\alpha}{1+\alpha}} + K\log T\Big).$$

Proof. Let $\mathcal{E}$ be the event that every completed estimate satisfies $|\widehat{\mu}_\ell(a) - \mu(a)| \le \epsilon_\ell$. Conditional on the history before phase ℓ, its active set is fixed. Hence

$$\Pr(\mathcal{E}^c) \le \sum_{\ell\ge1}|\mathcal{V}_\ell|\delta_\ell \le \frac{\delta}{2}\sum_{\ell\ge1}(\ell+1)^{-2} < \delta.$$

Fix $a^\star \in \arg\max_a \mu(a)$. On $\mathcal{E}$, induction shows that $a^\star$ is never eliminated: for every active a,

$$\widehat{\mu}_\ell(a^\star) \ge \mu^\star - \epsilon_\ell \ge \mu(a) - \epsilon_\ell \ge \widehat{\mu}_\ell(a) - 2\epsilon_\ell.$$

If $a \in \mathcal{V}_{\ell+1}$, then

$$\mu(a) \ge \widehat{\mu}_\ell(a) - \epsilon_\ell \ge \widehat{\mu}_\ell(a^\star) - 3\epsilon_\ell \ge \mu^\star - 4\epsilon_\ell.$$

Thus every query in phase $\ell \ge 2$ incurs regret at most $4\epsilon_{\ell-1} = 8\epsilon_\ell$; the same bound holds in the first phase after enlarging the constant because means lie in $[-1, 1]$.

Let L be the last completed phase. Suppressing only polylogarithmic factors in T, K, u, Lemma B.4 gives

$$q_\ell = \widetilde{O}\Big(u^{\frac{1}{2\alpha}}\epsilon_\ell^{-\frac{1+\alpha}{2\alpha}} + 1\Big).$$

The geometric sum of phase budgets is bounded by a constant multiple of its last term (and by an additional logarithm when $\alpha = 1$). Hence the budget test in Algorithm 1 implies the terminal accuracy

$$\epsilon_L = \widetilde{O}\left(u^{\frac{1}{1+\alpha}}\left(\frac{K}{T}\right)^{\frac{2\alpha}{1+\alpha}}\right).$$

The exploitation regret is at most $4T\epsilon_L$. The exploration regret on $\mathcal{E}$ is

$$\sum_{\ell=1}^{L} 8\epsilon_\ell |\mathcal{V}_\ell| q_\ell \le \widetilde{O}\Big(u^{\frac{1}{1+\alpha}} K^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}} + K\log T\Big).$$

The $+1$ rounding in every arm–phase pair gives $K\log T$. Finally, $\mathsf{Reg}_T \le 2T$ deterministically, so the failure event contributes at most $2T\delta = 2/T$. □

C Omitted Algorithm and Proofs for Quantum Heavy-tailed Linear Bandits under Bounded Raw Moments

C.1 Lower-bound proofs

C.1.1 Finite-action embedding

The multi-armed hard family embeds in a linear model without using $K \gg d$. Let

$$a_1 = e_1, \qquad a_i = \frac{2}{3}e_1 + \frac{\sqrt{5}}{3}e_i, \quad i = 2, \dots, d. \tag{18}$$

Every action has unit norm. Set

$$\theta^{(0)} = \frac{3pB}{4}e_1, \qquad \theta^{(i)} = \theta^{(0)} + \frac{3pB}{2\sqrt{5}}e_i. \tag{19}$$

Under $\theta^{(0)}$, action a_1 has mean $3pB/4$ and every a_i, $i \ge 2$, has mean $pB/2$. Under $\theta^{(i)}$, the mean of a_i becomes pB, while all other means are unchanged. Assign the deterministic reference distribution to a_1, P_- to a baseline candidate, and P_+ to the elevated candidate. This is a valid linear-mean reward family; for small pB, all parameters have norm at most one.

Theorem 4.1 (Finite-action QLB lower bound, restated)**.** *For every $d \ge 2$ and $T \ge Cd$, there exists a fixed action set $\mathcal{A} \subset \mathbb{R}^d$ with $|\mathcal{A}| = d$ such that every quantum linear-bandit policy has raw-moment regret*

$$\mathsf{Reg}_T = \Omega\left(T\min\left\{1, u^{\frac{1}{1+\alpha}}\left(\frac{d}{T}\right)^{\frac{2\alpha}{1+\alpha}}\right\}\right).$$

For normalized $u = \Theta(1)$, the bound also contains the additive term $\Omega(d\log(T/d))$.

Proof. The mean vectors induced by equation 19 reproduce the multi-armed hard family exactly, so Lemma B.3 applies with $K = d$. Optimizing g, or fixing it at a sufficiently small universal constant when the optimum would be larger, gives the displayed raw-moment term. The bounded-reward $\Omega(d\log(T/d))$ construction embeds through the same action set as in Liu et al. (2026). The maximum of the polynomial and logarithmic lower bounds dominates half their sum. □

The finite embedding is enough for minimax lower bounds over the class of arbitrary action sets. The next result shows that finiteness itself is not the source of hardness. Radially extend the actions in equation 18 and set

C.1.2 Compact infinite action set

$$\mathcal{A}_{\mathrm{ray}} := \bigcup_{i=1}^{d} \{ta_i : t \in [0,1]\}. \tag{20}$$

This is a compact, infinite, continuously parameterized action set. For a candidate ray, define the reward at ta_i to be B with probability tr_i and zero otherwise, where $r_i \in \{p/2, p\}$ is the rare-event probability at a_i; on the reference ray use the deterministic reward $3tpB/4$. These distributions have the required linear means and satisfy the raw-moment bound.

Theorem C.1 (Continuous-action QLB lower bound)**.** *For every $d \geq 2$ and $T \geq Cd$, there is a compact infinite action set $\mathcal{A}_{\mathrm{ray}} \subset \mathbb{R}^d$ for which every quantum linear-bandit policy suffers*

$$\mathsf{Reg}_T = \Omega\left(T \min\left\{1, u^{\frac{1}{1+\alpha}} \left(\frac{d}{T}\right)^{\frac{2\alpha}{1+\alpha}}\right\}\right) \tag{21}$$

under the raw-moment assumption. In the normalized regime one may add $\Omega(d \log(T/d))$.

