Safe Quantum Stochastic Linear Bandits

Ruizhe Zhang
Department of Computer Science
Purdue University
West Lafayette, IN 47907
rzzhang@purdue.edu

Junyi Wu
Department of Mathematics
Purdue University
West Lafayette, IN 47907
wu2393@purdue.edu

Guang Lin
Department of Mathematics
Purdue University
West Lafayette, IN 47907
guanglin@purdue.edu

Abstract

We initiate the study of safe quantum stochastic linear bandits, in which a learner maximizes an unknown linear reward while satisfying unknown linear mean constraints at every interaction. Under coherent unitary access to the joint reward-and-constraint distribution, we develop QRS-COLTS, a staged quantum resampled-scaling analog of constrained linear Thompson sampling (COLTS). With high probability, QRS-COLTS plays only feasible actions and achieves regret that is polylogarithmic in the query horizon and scales only with the square root of the number of constraints. We also study a model in which the learner queries superpositions of actions while an unsafe-action flag, a "bomb" that terminates the interaction upon violation, is measured after every query. In this setting, the bomb provides exact membership access to the safe set, and query regret is evaluated over the action superposition. For this model, we propose BCO-QLinTS, which selects feasible actions via bomb-certified quantum convex optimization without estimating the constraint matrix. Its query-weighted regret remains polylogarithmic in the horizon, with the constraints entering the bound only through the geometry of the safe set.

1 Introduction

Sequential decision-making systems are increasingly deployed in settings where exploration must respect hard operational constraints. In adaptive clinical trials, a treatment policy must avoid unsafe toxicity levels; in recommendation systems, decisions may be constrained by budget or fairness requirements; and in robotics or power-grid control, unsafe actions can drive the system outside a stability envelope. The same tension arises in quantum-native control and simulation. For example, consider tuning a microwave pulse to read out a superconducting qubit, where a stronger pulse can make the two logical states easier to distinguish in a short measurement, yet even a pulse that reports the correct bit can drive the transmon into a higher energy level, compromising subsequent operations [HDC$^+$25]. Starting from a conservative baseline pulse with a certified leakage margin, the learner seeks to maximize readout accuracy while keeping the leakage probability below a prescribed threshold at every step. When such controlled quantum dynamics are simulated coherently on a quantum computer, the reward and constraint signals are generated by a reversible quantum circuit, and we ask how much coherent access helps a learner that must remain safe.

We study this question through the lens of safe stochastic linear bandits [AAT19, KB20, MAAT21, GCS24, GS26]. At each interaction, the learner chooses a classical action $a_t \in \mathcal{A} \subseteq \mathbb{R}^d$ and receives a random feedback vector

$$X_t = (R_t, C_t) \in [0,1]^{1+m},$$

where R_t is the reward and $C_t = (C_{t,1}, \ldots, C_{t,m})$ consists of noisy measurements of m safety-related quantities, each normalized to lie in $[0,1]$. While the feedback distribution may depend on the

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

chosen action, its expectation is assumed to be linear:

$$\mathbb{E}[R_t \mid a_t = a] = \langle \theta^\star, a \rangle , \qquad \mathbb{E}[C_t \mid a_t = a] = \Phi^\star a,$$

for an unknown reward vector $\theta^\star \in \mathbb{R}^d$ and an unknown constraint matrix $\Phi^\star \in \mathbb{R}^{m \times d}$.

Safety is defined in terms of these expected constraint values. Specifically, an action $a \in \mathcal{A}$ is *feasible* if $\Phi^\star a \leq \alpha$, where the inequality holds coordinate-wise, i.e., $\langle \phi_i^\star, a \rangle \leq \alpha_i$ for every $i \in [m]$, with $\phi_i^\star$ denoting the i-th row of $\Phi^\star$. Although individual noisy realizations C_t need not satisfy the thresholds, the *round-wise mean-safety* requirement [AAT19, MAAT21, GS26] requires that, with high probability, every selected action is feasible in expectation: $\Phi^\star a_t \leq \alpha$ for all $t \leq T$. This mean-constraint formulation captures settings where safety is governed by expected performance, risk, or resource metrics, including efficacy–toxicity tradeoffs in dose-finding trials [TC04], expected resource consumption [BKS18], fairness of exposure [WBSJ21], and stability under uncertain dynamics [BGH$^+$22]. It also encompasses binary failure indicators whose expectation bounds the probability of an unsafe event.

Performance is measured by cumulative regret relative to the best feasible action:

$$a^\star \in \operatorname*{arg\,max}_{a \in \mathcal{A}:\, \Phi^\star a \leq \alpha} \langle \theta^\star, a \rangle , \qquad \mathrm{Reg}(T) = \sum_{t=1}^{T} \langle \theta^\star, a^\star - a_t \rangle .$$

Following standard practice in safe bandits with round-wise constraints, we assume access to a known strictly safe baseline action $a_{\mathrm{safe}} \in \mathcal{A}$ with a known lower bound $\Gamma_0 > 0$ on its safety slack:

$$\min_{i \in [m]} \{\alpha_i - \langle \phi_i^\star, a_{\mathrm{safe}} \rangle\} \geq \Gamma_0.$$

This baseline provides the learner with a certified fallback action while it explores and expands its estimate of the unknown feasible region.

Classical algorithms for safe linear bandits, including confidence-set methods, conservative and stage-wise exploration, Thompson-sampling variants, and local-feedback or polytope models [AAT19, KB20, MTA20, MAAT21, LLSY21, VLA23, HTA24, GCS24, GS26], are fundamentally bottlenecked by classical Monte Carlo estimation. Even recent computationally efficient methods such as Constrained Linear Thompson Sampling (COLTS) [GS26], which replaces expensive robust subproblems with coupled perturbations and constrained optimization, observe only a single noisy sample of the reward and constraints per round. Because unconstrained stochastic linear bandits already face an $\Omega(d\sqrt{T})$ minimax regret lower bound for $T \gtrsim d^2$ [DHK08, BCB12, AYPS11], classical safe algorithms necessarily inherit this $\Omega(\sqrt{T})$ statistical barrier; for instance, the zero-violation variant S-COLTS has a leading regret term of $\widetilde{O}(\sqrt{d^3 T})$ for fixed baseline gap and slack [GS26].

In unconstrained settings, recent quantum bandit algorithms have shown that coherent oracle access can break this $\sqrt{T}$ barrier. Under quantum reward oracles, multi-armed and stochastic linear bandits admit polylogarithmic regret in the query horizon T [WZL$^+$23], with related quantum speedups established for convex bandits, Bayesian optimization and kernelized bandits, heavy-tailed bandits, and Lipschitz bandits [LZ22, DLV$^+$23, HMTU24, WGAW23, YKL26]. However, these unconstrained methods do not address the core challenge of safe exploration: certifying round-wise feasibility under unknown constraints. Doing so requires simultaneously handling reward uncertainty optimistically and constraint uncertainty conservatively across $(1+m)$-dimensional vector feedback. This motivates the central question of our work:

Can quantum access break the classical $\sqrt{T}$ regret barrier for safe linear bandits while keeping every selected action safe?

1.1 Main results

We make two main contributions. First, at the conceptual level, we initiate the study of safe quantum linear bandits and formulate three quantum feedback models with round-wise safety constraints: (i) a classical-action, coherent-feedback model; (ii) a classical-action, classical-feedback model, which serves as a baseline and admits no regret speedup; and (iii) a coherent-action bomb model evaluated under query-weighted regret. Second, at the algorithmic level, we establish positive results in the first and third models, showing that coherent quantum access enables polylogarithmic regret in the horizon T while preserving the safety guarantee of each model.

Safe quantum bandit models. We consider three quantum feedback models in this paper (Figure 1). Across all three models, the benchmark is the optimal feasible action satisfying the mean constraints $\Phi^\star a \le \alpha$, while the models differ in how the learner queries the feedback distribution and the safe set.

The first model extends the coherent reward oracle of [WZL+23] to vector-valued reward and constraint feedback. Intuitively, after selecting a classical action a, the learner can call a unitary quantum oracle $\mathcal{O}_a$ that prepares a coherent quantum sample ("qsample") of the joint feedback $X(a)$, as defined below.

Definition 1 (Classical-action, coherent-feedback model)**.** *To evaluate a classical action $a \in \mathcal{A}$, the quantum learner receives coherent unitary access to the feedback sampler for that action via a quantum oracle $\mathcal{O}_a$ and its inverse $\mathcal{O}_a^\dagger$:*

$$\mathcal{O}_a |0\rangle |0\rangle = \sum_\omega \sqrt{p_a(\omega)}\, |\omega\rangle\, |X(a,\omega)\rangle\,,$$

where the first register holds the coherent random seed ω and the second register encodes the feedback vector satisfying

$$\sum_\omega p_a(\omega) X(a,\omega) = \mathbb{E}[X_t \mid a_t = a] = (\langle \theta^\star, a\rangle\,, \Phi^\star a)\,.$$

Each call to $\mathcal{O}_a$ or $\mathcal{O}_a^\dagger$ counts as one of the T queries and incurs regret $\langle \theta^\star, a^\star - a\rangle$. Safety requires every queried action to satisfy $\Phi^\star a \le \alpha$ with high probability.

To instantiate Definition 1 in the qubit readout setting, let $a \in \Delta^{d-1}$ be a probability vector over d candidate readout circuits, and assume controlled access to U_j and $U_j^\dagger$, where $U_j |0\rangle = \sum_y \sqrt{p_j(y)}\, |y\rangle$ for $j \in [d]$. Here, an outcome $y = (b, \hat{b}, \ell)$ records the prepared logical bit b, the reported bit $\hat{b}$, and the post-readout energy level ℓ, with reward and leakage indicators $h_j(y) = (\mathbf{1}\{\hat{b} = b\},\ \mathbf{1}\{\ell \ge 2\})$. By preparing the superposition $\sum_{j=1}^d \sqrt{a_j}\, |j\rangle$, applying the j-th circuit conditioned on $|j\rangle$, and computing $h_j(y)$ into the feedback register, we obtain the oracle

$$\mathcal{O}_a |0\rangle |0\rangle = \sum_{j,y} \sqrt{a_j p_j(y)}\, |j, y\rangle\, |h_j(y)\rangle\,,$$

where each call to this joint circuit or its inverse counts as a single quantum query. Letting $v_j = \mathbb{E}_{y\sim p_j}[h_j(y)] \in [0,1]^2$, the expected feedback under policy a is $\sum_{j=1}^d a_j v_j$, which is linear in a, and any mixture of circuits with certified leakage margins provides a baseline action a_{safe}. Finally, rescaling both the scores and the leakage threshold by $1/\sqrt{d}$ ensures that the unknown parameter vectors have at most unit ℓ_2-norm, so regret in the original units is simply $\sqrt{d}$ times the normalized regret.

The second model provides only incoherent access to the feedback, as in the classical bandit setting, with quantum processing allowed only between queries.

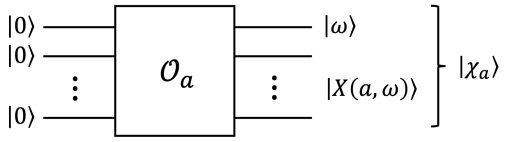
Definition 2 (Classical-action, classical-feedback model)**.** *At each round, the learner chooses a classical action $a \in \mathcal{A}$ and observes a single classical feedback sample $X(a,\omega)$ drawn from $\omega \sim p_a$. Equivalently, the query applies a quantum channel $\mathcal{E}_a$ followed by an immediate standard-basis measurement:*

$$\rho_0 \mapsto \mathcal{E}_a(\rho_0), \quad \rho_0 = |0\rangle\langle 0|, \qquad \textit{followed by standard-basis measurement}\,.$$

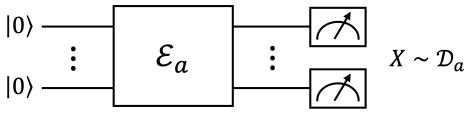
Because the measurement collapses the output state, the learner does not have reversible coherent access to the feedback randomness. This model subsumes the standard classical safe stochastic linear bandit setting.