Proof. Under the null instance, every nonzero point on candidate ray i has gap at least $3pB/4 - tpB/2 \geq pB/4 = g$. Under alternative i, every action outside ray i has gap at least g. Therefore the testing reduction in Lemma B.3 remains valid when $N_i(T)$ counts all queries to the ith ray. The common origin is assigned to no candidate ray; it is deterministic, uninformative, and has gap at least g on every hard instance.

It remains to check that choosing $t < 1$ cannot make testing easier. In the canonical two-dimensional reward-oracle representation, the two candidate oracles at scale t differ by the angle

$$\left|\arcsin\sqrt{tp} - \arcsin\sqrt{tp/2}\right| \leq C\sqrt{t}\,\sqrt{p} \leq C\sqrt{p}.$$

Thus every variable-t query has operator separation at most a query at the endpoint a_i. For a tester whose query may choose a different t in every round, the hybrid sum is bounded by $C\sqrt{p}$ times the total number of queries to the ray. Applying Lemma B.1 lower bounds the null expected number of such queries by $\Omega(p^{-1/2})$. The reduction in Lemma B.3, with $N_i(T)$ counting all selections on the ith ray, then applies unchanged. Summing over rays and optimizing g (with the same constant-gap saturation as in Theorem 3.1) proves the raw-moment term. Applying the same radial thinning to the bounded-reward point-versus-interval construction in Liu et al. (2026) gives the additive logarithmic term. □

Remark C.2 (What the continuous lower bound does not prove)**.** *The set equation 20 has only d informative directions. It proves a genuine infinite-action lower bound, but it is not a lower bound for a rich convex body containing many mixed directions. Such directions can act as aggregate measurements of several parameter coordinates, and controlling their quantum information is precisely the missing ingredient in a matching general-action lower bound.*

C.2 Finite action sets

The finite-action improvement uses an unbiased raw-moment scalar estimator, rather than a high-probability estimate for every support action.

Lemma C.3 (Unbiased raw-moment QTME)**.** *Suppose $\mathbb{E}|Y_a|^{1+\alpha} \leq u$. For every target $\rho \in (0, u^{1/(1+\alpha)}]$, there is a quantum procedure returning an estimate satisfying*

$$\mathbb{E}\widehat{\mu}(a) = \mu(a), \qquad \mathbb{E}|\widehat{\mu}(a) - \mu(a)|^{1+\alpha} \leq \rho^{1+\alpha}, \tag{22}$$

using

$$\widetilde{O}_\alpha\left(u^{\frac{1}{2\alpha}} \rho^{-\frac{1+\alpha}{2\alpha}}\right) \tag{23}$$

queries in expectation. Equivalently, an expected budget q can be chosen so that the mean-zero error e satisfies

$$\mathbb{E}e = 0, \qquad \mathbb{E}|e|^{1+\alpha} \leq \widetilde{O}_\alpha(uq^{-2\alpha}). \tag{24}$$

The procedure is a direct raw-moment corollary of the unbiased estimator of Luo et al. (2026); its random query count is part of their MLMC construction. The reduction and estimator internals are given in Section D.2.1 so that the main development uses only the raw-moment interface above.

Lemma C.4 (Small-support approximate G-design)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ span a d_ℓ-dimensional subspace. There are $S = \{a_1, \dots, a_s\} \subseteq \mathcal{V}$ and $\pi \in \Delta(S)$ such that*

$$s \le c_0 d_\ell, \qquad M = \sum_{i=1}^{s} \pi(a_i) a_i a_i^\top, \qquad \sup_{a \in \mathcal{V}} a^\top M^+ a \le c_1 d_\ell. \tag{25}$$

For every $a \in \mathrm{span}(\mathcal{V})$, define

$$\gamma_i(a) = \pi(a_i) a_i^\top M^+ a. \tag{26}$$

Then

$$a = \sum_{i=1}^{s} \gamma_i(a) a_i, \qquad \sum_{i=1}^{s} \frac{\gamma_i(a)^2}{\pi(a_i)} = a^\top M^+ a \le c_1 d_\ell. \tag{27}$$

Proof. Work in an orthonormal coordinate system for $\mathrm{span}(\mathcal{V})$, whose dimension is d_ℓ. The Kiefer–Wolfowitz equivalence theorem gives an optimal probability measure ξ with information matrix $M_\xi \succ 0$ satisfying

$$\sup_{a \in \mathcal{V}} a^\top M_\xi^{-1} a = d_\ell,$$

and every point receiving positive mass has leverage d_ℓ Kiefer & Wolfowitz (1960). By Carathéodory's theorem, the same matrix can first be represented by at most $d_\ell(d_\ell + 1)/2 + 1$ weighted actions.

For such a representation, write $b_i = \sqrt{\xi_i}\, M_\xi^{-1/2} a_i$. Then $\sum_i b_i b_i^\top = I_{d_\ell}$ and $\|b_i\|_2^2 = d_\ell \xi_i$ on the support. Constant-accuracy spectral sparsification selects $O(d_\ell)$ of these vectors and coefficients $s_i \ge 0$ so that

$$\tfrac{1}{2} I_{d_\ell} \preceq \sum_i s_i b_i b_i^\top \preceq 2 I_{d_\ell}.$$

Define $\widetilde{\pi}_i = s_i \xi_i$. Taking traces and using $\|b_i\|_2^2 = d_\ell \xi_i$ shows $\sum_i \widetilde{\pi}_i \in [1/2, 2]$. After normalizing these weights to sum to one, their information matrix M satisfies $c M_\xi \preceq M \preceq C M_\xi$ for universal constants. Consequently $\sup_{a \in \mathcal{V}} a^\top M^+ a \le c_1 d_\ell$ and the support has size at most $c_0 d_\ell$. This is the standard sparsified-design construction; finite-set algorithms and rounding guarantees are given, for example, by Allen-Zhu et al. (2021).

Returning to the ambient coordinates, $a \in \mathrm{range}(M)$ and hence

$$\sum_i \gamma_i(a) a_i = M M^+ a = a.$$

Moreover,

$$\sum_i \frac{\gamma_i(a)^2}{\pi(a_i)} = \sum_i \pi(a_i)(a_i^\top M^+ a)^2 = a^\top M^+ M M^+ a = a^\top M^+ a.$$

□

One constant-success replicate. Fix an active set $\mathcal{V}_\ell$, let $d_\ell = \dim(\mathrm{span}(\mathcal{V}_\ell))$, and compute the design in Lemma C.4. For target accuracy ϵ_ℓ, allocate expected query budgets

$$q_{\ell,i}^{\mathrm{raw}} = \left\lceil c_\alpha u^{\frac{1}{2\alpha}} d_\ell^{\frac{1+\alpha}{4\alpha}} \epsilon_\ell^{-\frac{1+\alpha}{2\alpha}} \sqrt{\pi(a_i)} \right\rceil, \tag{28}$$

Run the support estimators independently. With $e_i = \widehat{\mu}(a_i) - \mu(a_i)$, predict every $a \in \mathcal{V}_\ell$ by

$$\widetilde{\mu}(a) = \sum_{i=1}^{s} \gamma_i(a) \widehat{\mu}(a_i). \tag{29}$$

Lemma C.5 (Fixed-action prediction)**.** *Before timeout, the constant c_α can be chosen so that, for every fixed $a \in \mathcal{V}_\ell$,*

$$\mathbb{E}|\widetilde{\mu}(a) - \mu(a)|^{1+\alpha} \le \frac{\epsilon_\ell^{1+\alpha}}{48}. \tag{30}$$

One replicate has expected query cost at most

$$U_\ell^{\mathrm{raw}} = \widetilde{O}_\alpha\left(u^{\frac{1}{2\alpha}} d_\ell^{\frac{1+3\alpha}{4\alpha}} \epsilon_\ell^{-\frac{1+\alpha}{2\alpha}} + d_\ell \right), \tag{31}$$

Proof. The support errors are independent and mean zero, so the von Bahr–Esseen inequality von Bahr & Esseen (1965) and equation 24 give

$$\mathbb{E}\left|\sum_i \gamma_i(a)e_i\right|^{1+\alpha} \le C_\alpha u \sum_i |\gamma_i(a)|^{1+\alpha}(q_{\ell,i}^{\mathrm{raw}})^{-2\alpha}.$$

Substituting equation 28 and $\gamma_i(a) = \pi(a_i)a_i^\top M^+ a$ yields

$$\begin{aligned}\mathbb{E}|\widetilde{\mu}(a) - \mu(a)|^{1+\alpha} &\le C_\alpha c_\alpha^{-2\alpha}\epsilon_\ell^{1+\alpha} d_\ell^{-\frac{1+\alpha}{2}} \sum_i \pi(a_i)|a_i^\top M^+ a|^{1+\alpha} \\ &\le C_\alpha c_\alpha^{-2\alpha}\epsilon_\ell^{1+\alpha} d_\ell^{-\frac{1+\alpha}{2}} \left(\sum_i \pi(a_i)|a_i^\top M^+ a|^2\right)^{\frac{1+\alpha}{2}} \\ &\le C_\alpha c_\alpha^{-2\alpha} c_1^{\frac{1+\alpha}{2}} \epsilon_\ell^{1+\alpha}.\end{aligned}$$

The second line uses monotonicity of L_p norms under the probability weights π, since $1+\alpha \le 2$. Taking c_α sufficiently large proves equation 30. Finally, $\sum_i \sqrt{\pi(a_i)} \le \sqrt{s} \le \sqrt{c_0 d_\ell}$ gives equation 31. □

Timeout and simultaneous accuracy. The unbiased raw-moment estimator has only an expected query guarantee. In each replicate, interrupt all remaining calls once the total query count reaches $6U_\ell^{\mathrm{raw}}$ and mark that replicate as failed. Set all predictions from a timed-out replicate to zero, so the subsequent median is always well defined. Markov's inequality bounds the timeout probability by $1/6$. On an uninterrupted replicate, equation 30 and Markov's inequality give a fixed-action error probability at most $1/6$ after adjusting constants. Thus every fixed-action prediction is accurate to ϵ_ℓ with probability at least $2/3$, including timeout.

For each fixed action, a Chernoff bound makes its median fail with probability at most $\delta/[2n(\ell+1)^2]$. A union bound over actions and phases gives

$$\Pr(|\widehat{\mu}_\ell(a) - \mu(a)| \le \epsilon_\ell \text{ for every completed phase and active } a) \ge 1 - \delta. \tag{32}$$

Theorem C.6 (Finite-action quantum linear bandits)**.** *For a finite action set $\mathcal{A} \subseteq \mathbb{R}^d$ with $n = |\mathcal{A}|$, Algorithm 2 with $\delta = T^{-2}$ satisfies, under the raw-moment assumption,*

$$\mathsf{Reg}_T = \widetilde{O}_\alpha\Big(u^{\frac{1}{1+\alpha}} d^{\frac{1+3\alpha}{2(1+\alpha)}} (\log(nT))^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}} + d\log T \log(nT)\Big). \tag{33}$$

The tilde hides scalar-estimator and phase logarithms, but not the displayed dependence on n.

Proof. Condition on equation 32. The induction in the proof of Theorem 3.2 shows that $a^\star$ is never eliminated and that every $a \in \mathcal{V}_\ell$ has gap at most $8\epsilon_\ell$. In particular, all queried support actions have this gap. From equation 31 and $H_\ell = O(\log(nT))$, the accuracy-dependent part of the raw phase cap is

$$\widetilde{O}_\alpha\Big(u^{\frac{1}{2\alpha}} d^{\frac{1+3\alpha}{4\alpha}} \epsilon_\ell^{-\frac{1+\alpha}{2\alpha}} \log(nT)\Big).$$

If L is the last completed phase, the budget test and geometric growth give

$$\epsilon_L = \widetilde{O}_\alpha\Big(u^{\frac{1}{1+\alpha}} d^{\frac{1+3\alpha}{2(1+\alpha)}} (\log(nT))^{\frac{2\alpha}{1+\alpha}} T^{-\frac{2\alpha}{1+\alpha}}\Big).$$