Setting all constraint observations to zero and taking positive thresholds embeds the classical stochastic linear bandit in Definition 2. Its $\Omega(d\sqrt{T})$ minimax regret lower bound therefore applies. For best-arm identification, Buchholz et al. [BKS25] prove a classical-order query lower bound for their one-time stochastic channel oracle.

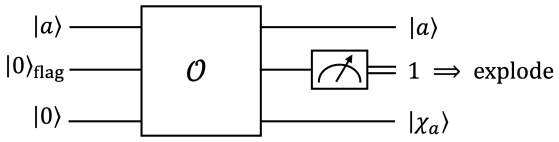
The third model additionally allows coherent superposition queries over actions, analogous to the quantum stochastic convex bandit model of [LZ22]. To enforce safety across action superpositions,



(a) Classical-action, coherent-feedback model



(b) Classical-action, classical-feedback model



(c) Coherent-action, coherent-feedback model with a safe-action bomb

Figure 1: Safe quantum bandit models. In Figure 1a, each action a has a corresponding unitary oracle $\mathcal{O}_a$ that prepares a coherent qsample of the stochastic feedback. In Figure 1b, each query applies a quantum channel $\mathcal{E}_a$ followed by measurement, yielding classical feedback. In Figure 1c, the learner queries actions in superposition (first register) while the oracle attaches a safety flag (second register) and prepares the reward qsample (third register); measuring 1 on the flag register triggers a bomb explosion and terminates the algorithm.

we adopt the bomb-query mechanism of [LL15], inspired by the Elitzur–Vaidman bomb-testing problem [EV93]. Both query directions compute the exact unsafe-action bit into a fresh flag. Measuring the flag terminates the algorithm on outcome 1. To survive, the learner must keep the query amplitude on unsafe actions sufficiently small throughout execution.

Definition 3 (Coherent-action, coherent-feedback model with a safe-action bomb)**.** *The quantum learner may query a superposition over actions. For each action $a \in \mathcal{A}$, let*

$$|\chi_a\rangle_X := \sum_{\omega \in \Omega_a} \sqrt{p_a(\omega)}\ |\omega\rangle\,|R(a,\omega)\rangle$$

denote a coherent qsample of the reward distribution. On a forward query with blank output registers, the oracle coherently attaches an unsafe-action flag and the reward qsample:

$$|a\rangle\,|0\rangle_f\,|0\rangle_X \mapsto |a\rangle\,|f(a)\rangle_f\,|\chi_a\rangle_X\,, \qquad f(a) = \mathbf{1}\{\Phi^\star a \not\leq \alpha\}.$$

Immediately after each forward or inverse query, the flag register $|\cdot\rangle_f$ is measured; if the outcome is 1, a bomb explosion occurs and the algorithm terminates. Following [LL15], we require the cumulative explosion probability over all T queries to be at most ε. Because a query may place amplitude across multiple actions, its regret is evaluated with respect to the action distribution of the query state (Definition 7).

Quantum algorithms. We establish positive results for the first and third models. Our first algorithm is QRS-COLTS (Quantum Resampled Scaling COLTS), a staged quantum analog of S-COLTS [GS26] in the classical-action, coherent-feedback model. QRS-COLTS preserves the perturb-and-scale structure of COLTS, including coupled Thompson perturbations, perturbed subproblems, and conservative scaling toward a known safe baseline, while replacing classical sample averaging within each stage with quantum mean estimation.

Theorem 1 (Informal main theorem: QRS-COLTS)**.** *Let $\mathcal{A} \subseteq \mathbb{R}^d$ be compact and convex with $\|a\|_2 \leq 1$ for all $a \in \mathcal{A}$. Assume the unknown reward vector and constraint rows satisfy $\|\theta^\star\|_2 \leq 1$ and $\|\phi_i^\star\|_2 \leq 1$ for each $i \in [m]$. Let*

$$a^\star \in \operatorname*{arg\,max}_{a \in \mathcal{A}:\, \Phi^\star a \leq \alpha} \langle \theta^\star, a \rangle$$

be an optimal safe action, and assume access to a known strictly safe action $a_{\mathrm{safe}} \in \mathcal{A}$ with safety slack at least $\Gamma_0 > 0$. In the classical-action, coherent-feedback model, QRS-COLTS with bounded multivariate quantum mean estimation satisfies, with probability at least $1 - 3\delta$:

1. Round-wise mean safety: *every queried action is feasible, $\Phi^\star a_t \leq \alpha$ for all $t \leq T$.*

2. Quantum regret bound:

$$\mathrm{Reg}(T) = \widetilde{O}\left(\sqrt{m+1}\, d^2 \log^{3/2}(T/\delta)\left(1 + \frac{\Delta(a_{\mathrm{safe}})}{\Gamma_0}\right)\right),$$

where $\Delta(a_{\mathrm{safe}}) = \langle \theta^\star, a^\star - a_{\mathrm{safe}} \rangle$ *is the reward suboptimality gap of the safe baseline.*

Here $\widetilde{O}(\cdot)$ *suppresses additional logarithmic factors from quantum mean estimation and resampling; see Theorem 5 and Corollary 3 for the formal statement and proof.*

The safety guarantee holds uniformly over the entire horizon: a single high-probability event of measure at least $1 - 3\delta$, encompassing all quantum mean-estimation, confidence-ellipsoid, and perturbation concentration events, guarantees $\Phi^\star a_t \leq \alpha$ for all $t \leq T$. Because δ enters the regret bound only polylogarithmically, the failure probability can be made negligibly small at minimal regret cost. We note that this guarantee applies to expected constraint values; see Section 5 for further discussion of its scope.

The classical cost of QRS-COLTS depends on the optimization subproblems. Across $S = \widetilde{O}(d \log T)$ stages, stage s solves at most $K_s = O(\log(s/\delta))$ linear optimization subproblems over $\mathcal{A}$ intersected with m sampled halfspaces (which reduce to linear programs whenever $\mathcal{A}$ is a polytope). Each stage also performs a rank-one update to the upper-triangular Cholesky factor U_s of the $d \times d$ design matrix $V_s = U_s^\top U_s$ to generate the coupled reward and constraint perturbations, and recomputes the $m+1$ ridge regression estimates in $O((m+1)d^2)$ operations. If each linear optimization subproblem takes $\mathcal{C}_{\mathrm{opt}}$ operations, the classical optimization and regression cost is

$$\widetilde{O}\big(d \log T \ \big(\mathcal{C}_{\mathrm{opt}} + (m+1)d^2\big)\big).$$

Consequently, quantum query savings translate into wall-clock speedups primarily when calls to the coherent quantum simulator dominate these classical optimization and regression updates. Achieving this statistical speedup requires more than simply plugging a quantum mean estimator into COLTS, as the stage-wise quantum estimates must be carefully synchronized with the unsaturation and conservative scaling arguments (Section 4).

Our second algorithmic result addresses the coherent-action bomb model. In this setting, the bomb flag directly certifies whether an action is safe, albeit at the risk of terminating the algorithm. This allows the learner to treat the flag as an exact membership oracle and optimize directly over the true safe set without ever estimating the constraint matrix $\Phi^\star$. We realize this approach in BCO-QLinTS (Bomb-Certified Optimistic Quantum Linear Thompson Sampling), which in each stage maximizes K_s perturbed linear reward objectives over the safe set via bomb-certified quantum convex optimization and estimates the reward at the most optimistic candidate action.

Theorem 2 (Informal main theorem: BCO-QLinTS)**.** *Consider the coherent-action, coherent-feedback model with a safe-action bomb under query-weighted regret. Assume that* $\|a\|_2 \leq 1$ *for all* $a \in \mathcal{A}$, $\|\theta^\star\|_2 \leq 1$, *and the true safe set* $\mathcal{S} = \{a \in \mathcal{A} : \Phi^\star a \leq \alpha\}$ *satisfies Assumption 2, with known* a_{safe}, r, R *such that* $\mathbb{B}(a_{\mathrm{safe}}, r) \subseteq \mathcal{S} \subseteq \mathbb{B}(a_{\mathrm{safe}}, R)$. *Then BCO-QLinTS has total explosion probability at most* ε *and performs reward estimation only at certified feasible actions. With probability at least* $1 - \delta$, *its query-weighted regret up to and including the last query satisfies*

$$R_{\mathrm{w}}(T) \leq \widetilde{O}\left(d^2 \log^{3/2}(2T/\delta) + \frac{d^4 \log^2(2T)}{\varepsilon}\right),$$

where $\widetilde{O}(\cdot)$ *suppresses logarithmic factors in* R/r *and in the inverse optimization tolerance* T^2*; see Theorem 8 for the formal statement and proof.*

In Theorem 2, the first term bounds the regret incurred during reward estimation, while the second term captures the bomb-query overhead of optimizing over $\mathcal{S}$. Specifically, BCO-QLinTS runs for at most $S = S_{\max} = \widetilde{O}(d \log T)$ stages and divides its per-stage explosion budget ε/S equally across $K_s = O(\log(s/\delta))$ convex optimization calls, each requiring $Q = \widetilde{O}(d)$ coherent membership queries. Simulating these calls via the Lin–Lin bomb-query framework [LL15] consumes $\widetilde{O}(Q^2 K_s^2 S/\varepsilon)$ bomb queries in stage s, which sums to $\widetilde{O}(d^2 S^2/\varepsilon) = \widetilde{O}(d^4 \log^2 T/\varepsilon)$ over all stages. Crucially, the number of constraints m does not appear in the regret bound: because the bomb flag aggregates feasibility across all m constraints into a single bit, the constraints enter only logarithmically through the geometric ratio R/r.

Table 1: Comparison of linear bandit algorithms that select one classical action per query. Here $\Delta = \Delta(a_{\mathrm{safe}})$ and $\Gamma = \min_i(\alpha_i - \langle \phi_i^\star, a_{\mathrm{safe}} \rangle) \geq \Gamma_0$.

Algorithm	Unknowns	Feedback per query	Action selection and safety	Regret
QLinUCB [WZL$^+$23]	$\theta^\star$	coherent reward sampler	optimistic maximization; no constraints	$\widetilde{O}(d^2 \log^{3/2} T)$
S-COLTS [GS26]	$\theta^\star, \Phi^\star$	one classical sample of (R, C)	perturbed subproblem, then scaling toward a_{safe}	$\widetilde{O}((1+\Delta/\Gamma)\sqrt{d^3 T})$ $+ \widetilde{O}(d^2 \Delta/\Gamma^2)$
QRS-COLTS (Theorem 1)	$\theta^\star, \Phi^\star$	coherent sampler of (R, C)	K_s perturbed subproblems, then scaling toward a_{safe}	$\widetilde{O}(\sqrt{m+1}\, d^2 \log^{3/2} T)$ $\times(1+\Delta/\Gamma_0)$

2 Related work

Classical safe and constrained linear bandits. Classical approaches to safe linear bandits with unknown linear constraints broadly fall into two families: pessimistic-optimistic confidence-set methods [AAT19, KB20, MTA20, LLSY21, VLA23, HTA24, GCS24] and randomized Thompson-sampling methods [MAAT21, GS26], with related approximate Thompson sampling techniques via Langevin Monte Carlo studied in [ZDML24, WZL$^+$25]. The closest predecessor, S-COLTS [GS26], uses coupled perturbations and constrained linear optimization, then scales toward a_{safe} to ensure round-wise safety. Our first algorithm, QRS-COLTS, retains this perturb-and-scale architecture while replacing classical sample averaging within each stage by multivariate quantum mean estimation (Table 1).

Quantum bandits and query models. Wan et al. [WZL$^+$23] initiated the study of regret minimization with quantum reward oracles, establishing polylogarithmic regret for multi-armed and stochastic linear bandits (QLinUCB; see Table 1). Subsequent works by Dai et al. [DLV$^+$23] and Hikima et al. [HMTU24] similarly select actions classically while querying coherent reward unitaries and their inverses; in particular, specializing the kernelized analysis of Hikima et al. to the linear kernel yields $O(d^{3/2}\,\mathrm{polylog}\, T)$ regret for unconstrained quantum linear bandits [LLL26, Appendix C]. Quantum speedups have also been established for best-arm identification [WYLC21], online exp-concave optimization [HLL$^+$24], stochastic convex bandits [LZ22], heavy-tailed bandits [WGAW23], Lipschitz bandits [YKL26], and bandits with knapsacks [SYH$^+$25]. Buchholz et al. [BKS25] give a classical-order lower bound for best-arm identification under their one-time stochastic channel oracle.

None of the above works address round-wise safety under unknown linear constraints. The closest constrained quantum formulation is the quantum bandits-with-knapsacks model of Su et al. [SYH$^+$25], which enforces cumulative resource budgets over the horizon rather than per-round feasibility constraints. Beyond the classical-action coherent-feedback model, we also introduce a coherent-action model equipped with a safe-action bomb oracle, inspired by the bomb-query framework of Lin and Lin [LL15]. Whereas QRS-COLTS queries individual classical actions and observes noisy numerical constraint feedback, the bomb oracle tests exact feasibility across action superpositions. Related coherent-action queries appear in the quantum convex bandit framework of Li and Zhang [LZ22], where regret is charged to a committed classical action at every query round, with the learner playing the previous phase's action while making coherent queries in the current phase. In contrast, our bomb model evaluates performance directly on the queried superpositions via query-weighted regret, integrating the suboptimality gap over safe action mass and assigning a unit penalty to unsafe mass, while separately bounding the cumulative explosion probability. When every query is supported on a single safe action, query-weighted regret reduces to standard cumulative regret.

Quantum mean estimation. Quantum Monte Carlo and amplitude estimation achieve the quadratic speedup $\widetilde{O}(1/\varepsilon)$ for scalar mean estimation [Mon15, KO23]. For bounded vector feedback, we use the multivariate estimator of Cornelissen, Hamoudi, and Jerbi [CHJ22, Theorem 3.3]. For $Z \in [0,1]^M$, it achieves ℓ_∞ error ε with $\widetilde{O}(\sqrt{M}/\varepsilon)$ queries after rescaling (Theorem 3). Marsden et al. [MOSZ25] study directional error guarantees in stochastic convex optimization.

Quantum convex optimization. Our coherent-action bomb algorithm BCO-QLinTS (Theorem 2) relies on quantum convex optimization to maximize perturbed linear objectives over the unknown safe set via membership queries. Quantum convex optimization with membership and separation oracles was studied by van Apeldoorn et al. [vAGGdW20] and Chakrabarti et al. [CCLW20], who showed that optimizing over a well-rounded convex body in $\mathbb{R}^d$ requires only $\widetilde{O}(d)$ quantum membership queries, compared to the best known quadratic classical upper bound. We combine this quantum membership-query primitive with the bomb-query simulation of Lin and Lin [LL15] to obtain a bomb-certified safe optimization subroutine.

3 Preliminaries

Notation. For a positive integer n, $[n] = \{1, \dots, n\}$. Vectors are columns, $\langle x, y\rangle = x^\top y$, and inequalities between vectors hold coordinate-wise. We write $\|x\|_2$ for the Euclidean norm, $\|x\|_A = \sqrt{x^\top A x}$ for a positive definite matrix A, $\mathbb{B}(x, r)$ for the closed Euclidean ball of radius r around x, $\mathbf{1}\{\cdot\}$ for the indicator function, and $\mathbf{1}_m$ for the all-ones vector in $\mathbb{R}^m$. As in Section 1, $a^\star$ is an optimal safe action, $\mu^\star = \langle \theta^\star, a^\star\rangle$ is its mean reward, and $\Delta(a) = \mu^\star - \langle \theta^\star, a\rangle$ is the gap of an action a. We use $\widetilde{O}(\cdot)$ to hide constant and polylogarithmic factors. A power of $\log T$ written inside $\widetilde{O}$ is the dependence that comes from the number of stages, and each theorem states which other logarithms it hides.

3.1 Quantum computing basics

We use only standard notions from the quantum circuit and query models; see, e.g., [NC10] for a comprehensive reference. A quantum register with basis indexed by a finite set Ω is a Hilbert space with computational basis $\{|\omega\rangle : \omega \in \Omega\}$. A pure state is a unit vector $|\psi\rangle$, while a mixed state is represented by a density matrix ρ, i.e., a positive semidefinite operator with trace one. Quantum gates are unitary maps acting on one or more registers. More general noisy operations are quantum channels, i.e., completely positive trace-preserving maps $\mathcal{E}$ on density matrices. A measurement is described by positive operators whose sum is the identity; measuring in the computational basis of $\sum_i c_i |i\rangle$ returns outcome i with probability $|c_i|^2$ and collapses the state accordingly.

A quantum query algorithm interleaves fixed quantum gates with calls to an oracle. In a coherent unitary oracle model, a query is a unitary transformation and may be used inside a larger coherent computation, typically along with its inverse. This is the setting in which amplitude-estimation and quantum-mean-estimation algorithms obtain their speedups. When only a measured feedback value is returned, the learner has no coherent access to the sampling randomness.

We use two kinds of action access. In a *classical-action* query, the learner first chooses a classical action a, and the oracle acts only on the feedback registers associated with that action. In a *coherent-action* query, the action itself is stored in a quantum register, for example $\sum_a \psi_a |a\rangle$, and the oracle acts coherently on all actions in superposition. If such a query is later assigned regret, the natural action weight is the marginal probability $|\psi_a|^2$ (or the corresponding diagonal marginal for a mixed state). This distinction is why the coherent-action model in Section A.3 uses query-weighted regret rather than ordinary played-action regret.

3.2 Quantum mean-estimation primitives

The quantum algorithms in this paper use quantum subroutines to estimate expectations of random variables. We state these subroutines as black boxes in the form needed for our bandit analysis.

A *quantum sampling oracle* for a random variable Y with distribution $\mathcal{D}$ is a unitary that prepares a coherent sample of Y (so-called *qsample*). For example, if $Y = Y(\omega)$ and $\omega \sim p$, then the oracle prepares

$$|0\rangle\,|0\rangle \mapsto \sum_\omega \sqrt{p(\omega)}\;|\omega\rangle\,|Y(\omega)\rangle\,.$$

Measuring the value register produces one ordinary classical sample from $\mathcal{D}$. The quantum advantage comes from using the state coherently before measurement.

We first recall the scalar mean-estimation primitive (see, e.g., [Mon15, KO23]).

Lemma 1 (Scalar quantum mean estimation)**.** *Let $Y \in [0,1]$ be a random variable with mean $\mu = \mathbb{E}[Y]$. Given a quantum sampling oracle for Y and its inverse, there is a quantum algorithm that outputs an estimate $\widehat{\mu}$ satisfying $\mathbb{P}\big(|\widehat{\mu} - \mu| > \varepsilon\big) \le \delta$ using at most $C_{\mathrm{qmc}}\varepsilon^{-1}\log(2/\delta)$ oracle queries for a universal constant $C_{\mathrm{qmc}} > 0$.*

Thus a scalar expectation can be estimated to accuracy ε with $\widetilde{O}(1/\varepsilon)$ coherent oracle queries, instead of the classical $\Theta(1/\varepsilon^2)$ samples required by Monte Carlo.

In our safe-bandit problem, however, the feedback is not scalar: one query returns a vector

$$X(a) = (R(a), C_1(a), \ldots, C_m(a)) \in [0,1]^{m+1}.$$

A direct approach is to apply Lemma 1 separately to all $m+1$ coordinates. This gives coordinate-wise accuracy ε at query cost $\widetilde{O}\big(\frac{m+1}{\varepsilon}\big)$. The query cost thus grows linearly with the number of constraints.

For vector feedback we use QBOUNDED [CHJ22, Theorem 3.3]. Rescaling the feedback vector to the unit Euclidean ball gives the following simultaneous coordinate-wise bound, as shown in the proof of Corollary 3.

Theorem 3 (Bounded multivariate quantum mean estimation)**.** *Let $Z \in [0,1]^M$ be a vector-valued random variable with mean $\mu = \mathbb{E}[Z]$. Given a quantum sampling oracle for Z and its inverse, for any target accuracy $0 < \varepsilon \le 1$ and failure probability $\delta \in (0,1)$, there is a quantum algorithm that outputs $\widehat{\mu} \in \mathbb{R}^M$ such that $\mathbb{P}\big(\|\widehat{\mu} - \mu\|_\infty > \varepsilon\big) \le \delta$ using $\widetilde{O}(\sqrt{M}/\varepsilon)$ oracle queries in the worst case.*

For our feedback vector $X(a)$, the dimension is $M = m + 1$. For accuracy $1/n$ in every coordinate, scalar estimation therefore needs $\widetilde{O}((m+1)n)$ queries, while the multivariate estimator needs $\widetilde{O}(\sqrt{m+1}\,n)$ queries. This gap gives the factor $\sqrt{m+1}$ in Theorem 1.

4 Technical overview

Our two algorithms exploit coherent quantum access in fundamentally different ways, leading to distinct proof architectures. QRS-COLTS operates closer to the classical safe-bandit paradigm: it commits to a single classical action before each query and must learn both the reward vector $\theta^\star$ and the constraint matrix $\Phi^\star$ from noisy feedback. In contrast, BCO-QLinTS operates in the coherent-action model, using the safe-action bomb as an exact membership oracle for the true feasible set $\mathcal{S}$ so that only the reward vector $\theta^\star$ needs to be learned.

QRS-COLTS in six steps. Quantum mean estimation enters in Steps 1–2, while Steps 3–6 adapt the S-COLTS analysis to a stage-wise schedule (Algorithm 1).

1. *Stages.* Because quantum mean estimation queries a fixed action repeatedly to produce a single estimate, QRS-COLTS updates its statistics once per stage. In stage s, it plays a single action a_s, estimates $\mu(a_s) = (\langle\theta^\star, a_s\rangle, \Phi^\star a_s)$ to coordinate-wise accuracy $1/n_s$, and incorporates the result into a weighted ridge regression. Per coordinate, achieving accuracy $1/n_s$ requires $\widetilde{O}(n_s)$ coherent queries rather than $\Theta(n_s^2)$ classical samples. Conditioning on the estimation event yields, for all $a \in \mathcal{A}$ and $i \in [m]$,
$$|\langle\widehat{\theta}_s - \theta^\star, a\rangle| \le \beta_s \|a\|_{V_s^{-1}}, \qquad |\langle\widehat{\phi}_{s,i} - \phi_i^\star, a\rangle| \le \beta_s \|a\|_{V_s^{-1}},$$
where $V_s = \lambda I_d + \sum_{j<s} n_j^2 a_j a_j^\top$ is the weighted design matrix with $\lambda = 1$ and $\beta_s = 1 + \sqrt{s-1}$ is the confidence radius.
2. *Vector feedback.* Estimating the $m+1$ coordinates independently would incur an $(m+1)$ overhead. Instead, the multivariate estimator of Theorem 3 achieves the same ℓ_∞-accuracy $1/n_s$ with only a $\sqrt{m+1}$ overhead (up to logarithmic factors), which suffices for the confidence bounds above.
3. *Coupled optimism.* A single standard normal vector $\zeta \sim \mathcal{N}(0, I_d)$ adds $\beta_s \left\langle \zeta, V_s^{-1/2} a \right\rangle$ to the estimated reward at a and subtracts it from each estimated constraint value. Given a stage history satisfying the confidence bounds, with constant probability at least $\pi_{\mathrm{opt}} = \mathbb{P}(Z \ge 1) \approx 0.16$ for $Z \sim \mathcal{N}(0,1)$, the optimal safe action $a^\star$ satisfies the perturbed constraints and achieves a perturbed reward of at least $\langle\theta^\star, a^\star\rangle$.