Both terminal exploitation and the geometric sum of exploration regret are at most the first term of equation 33. The ceilings in the $O(d)$ support allocations, repeated $O(\log(nT))$ times over $O(\log T)$ phases, give the additive term. The failure event contributes $2T\delta$. □

C.3 General action sets

Let $\mathcal{A} \subseteq \{a \in \mathbb{R}^d : \|a\|_2 \le 1\}$ be compact. For an active set $\mathcal{V}_\ell$ of rank d_ℓ, compute the design in Lemma C.4 and estimate every support mean to accuracy

$$\rho_\ell = \frac{\epsilon_\ell}{\sqrt{c_1 d_\ell}} \tag{34}$$

with failure probability $\delta/[2|S_\ell|(\ell+1)^2]$. Use QTME on every support action. On the resulting event, Cauchy–Schwarz gives, simultaneously for all $a \in \mathcal{V}_\ell$,

$$|\widetilde{\mu}_\ell(a) - \mu(a)| \le \left(\sum_{a_i \in S_\ell} \pi_\ell(a_i) |a_i^\top M_\ell^+ a|^2\right)^{1/2} \left(\sum_{a_i \in S_\ell} \pi_\ell(a_i) |e_i|^2\right)^{1/2}$$
$$\le \sqrt{c_1 d_\ell}\, \rho_\ell = \epsilon_\ell. \tag{35}$$

Thus the probability union bound is only over $O(d_\ell)$ support actions, never over the possibly uncountable active set.

Up to the displayed scalar confidence logarithms in equation 17, the phase-query cost is

$$Q_\ell^{\mathrm{raw}} = \widetilde{O}_\alpha\left(u^{\frac{1}{2\alpha}} d_\ell^{1+\frac{1+\alpha}{4\alpha}} \epsilon_\ell^{-\frac{1+\alpha}{2\alpha}} + d_\ell\right), \tag{36}$$

Theorem C.7 (General-action quantum linear bandits)**.** *Assume access to a small-support approximate G-design and a linear-optimization oracle on every active set. For every compact action set, under the raw-moment assumption Algorithm 3 with $\delta = T^{-2}$ satisfies*

$$\mathsf{Reg}_T = \widetilde{O}_\alpha\left(u^{\frac{1}{1+\alpha}} d^{\frac{1+5\alpha}{2(1+\alpha)}} T^{\frac{1-\alpha}{1+\alpha}} + d\log T\right). \tag{37}$$

The bound is independent of $|\mathcal{A}|$ and of every covering number.

Proof. The support-level failure probabilities are summable over actions and phases, so equation 35 holds in every completed phase with probability at least $1-\delta$. On this event, the same elimination induction as in Theorem 3.2 preserves $a^\star$ and bounds every active gap in phase ℓ by $8\epsilon_\ell$. For the raw model, the last-phase budget relation from equation 36 is

$$T \asymp u^{\frac{1}{2\alpha}} d^{1+\frac{1+\alpha}{4\alpha}} \epsilon_L^{-\frac{1+\alpha}{2\alpha}},$$

up to polylogarithmic factors. Therefore

$$T\epsilon_L = \widetilde{O}_\alpha\left(u^{\frac{1}{1+\alpha}} d^{\frac{1+5\alpha}{2(1+\alpha)}} T^{\frac{1-\alpha}{1+\alpha}}\right).$$

This bounds terminal exploitation; multiplying every phase cap by its $O(\epsilon_\ell)$ active gap and summing the geometric series gives the same order for exploration. Support rounding over $O(\log T)$ phases gives $d\log T$ up to confidence logarithms, and the failure event contributes $2T\delta$. □

D Extension to Bounded Central Moments

This appendix collects the complete extension to the translation-invariant central-moment model. It is deliberately separated from the main paper so that the main notation and algorithms use only the raw-moment scale u.

D.1 Central-moment model

For every action a, assume

$$\mathbb{E}|Y_a - \mu(a)|^{1+\alpha} \le v, \qquad \mu(a) = \mathbb{E}Y_a, \tag{38}$$

where $v > 0$ is known but $\mu(a)$ is not. Define

$$\mathcal{H}_\alpha^{\mathrm{cen}}(v) := \left\{\nu : \ |\mathbb{E}_\nu Y| \le 1,\ \mathbb{E}_\nu |Y - \mathbb{E}_\nu Y|^{1+\alpha} \le v\right\}. \tag{39}$$

Unlike the raw condition, equation 38 is invariant under translations. It controls deviations about an unknown center and therefore does not justify truncating Y_a about zero. A central-moment estimator must first locate $\mu(a)$ coarsely and then estimate a truncated residual.

For reference, Jensen's inequality and $|x-y|^{1+\alpha} \le 2^\alpha(|x|^{1+\alpha} + |y|^{1+\alpha})$ give

$$\mathbb{E}|Y_a|^{1+\alpha} \le u \quad \Longrightarrow \quad \mathbb{E}|Y_a - \mu(a)|^{1+\alpha} \le 2^{1+\alpha} u. \tag{40}$$

Conversely, when $|\mu(a)| \le 1$,

$$\mathbb{E}|Y_a - \mu(a)|^{1+\alpha} \le v \quad \Longrightarrow \quad \mathbb{E}|Y_a|^{1+\alpha} \le 2^\alpha(v+1). \tag{41}$$

These implications compare the model classes; throughout this appendix the direct central results use v rather than silently replacing it by u.

D.2 CENTRAL-MOMENT QUANTUM MEAN ESTIMATION

The high-probability estimator of Luo et al. (2026) uses a center–truncate–estimate procedure. The initial center is essential because $\mu(a)$ is unknown.

Algorithm 4 Scalar central-moment QHTME (specialization of Luo et al. (2026))

Input: oracle $\mathcal{O}_a$, v, α, accuracy ρ, and failure probability δ.

1. Set $\bar{\rho} \leftarrow \min\{\rho, v^{1/(1+\alpha)}\}$ and use $N_{\mathrm{c}} = C_\alpha \lceil \log(3/\delta) \rceil$ measured oracle samples in CHTME to obtain a rough center ζ satisfying $|\zeta - \mu(a)| \le C_\alpha v^{1/(1+\alpha)}$ with probability at least $1 - \delta/3$.
2. Define $Z = (Y_a - \zeta)\mathbf{1}\{|Y_a - \zeta| \le B\}$ with
$$B = \left(\frac{3 \cdot 2^{1+\alpha} v}{\bar{\rho}}\right)^{1/\alpha}, \qquad L = 2^{1+\alpha} v B^{1-\alpha}.$$
3. Apply the bounded-variance quantum mean estimator to Z with budget
$$N_{\mathrm{q}} = \left\lceil \max\left\{ \log\frac{3}{\delta}, \frac{3\sqrt{L \log(3/\delta)}}{2\bar{\rho}} \right\} \right\rceil$$
and failure probability $\delta/3$, obtaining $\widehat{\mu}_Z$.
4. Output $\widehat{\mu}(a) = \zeta + \widehat{\mu}_Z$.

Lemma D.1 (Central-moment QHTME Luo et al. (2026))**.** *Under equation 38, for every $\rho > 0$, Algorithm 4 satisfies*
$$\Pr(|\widehat{\mu}(a) - \mu(a)| > \rho) \le \delta$$
and has deterministic query bound
$$Q_{\mathrm{cen}}(\rho, \delta; v) = \widetilde{O}_\alpha\left(v^{\frac{1}{2\alpha}} \rho^{-\frac{1+\alpha}{2\alpha}} \sqrt{\log\frac{1}{\delta}} + \log\frac{1}{\delta} \right). \tag{42}$$
For $\rho > v^{1/(1+\alpha)}$, running the estimator at fallback accuracy $v^{1/(1+\alpha)} \le \rho$ gives the same displayed form. The tilde hides logarithms in the accuracy ratio but not the principal confidence factors.

Remark D.2 (CHTME is only the pilot)**.** *CHTME is a classical robust mean estimator applied to measured oracle samples. It locates the unknown center within the noise scale. The accuracy-dependent quantum speedup comes from estimating the truncated residual in Steps 2–3. The pilot can be cached after the required confidence has been reached.*

For finite-action linear bandits we need an unbiased estimator rather than a high-probability scalar estimate.

Lemma D.3 (Unbiased central-moment QUHTME Luo et al. (2026))**.** *Under equation 38, for every $\rho \in (0, v^{1/(1+\alpha)}]$, QUHTME returns an estimate satisfying*
$$\mathbb{E}\widehat{\mu}(a) = \mu(a), \qquad \mathbb{E}|\widehat{\mu}(a) - \mu(a)|^{1+\alpha} \le \rho^{1+\alpha}, \tag{43}$$
using
$$\widetilde{O}_\alpha\left(v^{\frac{1}{2\alpha}} \rho^{-\frac{1+\alpha}{2\alpha}} \right) \tag{44}$$
queries in expectation. Equivalently, for an expected budget q, its mean-zero error e satisfies
$$\mathbb{E}e = 0, \qquad \mathbb{E}|e|^{1+\alpha} \le \widetilde{O}_\alpha(v q^{-2\alpha}). \tag{45}$$
The random query count is part of the MLMC construction. The centering and its CHTME-type pilot are internal; the algorithm never receives $\mu(a)$.

D.2.1 DERIVATION OF THE RAW UNBIASED INTERFACE

Under the raw assumption, equation 40 supplies central scale $2^{1+\alpha}u$. Applying Lemma D.3 with that scale and absorbing the α-dependent constant gives equation 22–equation 24. This proves Lemma C.3. The main paper therefore needs neither the unknown center nor the symbol v in its finite-action algorithm.

D.3 CENTRAL-MOMENT QUANTUM MULTI-ARMED BANDITS

For $P_r = (1-r)\delta_0 + r\delta_B$,

$$\mathbb{E}_{P_r}|Y - \mathbb{E}Y|^{1+\alpha} = B^{1+\alpha}\big[r(1-r)^{1+\alpha} + (1-r)r^{1+\alpha}\big]. \tag{46}$$

Set $B = (v/(2p))^{1/(1+\alpha)}$ and take $r \in \{p/2, p\}$. Since the bracket in equation 46 is at most $2r$, both rare-event distributions satisfy equation 38. Their mean separation is

$$\Delta = \Theta\Big(v^{\frac{1}{1+\alpha}} p^{\frac{\alpha}{1+\alpha}}\Big),$$

whereas their oracle separation remains $\Theta(\sqrt{p})$. Hence the stopped hybrid proof of Theorem B.2 gives

$$\mathbb{E}_{P_-} N = \Omega\Big(v^{\frac{1}{2\alpha}} \Delta^{-\frac{1+\alpha}{2\alpha}}\Big). \tag{47}$$

Theorem D.4 (Central-moment QMAB bounds)**.** *For $T \geq CK$,*

$$\inf_{\mathsf{Alg}} \sup_{\nu \in (\mathcal{H}_\alpha^{\mathrm{cen}}(v))^K} \mathsf{Reg}_T(\mathsf{Alg}, \nu) \geq cT \min\left\{1, v^{\frac{1}{1+\alpha}} \left(\frac{K}{T}\right)^{\frac{2\alpha}{1+\alpha}}\right\}.$$

When $v = \Theta(1)$, the lower bound also contains $\Omega(K \log(T/K))$. The central-moment version of Algorithm 1, obtained by setting

$$q_\ell = Q_{\mathrm{cen}}(\epsilon_\ell, \delta_\ell; v),$$

satisfies

$$\mathsf{Reg}_T = \widetilde{O}_\alpha\Big(v^{\frac{1}{1+\alpha}} K^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}} + K \log T\Big).$$

Proof. For the lower bound, use equation 47 in the one-sided bandit-to-testing reduction and balance its two branches exactly as in Theorem 3.1. The bounded-reward construction supplies the additive logarithmic term. For the upper bound, Lemma D.1 gives the same $\epsilon_\ell^{-(1+\alpha)/(2\alpha)}$ accuracy exponent as raw QTME. Repeating the elimination proof with Q_{cen} yields the displayed rate. The pilot and additive confidence term affect only logarithmic factors and the $K \log T$ rounding term. □

D.4 CENTRAL-MOMENT QUANTUM LINEAR BANDITS

D.4.1 LOWER BOUNDS

Use the actions in equation 18, the parameters in equation 19, and the central rare-event distributions above.