Algorithm 1 QRS-COLTS

Initialize $V_1 = \lambda I_d$, $\widehat{\theta}_1 = 0$, $\widehat{\Phi}_1 = 0$, query counter $t = 0$, stage $s = 1$.
while $t < T$ **do**
  Compute β_s and $M_s(a_{\text{safe}}) = B\,\beta_s \|a_{\text{safe}}\|_{V_s^{-1}}$.
  if $M_s(a_{\text{safe}}) > \Gamma_0/3$ **then**
    $a_s \leftarrow a_{\text{safe}}$.
  else
    Draw K_s coupled perturbations, solve their perturbed subproblems, and keep the feasible candidate b_s with the highest sampled value.
    Scale b_s toward a_{safe} via (1); if no subproblem is feasible or no step passes, set $a_s \leftarrow a_{\text{safe}}$.
  end if
  If $a_s = 0$, play it for the remaining queries and stop. Otherwise set $n_s \leftarrow \lceil c_{\text{len}}/\|a_s\|_{V_s^{-1}} \rceil$ and compute the QME budget q_s for accuracy $1/n_s$ and failure probability $\delta_s = 6\delta/(\pi^2 s^2)$.
  If the stage query cost q_s exceeds $T - t$, spend the remaining $T - t$ queries at a_s and stop.
  Otherwise run QME at a_s using q_s queries and update V_{s+1}, $\widehat{\theta}_{s+1}$, $\widehat{\Phi}_{s+1}$, $t \leftarrow t + q_s$, and $s \leftarrow s + 1$.
end while

4. *Resampling.* Because each stage commits many queries to a single action, constant-probability optimism per stage is insufficient. Drawing $K_s = O(\log(s/\delta))$ independent perturbations and selecting the feasible subproblem solution b_s with the highest perturbed reward ensures, with probability at least $1 - \delta$, optimism at every sampling stage with valid confidence bounds. On these stages, if the perturbation norms are also bounded, the selected candidate satisfies $\Delta(b_s) \lesssim B\,\beta_s \|b_s\|_{V_s^{-1}}$, where $B = \widetilde{O}(\sqrt{d})$ bounds the perturbation norms.

5. *Scaling.* At any stage with $M_s(a_{\text{safe}}) := B\,\beta_s \|a_{\text{safe}}\|_{V_s^{-1}} \le \Gamma_0/3$, QRS-COLTS plays $a_s = (1 - \rho_s)a_{\text{safe}} + \rho_s b_s$ using the largest $\rho_s \in [0, 1]$ that satisfies the conservative safety check

$$\widehat{\Phi}_s a_s + \beta_s \|a_s\|_{V_s^{-1}} \mathbf{1}_m \le \alpha. \tag{1}$$

   On the confidence event, any action passing (1) is feasible. On the joint event of confidence, bounded perturbation norms, and global optimism, Lemma 9 gives

$$\Delta(a_s) \lesssim (1 + \Delta(a_{\text{safe}})/\Gamma_0)\, B\,\beta_s \|a_s\|_{V_s^{-1}}.$$

   On this joint event, if $a_s = 0$, the right-hand side vanishes, certifying that $a_s = 0$ is optimal so Algorithm 1 can play it for all remaining queries.

6. *Stage lengths.* Setting $n_s = \lceil c_{\text{len}}/\|a_s\|_{V_s^{-1}} \rceil$ for a constant $c_{\text{len}} > 0$ caps the stage-s regret at

$$\widetilde{O}\big(\sqrt{m+1}\, B\,\beta_s(1 + \Delta(a_{\text{safe}})/\Gamma_0)\big).$$

   Because each completed stage increases $\log\det V_s$ by at least a constant, the total number of stages is bounded by $S = \widetilde{O}(d \log T)$.

Since $\beta_s = O(\sqrt{s})$, summing the per-stage regret over $s \le S$ yields $B \sum_{s \le S} \beta_s = \widetilde{O}(\sqrt{d}\, S^{3/2}) = \widetilde{O}(d^2 \log^{3/2} T)$, which accounts for the d^2 factor in Theorem 1. Here, the $\sqrt{s}$ growth in β_s stems from bounding the accumulated estimation errors, while the extra $\sqrt{d}$ factor relative to the unconstrained $d^{3/2}\,\mathrm{polylog}\, T$ bound comes from the perturbation norm B. Compared with the leading $\widetilde{O}(\sqrt{d^3 T})$ regret of S-COLTS at fixed baseline gap and slack, QRS-COLTS achieves an exponential improvement in T while incurring a d^2 versus $d^{3/2}$ dimension dependence.

BCO-QLinTS: exact safe-set access through a bomb. Our second algorithm operates in the coherent-action bomb model, where the learner may query superpositions over actions and the oracle computes the unsafe-action flag $f(a) = \mathbf{1}\{a \notin \mathcal{S}\}$. While the flag provides exact membership access to the true safe set $\mathcal{S}$, any query placing nonzero amplitude on unsafe actions risks triggering an explosion and terminating the algorithm. Safety thus reduces to a bomb-certified membership constraint: BCO-QLinTS optimizes directly over $\mathcal{S}$ using a bomb-safe quantum convex optimization routine, bypassing the need to estimate the m rows of $\Phi^\star$.

Specifically, in stage s, BCO-QLinTS draws K_s perturbed reward vectors $\widetilde{\theta}_s^{(1)}, \ldots, \widetilde{\theta}_s^{(K_s)}$ around its current reward estimate. For each $k \in [K_s]$, it invokes bomb-certified quantum convex optimization to find a safe action $a_s^{(k)} \in \mathcal{S}$ that maximizes $\langle \widetilde{\theta}_s^{(k)}, a \rangle$ over $\mathcal{S}$ up to additive error γ_s, and selects the candidate pair maximizing $\langle \widetilde{\theta}_s^{(k)}, a_s^{(k)} \rangle$. If some draw is optimistic at $a^\star$ and all calls return γ_s-optimal actions, the selected objective value is at least $\langle \theta^\star, a^\star \rangle - \gamma_s$, and the reward confidence bound at the selected action controls its suboptimality gap. Finally, the learner runs quantum mean estimation of the reward at this certified safe action, which incurs zero explosion probability.

Because per-query rewards and unsafe penalties are bounded in $[0, 1]$, each bomb-optimization query contributes at most 1 to the query-weighted regret. Summing the query complexity of the bomb-certified optimization calls across all stages yields the $\widetilde{O}(d^4 \log^2(2T)/\varepsilon)$ term in Theorem 2, while summing the regret incurred during the reward-estimation phases yields the $\widetilde{O}(d^2 \log^{3/2}(2T/\delta))$ term.

5 Discussion and open questions

We view this work as a first step toward quantum algorithms for safe online learning. With classical actions and coherent feedback, QRS-COLTS preserves round-wise mean safety while improving the horizon dependence from $\sqrt{T}$ to $\operatorname{polylog} T$. With coherent actions and a bomb oracle, BCO-QLinTS bypasses learning the m constraint rows altogether, trading constraint estimation for a bomb-certified optimization overhead.

Limitations. First, our guarantees bound quantum query complexity; as discussed after Theorem 1, query savings translate to wall-clock speedups only when coherent simulator calls dominate the classical optimization and regression updates. Second, QRS-COLTS enforces round-wise mean safety ($\Phi^\star a_t \le \alpha$ for all $t \le T$ with probability at least $1 - 3\delta$), so individual noisy realizations $C_{t,i}$ may still exceed α_i. In particular, when a constraint encodes a binary adverse-event indicator, its expectation is the event probability: a positive threshold $\alpha_i > 0$ bounds the failure probability per round rather than excluding failures almost surely, whereas setting $\alpha_i = 0$ would preclude a strictly feasible baseline with positive slack $\Gamma_0 > 0$. Guaranteeing sample-path safety for every realization requires the constraint feedback to be supported within the safe range.

Lower bounds and open problems. Identically zero costs with positive thresholds recover the classical $\Omega(d\sqrt{T})$ minimax lower bound. For unconstrained coherent feedback, Liu et al. [LLL26] prove an $\Omega(d \log(1 + T/d))$ lower bound on expected regret for a d-action instance and an $O(d^{3/2} \operatorname{polylog} T)$ upper bound for general action sets. Extending the lower bound to our convex safe action sets or to bomb-query regret requires a separate argument. Open questions include (i) for QRS-COLTS (Theorem 1) and the reward-learning term of BCO-QLinTS (Theorem 2), whether the $\widetilde{O}(d^2 \log^{3/2} T)$ scaling (stemming from $\beta_s = O(\sqrt{s})$ over $S = \widetilde{O}(d \log T)$ stages and the perturbation bound $B = \widetilde{O}(\sqrt{d})$; see Lemmas 2, 3 and 10) can be improved to match the $d^{3/2}$ dependence for general action sets, and whether a single logarithm in T is achievable; (ii) whether the $\sqrt{m+1}$ overhead from multivariate mean estimation (Theorem 3 and Corollary 3) and the $(1 + \Delta(a_{\mathrm{safe}})/\Gamma_0)$ scaling factor (Lemma 9) are information-theoretically necessary; and (iii) in the bomb model, whether the $\widetilde{O}(d^4 \log^2 T/\varepsilon)$ optimization overhead in Theorem 2 induced by the quadratic bomb-query simulation [LL15] (Lemma 13) can be improved via direct bomb-safe convex optimization. Beyond tightening these rates, natural directions include studying hybrid quantum feedback models (such as noisy safety flags or coherent rewards paired with classical constraint observations) and extending safe quantum exploration to contextual bandits, knapsack constraints, Bayesian optimization, and reinforcement learning.

Acknowledgments and Disclosure of Funding

Guang Lin gratefully acknowledges the support from DARPA MATHBAC program under grant number: HR0011-26-3-E152, the National Science Foundation (DMS-2533878, DMS-2053746, DMS-2134209, ECCS-2328241, CBET-2347401 and OAC-2311848), and U.S. Department of Energy (DOE) Office of Science Advanced Scientific Computing Research program under the "Uncertainty Quantification for Multifidelity Operator Learning (MOLUcQ)" project (Project No. 81739), DE-

SC0023161, the SciDAC LEADS Institute, and DOE–Fusion Energy Science, under grant number: DE-SC0024583.

References

[AAT19] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32:9252–9262, 2019.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24:2312–2320, 2011.

[BCB12] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[BGH+22] Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.

[BKS18] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM*, 65(3):13:1–13:55, 2018.

[BKS25] Simon Buchholz, Jonas M. Kübler, and Bernhard Schölkopf. Multi-armed bandits and quantum channel oracles. *Quantum*, 9:1672, 2025.

[CCLW20] Shouvanik Chakrabarti, Andrew M. Childs, Tongyang Li, and Xiaodi Wu. Quantum algorithms and lower bounds for convex optimization. *Quantum*, 4:221, 2020.

[CHJ22] Arjan Cornelissen, Yassine Hamoudi, and Sofiene Jerbi. Near-optimal quantum algorithms for multivariate mean estimation. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 33–43, 2022.

[DHK08] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 355–366, 2008.

[DLV+23] Zhongxiang Dai, Gregory Kang Ruey Lau, Arun Verma, Yao Shu, Bryan Kian Hsiang Low, and Patrick Jaillet. Quantum bayesian optimization. *Advances in neural information processing systems*, 36:20179–20207, 2023.

[EV93] Avshalom C Elitzur and Lev Vaidman. Quantum mechanical interaction-free measurements. *Foundations of physics*, 23(7):987–997, 1993.

[GCS24] Aditya Gangrade, Tianrui Chen, and Venkatesh Saligrama. Safe linear bandits over unknown polytopes. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1755–1795. PMLR, 2024.