Theorem D.5 (Central-moment QLB lower bounds)**.** *For every $d \geq 2$ and $T \geq Cd$, there is a fixed action set with $|\mathcal{A}| = d$ such that every QLB policy suffers*

$$\mathsf{Reg}_T = \Omega\left(T \min\left\{1, v^{\frac{1}{1+\alpha}} \left(\frac{d}{T}\right)^{\frac{2\alpha}{1+\alpha}}\right\}\right).$$

The same statement holds on the compact infinite set $\mathcal{A}_{\mathrm{ray}}$ in equation 20. When $v = \Theta(1)$, both settings also admit the additive lower bound $\Omega(d \log(T/d))$.

Proof. The finite embedding reproduces the central-moment QMAB hard family with $K = d$. On the ray extension, assign reward B with probability tr_i at ta_i. Equation equation 46 bounds the central moment, and the endpoint dominates the oracle separation for every $t \in [0, 1]$. The proof of Theorem C.1 then applies with v in place of u. □

D.4.2 FINITE ACTION SETS

For the design in Lemma C.4, allocate

$$q_{\ell,i}^{\mathrm{cen}} = \left\lceil c_\alpha v^{\frac{1}{2\alpha}} d_\ell^{\frac{1+\alpha}{4\alpha}} \epsilon_\ell^{-\frac{1+\alpha}{2\alpha}} \sqrt{\pi(a_i)} \right\rceil. \tag{48}$$

Independent QUHTME estimates and the von Bahr–Esseen argument give the same fixed-action prediction guarantee as equation 30, with one-replicate expected cost

$$U_\ell^{\mathrm{cen}} = \widetilde{O}_\alpha\Big(v^{\frac{1}{2\alpha}} d_\ell^{\frac{1+3\alpha}{4\alpha}} \epsilon_\ell^{-\frac{1+\alpha}{2\alpha}} + d_\ell\Big). \tag{49}$$

Indeed, equation 45 yields

$$\mathbb{E}\left|\sum_i \gamma_i(a) e_i\right|^{1+\alpha} \le C_\alpha v \sum_i |\gamma_i(a)|^{1+\alpha} (q_{\ell,i}^{\mathrm{cen}})^{-2\alpha},$$

and substituting equation 48 reduces the right-hand side to $C_\alpha c_\alpha^{-2\alpha} \epsilon_\ell^{1+\alpha}$. Timeout at $6U_\ell^{\mathrm{cen}}$, followed by the same action-wise median amplification as Algorithm 2, gives simultaneous phase accuracy.

Theorem D.6 (Central-moment finite-action QLB upper bound)**.** *For $n = |\mathcal{A}|$, the preceding central-moment variant of Algorithm 2 satisfies*

$$\mathsf{Reg}_T = \widetilde{O}_\alpha\Big(v^{\frac{1}{1+\alpha}} d^{\frac{1+3\alpha}{2(1+\alpha)}} (\log(nT))^{\frac{2\alpha}{1+\alpha}} T^{\frac{1-\alpha}{1+\alpha}} + d \log T \log(nT)\Big). \tag{50}$$

Proof. Use equation 49 in place of equation 31 in the proof of Theorem C.6. Inverting the last-phase budget produces the displayed power of v, while all dimension, cardinality, and horizon exponents are unchanged. □

D.4.3 General compact action sets

Use QHTME at every design support point with accuracy $\rho_\ell = \epsilon_\ell / \sqrt{c_1 d_\ell}$. The support union bound and the deterministic reconstruction argument in equation 35 remain unchanged. The phase-query cost is

$$Q_\ell^{\mathrm{cen}} = \widetilde{O}_\alpha\Big(v^{\frac{1}{2\alpha}} d_\ell^{1+\frac{1+\alpha}{4\alpha}} \epsilon_\ell^{-\frac{1+\alpha}{2\alpha}} + d_\ell\Big). \tag{51}$$

Theorem D.7 (Central-moment general-action QLB upper bound)**.** *Under the design and linear-optimization oracle assumptions of Theorem C.7, the central-moment variant of Algorithm 3 satisfies*

$$\mathsf{Reg}_T = \widetilde{O}_\alpha\Big(v^{\frac{1}{1+\alpha}} d^{\frac{1+5\alpha}{2(1+\alpha)}} T^{\frac{1-\alpha}{1+\alpha}} + d \log T\Big). \tag{52}$$

The guarantee is independent of $|\mathcal{A}|$ and every covering number.

Proof. Use equation 51 in the last-phase calculation of Theorem C.7. Solving the budget relation for ϵ_L gives the displayed power of v and the same exponents of d and T. □

E Additional Experimental Details and Results

Experimental setting.

- (MAB) We consider stochastic K-armed bandits with $K \in \{5, 10\}$. The optimal mean is 0.6, while the suboptimality gaps are linearly spaced between 0.08 and 0.24; arm labels are randomly permuted for each (α, K). Rewards follow a location-shifted symmetric Pareto distribution. For $p = 1 + \alpha$, where $\alpha \in \{0.25, 0.5, 0.75, 1\}$, the Pareto shape is $p + 0.5$, and the scale is chosen such that $\mathbb{E}|X_a - \mu_a|^p = 0.25$. We compare successive-elimination algorithms using a truncated empirical mean (classical raw), median-of-means (classical central), and idealized quantum raw and central estimators. Each quantum estimate is sampled uniformly from $[\mu_a - \epsilon_\ell, \mu_a + \epsilon_\ell]$, with its theoretical query cost charged to the horizon. The horizon is 10^5 for $\alpha = 0.25$ and 10^4 otherwise. Each curve reports the mean cumulative pseudo-regret over 500 runs with pointwise 95% confidence intervals, where $R_t = \sum_a N_a(t)(\mu^\star - \mu_a)$.
- (Finite Linear Bandits.) We consider two finite-action linear bandit instances, each with $n = 128$ unit-norm actions, horizon $T = 256{,}000$, and the one-sparse parameter $\theta^\star = 0.6 e_1$, so that the expected reward of action a is $a^\top \theta^\star = 0.6 a_1$. Both action sets follow a

deterministic clustered geometry that contains the optimal action e_1 and a unique runner-up with gap 0.08. In the $d = 2$ instance, reported for $\alpha \in \{0.25, 0.5, 0.75, 1\}$ in Figure 3, the remaining gaps lie in $[0.12, 0.24]$. In the $d = 8$ instance, reported for $\alpha = 0.75$ in Figure 1(c), the action set additionally contains six anchor actions with gap 0.24 that ensure full rank, and 120 further actions with gaps in $[0.12, 0.24]$ whose orthogonal components are concentrated near e_2. In both instances, rewards are $Y_t = a_t^\top \theta^\star + \xi_t$, where ξ_t is symmetric Pareto noise with shape $1.5 + \alpha$, scaled such that $\mathbb{E}|\xi_t|^{1+\alpha} = 0.25$. Exploration allocations are computed from a G-optimal design. We compare classical raw-moment, classical central-moment, idealized quantum raw-moment, and idealized quantum central-moment estimators within the same phased-elimination framework. The quantum estimator is an unbiased bounded-uniform emulator whose $(1 + \alpha)$-absolute error moment follows the principal $q^{-2\alpha}$ rate. We use $\delta = 0.05$, set the nominal budget and median-amplification constants to one, and run 200 paired seeds per configuration. The curves report mean cumulative pseudo-regret, and the shaded regions are pointwise normal 95% confidence intervals given by the mean plus or minus 1.96 standard errors.

- (General Linear Bandits.) We evaluate general-action linear bandits on the continuous action set $\mathcal{A} = [-1/\sqrt{d}, 1/\sqrt{d}]^d$ with $d = 4$, horizon $T = 256{,}000$, and the hard parameter $\theta = 0.75(1, 1, 1, 0.05)/\|(1, 1, 1, 0.05)\|_2$, whose small fourth coordinate makes the optimal direction difficult to identify. Rewards follow $Y_t(a) = a^\top \theta + \xi_t$, where ξ_t is symmetric Pareto noise with shape $1 + \alpha + 0.5$, scaled so that $\mathbb{E}|\xi_t|^{1+\alpha} = 0.25$. We consider $\alpha \in \{0.25, 0.5, 0.75, 1\}$ and compare classical raw-moment, classical central-moment, quantum raw-moment, and quantum central-moment estimators within the same phased elimination algorithm. Each phase uses an approximate continuous G-optimal design with tolerance 0.1, satisfying the numerical certificate $G/d \leq 1.1$. Classical support means are estimated by median-of-means, whereas the quantum methods use an idealized confidence-range emulator with every query charged as one bandit round. Quantum raw employs the rate-normalized median-amplified construction with $C_{\mathrm{amp}} = 1$: independent constant-confidence estimates are perturbed within their certified ranges and aggregated by an odd median. Thus, these experiments compare normalized query-complexity rates rather than simulate a gate-level, finite-constant QTME implementation. Results are averaged over 200 independent seeds, and the shaded regions denote pointwise 95% Monte Carlo confidence intervals.

Results analysis.

- (MAB) At the displayed horizons, both idealized quantum methods achieve lower final mean cumulative pseudo-regret than their matched classical baselines in all eight settings in Figure 2. The reduction ranges from 65.1% to 98.2% for the raw-moment methods and from 87.3% to 97.3% for the central-moment methods. For the heaviest-tailed case, $\alpha = 0.25$, quantum raw attains regrets of $1{,}316$ and $2{,}605$ for $K = 5$ and $K = 10$, compared with $3{,}771$ and $8{,}459$ for classical raw; quantum central further reduces these values to 115 and 224. For $\alpha \geq 0.5$, the final quantum regrets range from 16 to 159, substantially below the corresponding classical values. Quantum central gives the lowest regret in seven of the eight panels, while quantum raw is marginally better for $\alpha = 1, K = 10$. Since the quantum estimates are sampled within the true error interval rather than generated by an explicit quantum circuit, these results should be interpreted as an estimator-rate comparison rather than hardware-level evidence.

- (Finite Linear Bandits.) In Figure 3, quantum central attains the lowest mean terminal regret in every panel. At $\alpha = 0.25$, the endpoint regrets are $8{,}853.606$, $1{,}882.770$, $5{,}568.095$, and 726.512 for classical raw, classical central, quantum raw, and quantum central, respectively. In the primary paired comparison, quantum central improves over classical central by $1{,}156.258$ regret units: the paired difference, quantum central minus classical central, is $-1{,}156.258$, with a 95% confidence interval of $[-1{,}221.294, -1{,}091.221]$. This corresponds to a 61.41% reduction in mean regret, with quantum central outperforming classical central on 94% of paired seeds. For $\alpha = 0.5, 0.75$, and 1, quantum-central regret is 159.638, 103.193, and 89.659, compared with $1{,}954.643$, 679.674, and 438.811 for classical central, yielding reductions of 91.83%, 84.82%, and 79.57%, respectively. Quantum raw also improves over classical raw for every α, although it remains worse than classical

central at $\alpha = 0.25$, indicating that the centered-moment assumption is particularly consequential in the heaviest-tailed regime. Thus, the results support the algorithmic comparison under the stated idealized emulator rather than a quantum-hardware claim.

- (General Linear Bandits.) In Figure 4, quantum raw achieves lower terminal cumulative pseudo-regret than both classical baselines for every value of α. In units of 10^3, its final regrets are 21.540, 18.341, 7.958, and 4.233 for $\alpha = 0.25, 0.5, 0.75$, and 1, respectively, compared with 39.635, 18.620, 9.758, and 14.746 for the better of the two classical methods. These correspond to reductions of approximately 45.7%, 1.5%, 18.4%, and 71.3%. Quantum central performs best overall, attaining terminal regrets of 7.749, 3.421, 1.308, and 0.746. The apparently small separation between quantum raw and classical raw at $\alpha = 0.5$ is statistically significant: the paired difference is 0.279×10^3, with 95% confidence interval $[0.233, 0.326] \times 10^3$. This proximity results from the atomic phase schedule rather than convergence to the same rate. Classical raw completes only two phases and then exploits the optimal action with zero additional regret, whereas quantum raw completes six substantially longer but lower-gap exploration phases. Consequently, the non-monotone improvement across α reflects finite-horizon phase-boundary effects, while the overall results consistently demonstrate the lower normalized query cost of the quantum estimators.