[GS26] Aditya Gangrade and Venkatesh Saligrama. Constrained linear thompson sampling. *Advances in Neural Information Processing Systems*, 38:121846–121892, 2026.

[HDC+25] S. Hazra, W. Dai, T. Connolly, P. D. Kurilovich, Z. Wang, L. Frunzio, and M. H. Devoret. Benchmarking the readout of a superconducting qubit for repeated measurements. *Physical Review Letters*, 134(10):100601, 2025.

[HLL+24] Jianhao He, Chengchang Liu, Xutong Liu, Lvzhou Li, and John C. S. Lui. Quantum algorithm for online exp-concave optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17946–17971. PMLR, 2024.

[HMTU24] 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*, pages 1640–1657. PMLR, 2024.

[HTA24] Spencer Hutchinson, Berkay Turan, and Mahnoosh Alizadeh. Directional optimism for safe linear bandits. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 658–666. PMLR, 2024.

[KB20] Kia Khezeli and Eilyan Bitar. Safe linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10202–10209, 2020.

[KO23] Robin Kothari and Ryan O'Donnell. Mean estimation when you have the source code; or, quantum Monte Carlo methods. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1186–1215. SIAM, 2023.

[LL15] Cedric Yen-Yu Lin and Han-Hsuan Lin. Upper bounds on quantum query complexity inspired by the Elitzur–Vaidman bomb tester. In *30th Conference on Computational Complexity (CCC 2015)*, volume 33 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 537–566. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2015.

[LLL26] Maoli Liu, Zhuohua Li, and John C. S. Lui. Quantum multi-armed bandits and linear bandits: Lower bounds and algorithms. arXiv preprint arXiv:2608.14319, 2026. To appear in the 38th ACM-SIAM Symposium on Discrete Algorithms (SODA 2027).

[LLSY21] Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34:24075–24086, 2021.

[LZ22] Tongyang Li and Ruizhe Zhang. Quantum speedups of optimizing approximately convex functions with applications to logarithmic regret stochastic convex bandits. *Advances in Neural Information Processing Systems*, 35:3152–3164, 2022.

[MAAT21] Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear Thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69:3755–3767, 2021.

[Mon15] Ashley Montanaro. Quantum speedup of Monte Carlo methods. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2181):20150301, 2015.

[MOSZ25] Annie Marsden, Liam O'Carroll, Aaron Sidford, and Chenyi Zhang. Isotropic noise in stochastic and quantum convex optimization. *Advances in Neural Information Processing Systems*, 38:111482–111523, 2025.

[MTA20] Ahmadreza Moradipari, Christos Thrampoulidis, and Mahnoosh Alizadeh. Stage-wise conservative linear bandits. *Advances in neural information processing systems*, 33:11191–11201, 2020.

[NC10] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 10th anniversary edition, 2010.

[SYH+25] Yuexin Su, Ziyi Yang, Peiyuan Huang, Tongyang Li, and Yinyu Ye. Quantum algorithms for bandits with knapsacks with improved regret and time complexities. *arXiv preprint arXiv:2507.04438*, 2025.

[TC04] Peter F. Thall and John D. Cook. Dose-finding based on efficacy–toxicity trade-offs. *Biometrics*, 60(3):684–693, 2004.

[vAGGdW20] Joran van Apeldoorn, András Gilyén, Sander Gribling, and Ronald de Wolf. Convex optimization using quantum oracles. *Quantum*, 4:220, 2020.

[VLA23] K. Nithin Varma, Sahin Lale, and Anima Anandkumar. Stochastic linear bandits with unknown safety constraints and local feedback. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

[WBSJ21] Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10686–10696. PMLR, 2021.

[WGAW23] Yulian Wu, Chaowen Guan, Vaneet Aggarwal, and Di Wang. Quantum heavy-tailed bandits. *arXiv preprint arXiv:2301.09680*, 2023.

[WYLC21] Daochen Wang, Xuchen You, Tongyang Li, and Andrew M. Childs. Quantum exploration algorithms for multi-armed bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):10102–10110, May 2021.

[WZL+23] Zongqi Wan, Zhijie Zhang, Tongyang Li, Jialin Zhang, and Xiaoming Sun. Quantum multi-armed bandits and stochastic linear bandits enjoy logarithmic regrets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):10087–10094, Jun. 2023.

[WZL+25] Weixin Wang, Haoyang Zheng, Guang Lin, Wei Deng, and Pan Xu. Rethinking langevin thompson sampling from a stochastic approximation perspective. *arXiv preprint arXiv:2510.05023*, 2025.

[YKL26] Bongsoo Yi, Yue Kang, and Yao Li. Quantum lipschitz bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(33):27844–27851, Mar. 2026.

[ZDML24] Haoyang Zheng, Wei Deng, Christian Moya, and Guang Lin. Accelerating approximate Thompson sampling with underdamped Langevin Monte Carlo. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2611–2619. PMLR, 2024.

Appendix

Roadmap. In Section A, we give formal definitions of the three quantum feedback models discussed in the main text: the classical-action coherent-feedback model, the classical-action classical-feedback model, and the coherent-action bomb model with query-weighted regret. In Section B, we present the full QRS-COLTS algorithm and prove the classical-action coherent-feedback result. In Section C, we give the full treatment of the coherent-action bomb model: we define the query-weighted regret formulation, state the bomb-certified safe-optimization primitive, present the BCO-QLinTS algorithm, and prove its explosion-probability and regret guarantees.

A Quantum models for safe stochastic linear bandits

In this section, we provide more details about our quantum action and feedback models for the safe bandit problems.

A.1 Classical-action coherent-feedback model

Our first positive result uses the coherent quantum reward-oracle model of [WZL+23], extended to vector-valued reward-and-constraint feedback. This model keeps the ordinary bandit semantics: before every query, the learner selects a single classical action. Only the access to the feedback distribution of that action is quantum.

For each action a, let Ω_a be a latent outcome space and let p_a be the probability distribution over outcomes. A classical interaction draws $\omega \sim p_a$ and returns

$$X(a,\omega) = (R(a,\omega), C_1(a,\omega), \dots, C_m(a,\omega)) \in [0,1]^{m+1},$$

with mean

$$\mu(a) = \mathbb{E}_{\omega\sim p_a}[X(a,\omega)] = (\langle \theta^\star, a\rangle, \Phi^\star a).$$

The learner knows neither p_a nor the map $\omega \mapsto X(a,\omega)$; it accesses them only through the oracle below. We write $\mathcal{D}_a$ for the law of $X(a,\omega)$ when $\omega \sim p_a$.

Definition 4 (Classical-action coherent-feedback oracle)**.** *For each $a \in \mathcal{A}$, the learner has coherent access to a unitary $\mathcal{O}_a$ satisfying*

$$\mathcal{O}_a |0\rangle |0\rangle = \sum_{\omega\in\Omega_a} \sqrt{p_a(\omega)}\ |\omega\rangle |X(a,\omega)\rangle .$$

Each use of $\mathcal{O}_a$ or $\mathcal{O}_a^\dagger$ counts as one query. The action a is chosen classically before the query.

Measuring the feedback register of $\mathcal{O}_a |0\rangle$ produces exactly one classical feedback sample from $\mathcal{D}_a$. The advantage of this model is that the learner may instead use $\mathcal{O}_a$ coherently, for example inside quantum mean estimation, before any final measurement. Regret is still the standard played-action regret $\mathrm{Reg}(T) = \sum_t \Delta(a_t)$, and round-wise mean safety still means $\Phi^\star a_t \le \alpha$ for every classically chosen query action. Thus this model changes the statistical estimation rate while preserving the usual online safe-bandit interpretation.

A.2 Classical-action classical-feedback model

The second model keeps the same classical-action semantics but removes coherent access to the feedback randomness.

Definition 5 (Classical-action classical-feedback model)**.** *At each query, the learner chooses a classical action a and obtains one feedback sample $X(a,\omega)$, with $\omega \sim p_a$, or interacts with a quantum channel whose output is measured and returned as a classical sample so that the sampling randomness is not available as a reversible unitary sampler. Regret and safety are the standard played-action notions from Section 1.*

This model includes the usual classical stochastic safe linear bandit model. When the feedback randomness is not coherently accessible, quantum algorithms need not inherit the $\widetilde{O}(1/\varepsilon)$ mean-estimation rate [BKS25]. This model is the classical statistical regime against which we compare the coherent models.

A.3 Coherent-action safe-action bomb model and query-weighted regret

The third model uses query-weighted regret. It allows the learner to query a superposition of actions and gives exact, but dangerous, access to the true safe set. Because a single query may place amplitude on many actions, there is no single played action at which to evaluate ordinary regret, so we use query-weighted regret.

Let

$$\mathcal{S} := \{a \in \mathcal{A} : \Phi^\star a \le \alpha\}$$

be the true safe set, and define the unsafe-action predicate

$$f(a) := \mathbf{1}\{a \notin \mathcal{S}\} = \mathbf{1}\{\Phi^\star a \not\le \alpha\}.$$

In this model the oracle reports safety through a flag for $f(a)$ in place of noisy constraint coordinates. Because the flag is measured, a query with nonzero amplitude on unsafe actions risks an explosion.

For each action a, let

$$|\chi_a\rangle_X := \sum_{\omega \in \Omega_a} \sqrt{p_a(\omega)}\, |\omega\rangle\, |R(a,\omega)\rangle$$

denote a coherent qsample of the reward distribution, so that $\mathbb{E}_{\omega\sim p_a}[R(a,\omega)] = \langle \theta^\star, a\rangle$.

Definition 6 (Safe-action bomb oracle)**.** *A coherent-action safe-action bomb oracle maps*

$$|a\rangle\, |0\rangle_f\, |0\rangle_X \mapsto |a\rangle\, |f(a)\rangle_f\, |\chi_a\rangle_X\,.$$

Write $W_a\,|0\rangle = |\chi_a\rangle$. A query applies W_a or $W_a^\dagger$ to the reward register, controlled by the action register, computes $f(a)$ into a fresh flag qubit, and measures the flag; either direction counts as one query. Outcome 1 is a bomb explosion and the algorithm terminates; outcome 0 projects the action register onto the safe subspace after the chosen forward or inverse reward operation.

To define regret, let ρ_t be the joint pre-query state at query t, including the action register and arbitrary work registers. Its action marginal is the probability measure π_t on $\mathcal{A}$ defined by

$$\pi_t(E) = \mathrm{Tr}[(\Pi_E \otimes I)\rho_t]$$

for every measurable $E \subseteq \mathcal{A}$, where Π_E projects the action register onto actions in E. For a pure state $\sum_a \psi_{t,a}\,|a\rangle\,|\mathrm{work}_a\rangle$ over a finite action set with normalized work states, this reduces to $\pi_t(a) = |\psi_{t,a}|^2$.

Definition 7 (Query-weighted regret and explosion probability)**.** *For an action a, let $\Delta_+(a) := (\mu^\star - \langle\theta^\star, a\rangle)_+$ be the positive part of its gap $\Delta(a) = \mu^\star - \langle\theta^\star, a\rangle$, where $\mu^\star = \langle\theta^\star, a^\star\rangle$. The per-query weighted loss in the coherent-action bomb model is*

$$L(\pi_t) := \int_{\mathcal{S}} \Delta_+(a)\, \pi_t(da) + \pi_t(\mathcal{A}\setminus\mathcal{S}),$$

and the query-weighted regret is

$$R_{\mathrm{w}}(T) := \sum_{t=1}^{T} L(\pi_t).$$

Here π_t is conditional on the preceding history and on reaching query t. The conditional probability of explosion at that query is $\pi_t(\mathcal{A}\setminus\mathcal{S})$. The sum includes the query at which the algorithm terminates, if any, and $L(\pi_t) = 0$ after termination.

The unit penalty on unsafe mass is a normalization, since rewards lie in $[0,1]$. If π_t is a point mass on a safe action a_t, then $L(\pi_t) = \Delta_+(a_t) = \Delta(a_t)$, so query-weighted regret reduces to ordinary played-action regret. In general, however, this model is not directly comparable to the classical-action model: it grants coherent access over actions and the bomb flag. BCO-QLinTS (Section C.1) uses the safety flag to optimize directly over $\mathcal{S}$.

B QRS-COLTS: algorithm and proof details

This appendix contains the full algorithmic specification and proof of the classical-action coherent-feedback result stated informally in Theorem 1. The organization mirrors the proof structure described in the main text: we first state the staged algorithm and its abstract mean-estimation assumption, then give a proof overview, and finally provide the detailed safety and regret analysis.

B.1 Algorithm and main guarantee

Throughout, $\mathcal{A}$ is compact and convex, $\|a\|_2 \le 1$, $\lambda = 1$, and $c_{\mathrm{len}} > 0$ is fixed. We now state the abstract estimation condition used by the analysis. It is satisfied by coordinate-wise QMC with $g(m) = m + 1$ and by the multivariate estimator with $g(m) = \sqrt{m+1}$ up to logarithmic factors (Corollaries 2 and 3).

Assumption 1 (Stage-wise vector QME primitive)**.** *At stage s, after choosing a_s, the algorithm returns $\widehat{\mu}_s$ satisfying*

$$\|\widehat{\mu}_s - \mu(a_s)\|_\infty \le \frac{1}{n_s}$$

with probability at least $1 - \delta_s$, and uses a deterministic number of queries q_s obeying

$$q_s = \widetilde{\Theta}(g(m)n_s).$$

Theorem 4 (Round-wise mean safety and polylogarithmic query regret)**.** *Fix $\delta \in (0,1)$ and run QRS-COLTS for T coherent oracle queries. Suppose $\|\theta^\star\|_2 \le 1$, $\|\phi_i^\star\|_2 \le 1$, there is a known strictly safe action with slack lower bound Γ_0, and Assumption 1 holds. Then, with probability at least $1 - 3\delta$, QRS-COLTS is safe at every query and*

$$\mathrm{Reg}(T) = \widetilde{O}\left(g(m)\, d^2\, \log^{3/2}(T/\delta) \left(1 + \frac{\Delta(a_{\mathrm{safe}})}{\Gamma_0}\right)\right),$$

where the $\widetilde{O}(\cdot)$ hides constants depending on the fixed c_{len} and additional logarithmic factors in d, m, T, and $1/\delta$. Theorem 5 states the bound with the stage-count, perturbation-norm, and warm-up terms kept separate.

Instantiations and query accounting. The factor $g(m)$ comes only from estimating the $(m+1)$-dimensional feedback mean. With scalar QMC, one estimates the reward and each constraint coordinate separately. To obtain $\|\widehat{\mu}_s - \mu(a_s)\|_\infty \le 1/n_s$ with stage failure probability δ_s, it suffices to run the scalar routine on each coordinate with failure probability $\delta_s/(m+1)$, leading to

$$q_s = \widetilde{O}((m+1)n_s).$$

The multivariate estimator uses the fact that one coherent query produces the entire vector feedback sample. Theorem 3 with accuracy $1/n_s$ and failure probability δ_s gives the same ℓ_∞ accuracy with

$$q_s = \widetilde{O}(\sqrt{m+1}\, n_s),$$

up to logarithmic factors. This explains the two possible values of $g(m)$ in Theorem 4; the regret theorem itself is written for any vector-mean primitive satisfying Assumption 1.

B.1.1 QRS-COLTS

At the start of stage s, the data consist of the completed stages $j < s$. Write $\widehat{r}_j$ for the estimated reward and $\widehat{c}_{j,i}$ for the estimated value of constraint i. The regression weight of observation j is n_j^2, since its coordinate error is at most $1/n_j$ on the estimation event. The design matrix, radius, and ridge estimates are

$$V_s = \lambda I + \sum_{j<s} n_j^2 a_j a_j^\top, \qquad \beta_s = \sqrt{\lambda} + \sqrt{s-1},$$
$$\widehat{\theta}_s = V_s^{-1} \sum_{j<s} n_j^2 a_j \widehat{r}_j, \qquad \widehat{\phi}_{s,i} = V_s^{-1} \sum_{j<s} n_j^2 a_j \widehat{c}_{j,i}.$$

The ith row of $\widehat{\Phi}_s$ is $\widehat{\phi}_{s,i}^\top$. When $M_s(a_{\mathrm{safe}}) > \Gamma_0/3$, the algorithm plays a_{safe} for the whole stage and skips the subproblems. Otherwise it draws K_s coupled perturbations $\zeta_s^{(k)} \sim \mathcal{N}(0, I_d)$, forms

$$\widetilde{\theta}_s^{(k)} = \widehat{\theta}_s + \beta_s V_s^{-1/2} \zeta_s^{(k)}, \qquad \widetilde{\Phi}_s^{(k)} = \widehat{\Phi}_s - \beta_s \mathbf{1}_m (\zeta_s^{(k)})^\top V_s^{-1/2},$$

and solves the perturbed subproblem

$$b_s^{(k)} \in \arg\max_{a \in \mathcal{A}} \left\langle \widetilde{\theta}_s^{(k)}, a \right\rangle \quad \text{s.t.} \quad \widetilde{\Phi}_s^{(k)} a \le \alpha.$$

The algorithm discards infeasible subproblems, lets b_s be the remaining solution with the largest perturbed objective value, and scales it toward a_{safe}:

$$a_s(\rho) = (1-\rho)a_{\mathrm{safe}} + \rho b_s,$$

choosing the largest $\rho_s \in [0,1]$ such that $a_s(\rho_s)$ passes the conservative test (1),

$$\widehat{\Phi}_s a_s(\rho_s) + \beta_s \left\|a_s(\rho_s)\right\|_{V_s^{-1}} \mathbf{1}_m \le \alpha.$$

The played action is $a_s = a_s(\rho_s)$; if every subproblem is infeasible or no ρ passes (1), it is a_{safe} instead. If $a_s = 0$, then $\|a_s\|_{V_s^{-1}} = 0$ and (2) is undefined; the algorithm plays a_s for the remaining queries, which incurs no regret on the good event (see the proof of Theorem 5). Otherwise it sets

$$n_s = \left\lceil \frac{c_{\mathrm{len}}}{\|a_s\|_{V_s^{-1}}} \right\rceil, \tag{2}$$

and computes the query budget q_s. If this exceeds the remaining budget, it spends the remaining queries at a_s and stops. Otherwise it runs QME at accuracy $1/n_s$ and updates

$$V_{s+1} = V_s + n_s^2 a_s a_s^\top.$$

The failure and resampling schedules used in the proof are

$$\delta_s = \frac{6\delta}{\pi^2 s^2}, \tag{3}$$

and

$$K_s = \left\lceil \frac{1}{\pi_{\mathrm{opt}}} \log \frac{\pi^2 s^2}{\delta} \right\rceil, \tag{4}$$

where $\pi_{\mathrm{opt}} \approx 0.16$ is the local-optimism probability of Lemma 4.

B.2 Proof overview

We outline the proof first, since it needs more than substituting quantum mean estimation into COLTS; the details follow in Section B.3.

Stage-wise confidence. At the end of stage s, QRS-COLTS has queried a single action a_s enough times to estimate the vector mean $\mu(a_s) = (\langle \theta^\star, a_s\rangle, \Phi^\star a_s)$ to accuracy $1/n_s$. Conditional on the QME success event, the normalized stage errors are bounded by one. A ridge-regression calculation then gives, simultaneously for the reward vector and all constraint rows,

$$|\langle \widehat{\theta}_s - \theta^\star, a\rangle| \le \beta_s \|a\|_{V_s^{-1}}, \qquad |\langle \widehat{\phi}_{s,i} - \phi_i^\star, a\rangle| \le \beta_s \|a\|_{V_s^{-1}}, \tag{5}$$

for every action a and every constraint i. This is the point at which the quantum estimator enters the statistical proof: it controls how many oracle queries are needed to make one staged observation accurate enough for (5). The remainder of the argument is deterministic on the confidence event and the perturbation events below.

Optimism and unsaturation. The COLTS perturbation couples the objective and constraints by drawing a Gaussian vector ζ_s and adding $\beta_s V_s^{-1/2}\zeta_s$ to the reward vector and subtracting its transpose from every constraint row. Given a stage-entry history satisfying the confidence bounds, the true optimal safe action $a^\star$ is both feasible for the perturbed subproblem and optimistic in its sampled reward with probability at least π_{opt} (Lemma 4). We call this local optimism. Since a single stage may contain many regret-incurring queries, we resample K_s perturbations and keep the most optimistic subproblem solution. The schedule (4) makes the probability of missing optimism summable over stages. On the joint event of confidence, bounded perturbation norms, and global optimism, Lemma 8 bounds the gap of the selected candidate b_s:

$$\Delta(b_s) \lesssim B\beta_s \|b_s\|_{V_s^{-1}}, \tag{6}$$

where B is a high-probability bound on all perturbation norms (Lemma 3).

Scaling and stage regret. Optimism alone does not give round-wise safety, because the perturbed feasible set may contain actions that are unsafe for the true constraints. QRS-COLTS therefore plays a scaled action $a_s = (1-\rho_s)a_{\text{safe}} + \rho_s b_s$, where ρ_s is chosen so that the conservative constraint (1) holds. By (5), this certificate implies $\Phi^\star a_s \le \alpha$. At stages where $M_s(a_{\text{safe}}) \le \Gamma_0/3$, Lemma 9 gives the following bound on the same joint event.

$$\Delta(a_s) \lesssim \left(1 + \frac{\Delta(a_{\text{safe}})}{\Gamma_0}\right) B\beta_s \|a_s\|_{V_s^{-1}}. \tag{7}$$

If this stage is completed and $a_s \neq 0$, its length satisfies $n_s \asymp 1/\|a_s\|_{V_s^{-1}}$. Thus, if QME costs $q_s = \widetilde{O}(g(m)n_s)$ queries, the regret incurred in stage s is

$$q_s\Delta(a_s) = \widetilde{O}\left(g(m)B\beta_s\left(1 + \frac{\Delta(a_{\text{safe}})}{\Gamma_0}\right)\right).$$

Finally, the design update $V_{s+1} = V_s + n_s^2 a_s a_s^\top$ ensures that every completed stage increases the log-determinant potential by a constant. Hence the number of stages is only polylogarithmic in the query budget, and summing the stage bound yields Theorem 4.

B.3 Detailed proofs

We now state and prove the safety and regret guarantees. The analysis is separated into: (i) a QME success event implying deterministic bounded regression errors, (ii) local-to-global optimism for the resampled coupled perturbations, (iii) a scaling-to-safety lemma controlling the loss due to scaling, and (iv) a stage-count bound via a log-determinant potential.

B.3.1 Confidence event for staged QME

Recall the stage-wise failure budget $\delta_s = 6\delta/(\pi^2 s^2)$ from (3), so that $\sum_{s\ge1}\delta_s \le \delta$.

For coordinate-wise scalar QMC applied to the $m+1$ coordinates of $X(a_s)$, we will further allocate

$$\delta_{s,\text{coord}} := \frac{\delta_s}{m+1},$$

so that a union bound over coordinates yields stage-wise success probability at least $1-\delta_s$.

Definition 8 (QME success event)**.** *Let* $\mathrm{QMEgood}(\delta)$ *be the event that at every completed estimation stage s, the stage mean-estimation routine returns $\widehat{\mu}_s$ satisfying*

$$\|\widehat{\mu}_s - \mu(a_s)\|_\infty \le \frac{1}{n_s}.$$

Conditional on the history at the start of stage s, a stage that completes its budget fails with probability at most δ_s, and the estimate of a truncated stage is discarded. A union bound over stages gives

$$\mathbb{P}\big(\mathrm{QMEgood}(\delta)^c\big) \le \sum_{s\ge1}\delta_s = \frac{6\delta}{\pi^2}\sum_{s\ge1}s^{-2} = \delta.$$

The following lemma gives confidence bounds on $\mathrm{QMEgood}(\delta)$.