Figure 2: Multi-Armed Bandits under Pareto Distribution

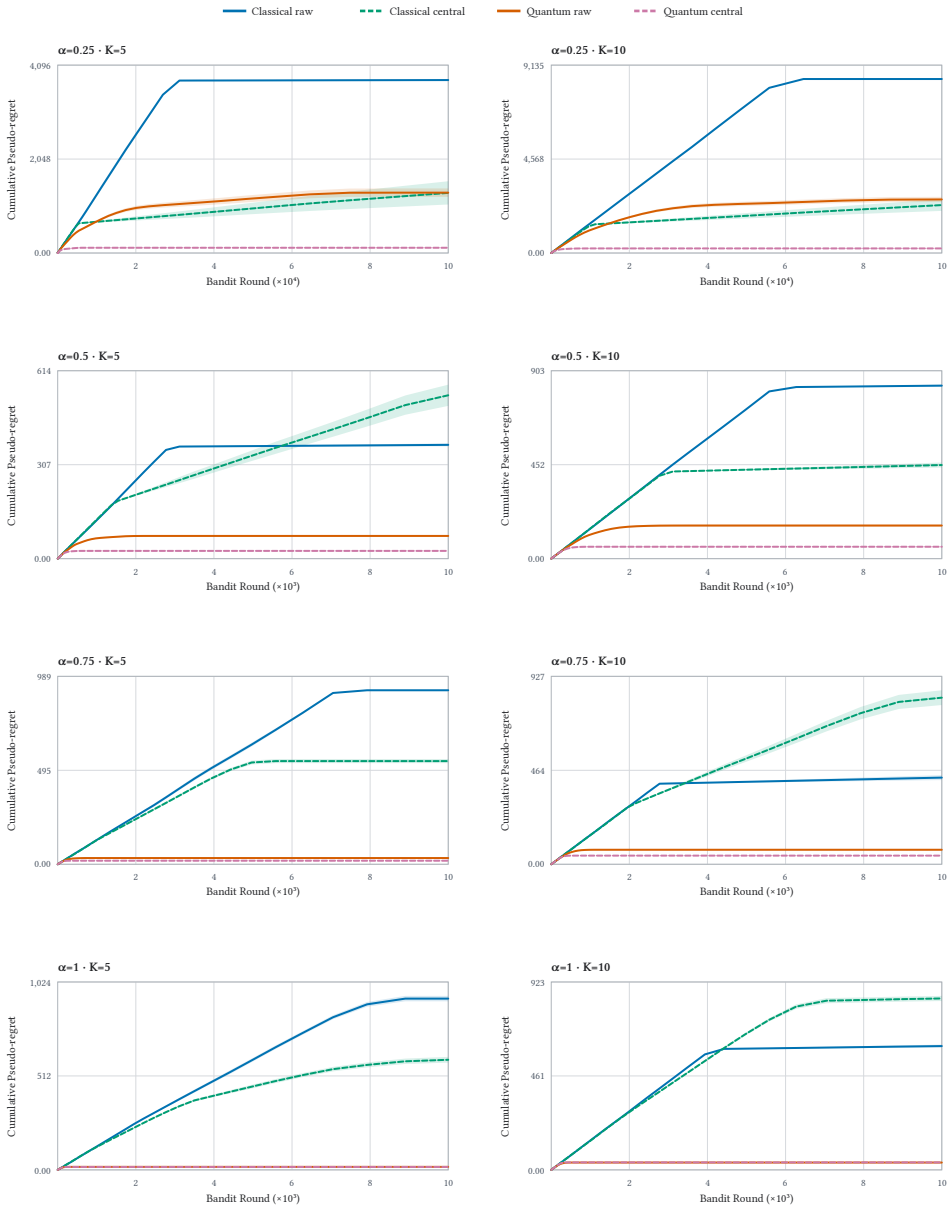


Figure 3: Linear Bandits with Finite Action under Pareto Distribution

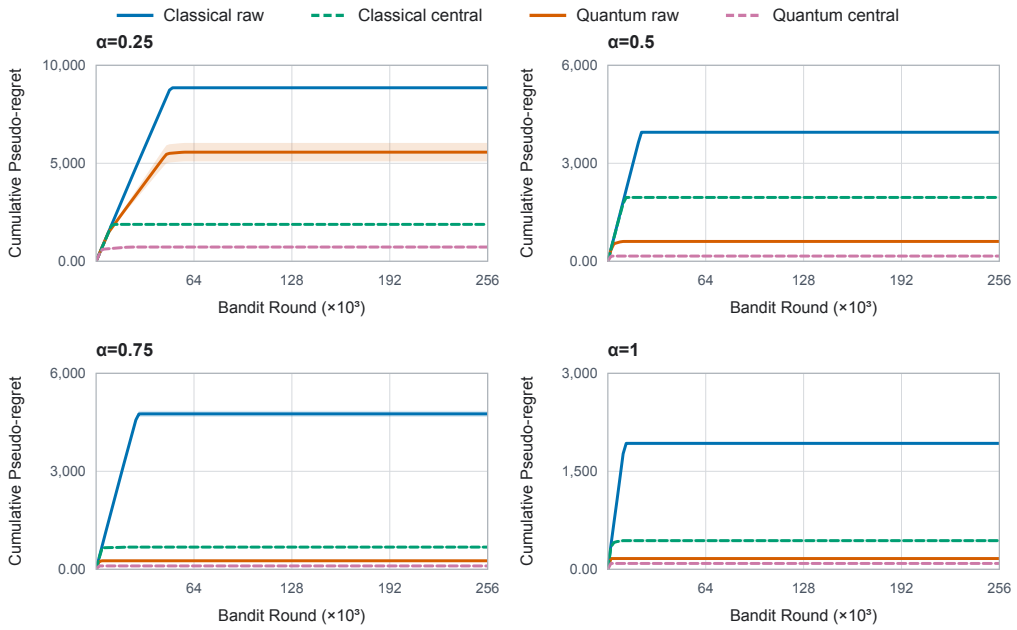


Figure 4: Linear Bandits with General Action under Pareto Distribution