Lemma 2 (Confidence bounds from bounded stage errors)**.** *Assume $\|\theta^\star\|_2 \le 1$ and $\|\phi_i^\star\|_2 \le 1$ for all $i \in [m]$. On the event* $\mathrm{QMEgood}(\delta)$*, simultaneously for all stages $s \ge 1$, all actions $a \in \mathbb{R}^d$, and all constraints $i \in [m]$,*

$$\left|\left\langle \widehat{\theta}_s - \theta^\star, a\right\rangle\right| \le \beta_s \|a\|_{V_s^{-1}}, \tag{8}$$

$$\left|\left\langle \widehat{\phi}_{s,i} - \phi_i^\star, a\right\rangle\right| \le \beta_s \|a\|_{V_s^{-1}}, \tag{9}$$

where $\beta_s = \sqrt{\lambda} + \sqrt{s-1}$ and $V_s = \lambda I + \sum_{j=1}^{s-1} n_j^2 a_j a_j^\top$.

Proof. Fix one coordinate of the feedback vector. Write $w^\star$ for its parameter and $\widehat{y}_j$ for its estimate at stage j. To put the errors at different stages on the same scale, define

$$x_j = n_j a_j, \qquad y_j = n_j \widehat{y}_j, \qquad \xi_j = y_j - x_j^\top w^\star.$$

On $\mathrm{QMEgood}(\delta)$ each $|\xi_j|$ is at most one, because the error before multiplication by n_j is at most $1/n_j$. Stack the rows $x_j^\top$ in X and the responses y_j in y, for $j < s$. Using $y = Xw^\star + \xi$ in the ridge normal equations gives

$$\begin{aligned}\widehat{w}_s &= V_s^{-1} X^\top y \\ &= V_s^{-1}(V_s - \lambda I)w^\star + V_s^{-1} X^\top \xi, \\ \widehat{w}_s - w^\star &= -\lambda V_s^{-1} w^\star + V_s^{-1} X^\top \xi.\end{aligned}$$

The second equality uses $X^\top X = V_s - \lambda I$. We bound the two error terms separately. Since $V_s \succeq \lambda I$ and $\|w^\star\|_2 \le 1$, the regularization term satisfies

$$\|\lambda V_s^{-1} w^\star\|_{V_s} = \lambda \|w^\star\|_{V_s^{-1}} \le \sqrt{\lambda}\, \|w^\star\|_2 \le \sqrt{\lambda}.$$

For the estimation term, the nonzero eigenvalues of $X V_s^{-1} X^\top$ are $t^2/(t^2+\lambda)$, where t ranges over the nonzero singular values of X. Each is at most one. It follows that

$$\|V_s^{-1} X^\top \xi\|_{V_s}^2 = \xi^\top X V_s^{-1} X^\top \xi \le \|\xi\|_2^2 = \sum_{j<s} |\xi_j|^2 \le s - 1.$$

Adding the two bounds by the triangle inequality gives

$$\|\widehat{w}_s - w^\star\|_{V_s} \le \sqrt{\lambda} + \sqrt{s-1} = \beta_s.$$

To obtain the error in the direction of an arbitrary action a, apply Cauchy–Schwarz after inserting $V_s^{1/2} V_s^{-1/2} = I$:

$$|\langle \widehat{w}_s - w^\star, a \rangle| \le \|\widehat{w}_s - w^\star\|_{V_s}\, \|a\|_{V_s^{-1}} \le \beta_s \|a\|_{V_s^{-1}}.$$

The same event bounds all completed stage errors in all coordinates. Thus this argument gives both confidence inequalities simultaneously. □

Definition 9 (Confidence event)**.** *Let* $\mathrm{Con}(\delta)$ *denote the event that* (8) *and* (9) *hold for all* s *and all* a*. By Lemma 2,* $\mathbb{P}(\mathrm{Con}(\delta)) \ge 1 - \delta$*.*

B.3.2 Bounding perturbations and defining the uncertainty proxy

We will condition on a high-probability event that all Gaussian perturbations used by the algorithm have bounded Euclidean norm. Let N_{Gauss} be an upper bound on the total number of Gaussian vectors sampled by the algorithm; we may take $N_{\mathrm{Gauss}} = TK_T$ as a crude bound since each stage uses at most K_T resamples and $S \le T$.

Lemma 3 (Gaussian norm bound with explicit N_{Gauss})**.** *Let* $\zeta \sim \mathcal{N}(0, I_d)$*. For any* $\eta \in (0,1)$*,*

$$\mathbb{P}\big(\|\zeta\|_2 \ge \sqrt{d} + \sqrt{2\log(1/\eta)}\big) \le \eta.$$

Consequently, if the algorithm samples at most N_{Gauss} *Gaussian vectors in total (across all stages and resamples), and we set*

$$B(\delta) \;:=\; \sqrt{d} + \sqrt{2\log(2N_{\mathrm{Gauss}}/\delta)},$$

then by a union bound

$$\mathbb{P}\Big(\max_{1 \le \ell \le N_{\mathrm{Gauss}}} \big\|\zeta^{(\ell)}\big\|_2 \le B(\delta)\Big) \ge 1 - \delta/2.$$

Moreover, under the adaptive schedule (4) *and for any* $S \ge 1$*,*

$$\sum_{s=1}^{S} K_s \;\le\; S + \frac{S}{\pi_{\mathrm{opt}}}\Big(\log\frac{\pi^2}{\delta} + 2\log S\Big) \;\le\; \frac{3S}{\pi_{\mathrm{opt}}}\log\frac{\pi S}{\delta},$$

where the last inequality holds for all $S \ge 1$ *and* $0 < \delta < 1$*. In particular, if* $S_{\max}$ *is any deterministic upper bound on the number of stages, one may take*

$$N_{\mathrm{Gauss}} := \frac{3S_{\max}}{\pi_{\mathrm{opt}}}\log\frac{\pi S_{\max}}{\delta},$$

which yields $B(\delta) = \sqrt{d} + O(\sqrt{\log(S_{\max}/\delta)})$*.*

Proof. The Euclidean norm is a 1-Lipschitz function. Jensen's inequality bounds its mean, and Gaussian concentration bounds the deviation above that mean. Together they give

$$\mathbb{E}\|\zeta\|_2 \le \left(\mathbb{E}\|\zeta\|_2^2\right)^{1/2} = \sqrt{d},$$
$$\mathbb{P}\big(\|\zeta\|_2 \ge \sqrt{d} + t\big) \le e^{-t^2/2} \qquad (t \ge 0).$$

Taking $t = \sqrt{2\log(1/\eta)}$ proves the first claim. For the uniform bound, take $\eta = \delta/(2N_{\mathrm{Gauss}})$ and sum the failure probabilities over a deterministic list of N_{Gauss} Gaussian draws. An adaptive run uses only a prefix of this list.

It remains to bound the number of draws. By the ceiling in (4) and the monotonicity of $\log s$,

$$\begin{aligned}\sum_{s=1}^{S} K_s &\le S + \frac{1}{\pi_{\mathrm{opt}}} \sum_{s=1}^{S} \left(\log \frac{\pi^2}{\delta} + 2\log s\right) \\ &\le S + \frac{S}{\pi_{\mathrm{opt}}} \log \frac{\pi^2 S^2}{\delta} \le \frac{3S}{\pi_{\mathrm{opt}}} \log \frac{\pi S}{\delta}.\end{aligned}$$

For the last step, expand the logarithm and use $\delta < 1$ and $\log(\pi S/\delta) \ge 1$ to absorb the additive S. Substituting the deterministic stage bound $S_{\max}$ for S proves the claimed bound on $B(\delta)$. □

We define the stage-wise uncertainty proxy

$$M_s(a) := B(\delta)\, \beta_s\, \|a\|_{V_s^{-1}}. \tag{10}$$

B.3.3 Local optimism and adaptive resampling

We now show that, on the confidence event, each perturbation draw has constant probability of being optimistic in that $a^\star$ stays feasible and its sampled reward is at least $\langle \theta^\star, a^\star \rangle$.

Lemma 4 (Local optimism probability)**.** *Fix a stage s and condition on a history $\mathcal{F}_{s-1}$ satisfying the confidence bounds at this stage. Let $a^\star$ be an optimal feasible action. For a single perturbation draw $(\widetilde{\theta}_s, \widetilde{\Phi}_s)$ constructed as above,*

$$\mathbb{P}\Big(\Big\langle \widetilde{\theta}_s, a^\star \Big\rangle \ge \langle \theta^\star, a^\star \rangle \;\; \textit{and} \;\; \widetilde{\Phi}_s a^\star \le \alpha \;\Big|\; \mathcal{F}_{s-1} \Big) \;\ge\; \pi_{\mathrm{opt}},$$

where $\pi_{\mathrm{opt}} := \mathbb{P}(Z \ge 1)$ for $Z \sim \mathcal{N}(0,1)$ is a universal constant (numerically $\pi_{\mathrm{opt}} \approx 0.1587$).

Proof. Fix a history up to stage s on which both confidence inequalities hold. If $a^\star = 0$, the perturbation vanishes at $a^\star$ and the conclusion holds with probability one. Otherwise write $w = \|a^\star\|_{V_s^{-1}} > 0$ and define

$$Z = \frac{\langle \zeta, V_s^{-1/2} a^\star \rangle}{w}.$$

Conditional on the history, the vector $V_s^{-1/2} a^\star / w$ is a fixed unit vector, so Z is standard normal. The same scalar appears with opposite signs in the sampled reward and in every sampled constraint. For the reward,

$$\begin{aligned}\langle \widetilde{\theta}_s, a^\star \rangle &= \langle \widehat{\theta}_s, a^\star \rangle + \beta_s w Z \\ &\ge \langle \theta^\star, a^\star \rangle + \beta_s w (Z - 1).\end{aligned}$$

The equality is the reward perturbation rule, and the inequality is the lower confidence bound on the reward estimate. For each constraint row i, the upper confidence bound and true feasibility give

$$\begin{aligned}\langle \widetilde{\phi}_{s,i}, a^\star \rangle &= \langle \widehat{\phi}_{s,i}, a^\star \rangle - \beta_s w Z \\ &\le \langle \phi_i^\star, a^\star \rangle + \beta_s w (1 - Z) \\ &\le \alpha_i + \beta_s w (1 - Z).\end{aligned}$$

Thus the single event $Z \ge 1$ gives optimism in reward and feasibility in all m constraints. Its conditional probability is π_{opt}. □

Adaptive resampling. Recall the resampling budget K_s from (4) and write $\nu_s := \delta/(\pi^2 s^2)$, so that $K_s = \lceil \pi_{\mathrm{opt}}^{-1} \log(1/\nu_s) \rceil$.

Lemma 5 (Global optimism via adaptive resampling)**.** *Let K_s be as in* (4)*. Then, with probability at least $1-\delta$, for every stage that solves the perturbed subproblems and has valid confidence bounds, there exists at least one resample $k \in [K_s]$ such that $a^\star$ is feasible for $(\widetilde{\theta}_s^{(k)}, \widetilde{\Phi}_s^{(k)})$ and $\left\langle \widetilde{\theta}_s^{(k)}, a^\star \right\rangle \geq \langle \theta^\star, a^\star \rangle$. Consequently, the chosen b_s satisfies the* global optimism *inequality*

$$\left\langle \widetilde{\theta}_s, b_s \right\rangle \geq \langle \theta^\star, a^\star \rangle . \tag{11}$$

Proof. Fix a history in which the perturbed subproblems are used, and the stage confidence bounds hold. By Lemma 4, a single draw succeeds with probability at least π_{opt}. Since the K_s draws are conditionally independent, $1 - x \leq e^{-x}$ and the choice of K_s give

$$\mathbb{P}(\text{no successful draw at stage } s \mid \mathcal{F}_{s-1}) \leq (1-\pi_{\mathrm{opt}})^{K_s} \leq e^{-\pi_{\mathrm{opt}} K_s} \leq \nu_s.$$

Averaging over these histories and summing over stages, the probability that some stage has valid confidence bounds but no successful draw is at most $\sum_{s \geq 1} \nu_s = \delta/6 \leq \delta$.

Now let k be a successful draw, so that $a^\star$ is feasible for its subproblem, and let $(\widetilde{\theta}_s, b_s)$ be the selected pair. The selection rule, the optimality of $b_s^{(k)}$ in its subproblem, and the optimism of draw k give

$$\begin{aligned} \langle \widetilde{\theta}_s, b_s \rangle &\geq \langle \widetilde{\theta}_s^{(k)}, b_s^{(k)} \rangle \\ &\geq \langle \widetilde{\theta}_s^{(k)}, a^\star \rangle \geq \langle \theta^\star, a^\star \rangle. \end{aligned}$$

This proves (11). □

B.3.4 Scaling-to-safety and scaling loss

The scaling step enforces conservative feasibility based on the confidence radius. We first show this implies true feasibility.

Lemma 6 (Round-wise mean safety on the confidence event)**.** *On* $\mathrm{Con}(\delta)$*, any action a satisfying the conservative constraints $\widehat{\Phi}_s a + \beta_s \|a\|_{V_s^{-1}} \mathbf{1}_m \leq \alpha$ is feasible for the true constraints: $\Phi^\star a \leq \alpha$. In particular, QRS-COLTS plays only feasible actions on* $\mathrm{Con}(\delta)$.

Proof. Fix a stage and any action satisfying the conservative constraints. For each row i, the confidence bound and the conservative constraint give

$$\langle \phi_i^\star, a \rangle \leq \langle \widehat{\phi}_{s,i}, a \rangle + \beta_s \|a\|_{V_s^{-1}} \leq \alpha_i.$$

This proves feasibility for all rows. QRS-COLTS plays either an action that satisfies the conservative constraints or a_{safe}, which is feasible by assumption, so on $\mathrm{Con}(\delta)$ every played action is feasible. □

We next quantify how much scaling can shrink the step-size ρ_s.

Lemma 7 (Scaling loss inequalities)**.** *Let $B := B(\delta) \geq 2$ and $M_s(a_{\mathrm{safe}}) \leq \Gamma_0/3$, and let Γ be the true margin of the baseline,*

$$\Gamma = \Gamma(a_{\mathrm{safe}}) = \min_{i \in [m]} \big(\alpha_i - \langle \phi_i^\star, a_{\mathrm{safe}} \rangle\big) \geq \Gamma_0.$$

Fix a stage s and suppose the candidate b_s satisfies the sampled-feasibility condition $\widetilde{\Phi}_s b_s \leq \alpha$ for the perturbation that generated b_s. Let $a_s = (1-\rho_s) a_{\mathrm{safe}} + \rho_s b_s$ be chosen by the scaling rule (1)*. Then on the intersection of* $\mathrm{Con}(\delta)$ *and the Gaussian norm event of Lemma 3,*

$$(1-\rho_s)\,\Gamma \;\leq\; \frac{3(B+1)}{2B-3}\, M_s(a_s), \tag{12}$$

$$\rho_s\, M_s(b_s) \;\leq\; \frac{3B-2}{2B-3}\, M_s(a_s). \tag{13}$$

In particular, the constants approach $3/2$ as $B \to \infty$, and for $B = 2$ we recover $(1-\rho_s)\Gamma \leq 9 M_s(a_s)$ and $\rho_s M_s(b_s) \leq 4 M_s(a_s)$.

Proof. Write $u = a_{\text{safe}}$ and $b = b_s$. For each row i, let F_i be its conservative upper confidence bound,

$$F_i(x) = \langle \widehat{\phi}_{s,i}, x \rangle + \beta_s \|x\|_{V_s^{-1}}.$$

We bound the slack at u and the possible excess at b, and then use the convexity of F_i to find how far the action can move from u toward b.

Slack at the baseline. The confidence inequality at u gives

$$\begin{aligned} F_i(u) &\le \langle \phi_i^\star, u \rangle + 2\beta_s \|u\|_{V_s^{-1}} \\ &\le \alpha_i - \Gamma + \frac{2}{B} M_s(u) \le \alpha_i - \left(1 - \frac{2}{3B}\right)\Gamma. \end{aligned}$$

The second inequality uses the definition of Γ, and the last uses the stage test $M_s(u) \le \Gamma_0/3 \le \Gamma/3$. Thus, with $\kappa = 1 - 2/(3B) > 0$,

$$F_i(u) \le \alpha_i - \kappa\Gamma. \tag{14}$$

Possible excess at the candidate. The candidate satisfies the constraints sampled with its own perturbation ζ. Undoing the perturbation in each row gives

$$\begin{aligned} F_i(b) &= \langle \widetilde{\phi}_{s,i}, b \rangle + \beta_s \langle \zeta, V_s^{-1/2} b \rangle + \beta_s \|b\|_{V_s^{-1}} \\ &\le \alpha_i + (B+1)\beta_s \|b\|_{V_s^{-1}}. \end{aligned}$$

The sampled constraint bounds the first term by α_i, and Cauchy–Schwarz with the Gaussian norm event bounds the second by $B\beta_s\|b\|_{V_s^{-1}}$. Since $M_s(b) = B\beta_s\|b\|_{V_s^{-1}}$, the bound reads

$$F_i(b) \le \alpha_i + (1 + 1/B) M_s(b). \tag{15}$$

A feasible point on the segment. Put $a(\rho) = (1-\rho)u + \rho b$. Each F_i is convex, being the sum of a linear function and a norm, so for $0 \le \rho \le 1$, (14) and (15) give

$$\begin{aligned} F_i(a(\rho)) &\le (1-\rho)F_i(u) + \rho F_i(b) \\ &\le \alpha_i - (1-\rho)\kappa\Gamma + \rho(1 + 1/B) M_s(b). \end{aligned}$$

Hence $a(\rho)$ passes every constraint if

$$\rho(1 + 1/B) M_s(b) \le (1-\rho)\kappa\Gamma. \tag{16}$$

Solving the equality for ρ gives the feasible step

$$\bar{\rho} = \frac{\kappa\Gamma}{\kappa\Gamma + (1 + 1/B) M_s(b)} \in (0, 1].$$

Since ρ_s is the largest step that passes (1), it is at least $\bar{\rho}$. Hence

$$(1-\rho_s)\Gamma \le (1-\bar{\rho})\Gamma = \frac{1 + 1/B}{\kappa}\bar{\rho} M_s(b) \le \frac{1 + 1/B}{\kappa}\rho_s M_s(b).$$

Writing out the coefficient gives our first comparison,

$$(1-\rho_s)\Gamma \le c_B\, \rho_s M_s(b), \qquad c_B = \frac{3(B+1)}{3B-2}. \tag{17}$$

Bounds in terms of the played action. To eliminate $M_s(b)$ from (17), use $\rho_s b = a_s - (1-\rho_s)u$. The triangle inequality for M_s, followed by the stage test, gives

$$\rho_s M_s(b) \le M_s(a_s) + (1-\rho_s) M_s(u) \le M_s(a_s) + \tfrac{1}{3}(1-\rho_s)\Gamma. \tag{18}$$

Substituting this into (17) and collecting the terms containing $(1-\rho_s)\Gamma$ gives

$$(1 - c_B/3)(1-\rho_s)\Gamma \le c_B M_s(a_s), \qquad 1 - c_B/3 = \frac{2B-3}{3B-2} > 0.$$

Dividing by $1 - c_B/3$ proves (12). A final substitution in (18) yields

$$\rho_s M_s(b) \le \left(1 + \frac{B+1}{2B-3}\right) M_s(a_s) = \frac{3B-2}{2B-3} M_s(a_s),$$

which proves (13). □

B.3.5 Unsaturation and a per-stage gap bound

Here $\Delta(a) = \langle \theta^\star, a^\star - a \rangle$ is the signed gap of Section 3, which is affine in a; a candidate b_s may be infeasible and have $\Delta(b_s) < 0$. We now relate the true reward gap $\Delta(a_s)$ to the uncertainty proxy $M_s(a_s)$ on optimistic stages.

Lemma 8 (Candidate gap bound under global optimism)**.** *Assume* $\mathrm{Con}(\delta)$ *and the Gaussian norm event of Lemma 3. If stage* s *is globally optimistic in the sense of* (11)*, then the chosen candidate* b_s *satisfies*

$$\Delta(b_s) \le \Big(1 + \frac{1}{B(\delta)}\Big) M_s(b_s). \tag{19}$$

Proof. Let $\widetilde{\theta}_s$ be the perturbed reward vector paired with b_s. Subtracting $\langle \theta^\star, b_s \rangle$ from both sides of (11) gives

$$\begin{aligned}\Delta(b_s) &= \langle \theta^\star, a^\star \rangle - \langle \theta^\star, b_s \rangle \\ &\le \langle \widetilde{\theta}_s - \theta^\star, b_s \rangle \\ &= \langle \widetilde{\theta}_s - \widehat{\theta}_s, b_s \rangle + \langle \widehat{\theta}_s - \theta^\star, b_s \rangle.\end{aligned}$$

The first term comes from the perturbation. The Gaussian norm event holds for every draw, including the one selected after the subproblems are solved, so by Cauchy–Schwarz

$$\begin{aligned}|\langle \widetilde{\theta}_s - \widehat{\theta}_s, b_s \rangle| &= \beta_s |\langle \zeta_s, V_s^{-1/2} b_s \rangle| \\ &\le B\beta_s \|b_s\|_{V_s^{-1}} = M_s(b_s).\end{aligned}$$

The second term comes from the regression error, and by (8) its absolute value is at most $M_s(b_s)/B$. Adding the two bounds proves (19). □

Lemma 9 (Stage gap bound (unsaturation after scaling))**.** *Let* $B := B(\delta) \ge 2$*. Suppose* $M_s(a_{\mathrm{safe}}) \le \Gamma_0/3$*. Assume* $\mathrm{Con}(\delta)$*, the Gaussian norm event, and global optimism at stage* s*. Assume moreover that* b_s *is feasible for its sampled constraints* $\widetilde{\Phi}_s b_s \le \alpha$*. Let* $a_s = (1-\rho_s)a_{\mathrm{safe}} + \rho_s b_s$ *be the scaled action. Then*

$$\Delta(a_s) \;\le\; \left(\frac{(B+1)(3B-2)}{B(2B-3)} + \frac{3(B+1)}{2B-3} \cdot \frac{\Delta(a_{\mathrm{safe}})}{\Gamma(a_{\mathrm{safe}})} \right) M_s(a_s). \tag{20}$$

In particular, for $B \ge 2$*,*

$$\Delta(a_s) \;\le\; \Big(6 + 9\,\frac{\Delta(a_{\mathrm{safe}})}{\Gamma(a_{\mathrm{safe}})}\Big)\, M_s(a_s),$$

and as $B \to \infty$ *the leading constant approaches* $\frac{3}{2}\big(1 + \Delta(a_{\mathrm{safe}})/\Gamma(a_{\mathrm{safe}})\big)$*.*

Proof. Write $u = a_{\mathrm{safe}}$ and $\Gamma = \Gamma(u)$. Since the reward is linear and the two coefficients in $a_s = (1-\rho_s)u + \rho_s b_s$ sum to one, the signed gap satisfies

$$\Delta(a_s) = (1-\rho_s)\Delta(u) + \rho_s \Delta(b_s).$$

The first term is the reward lost by keeping weight $1-\rho_s$ on the baseline. Since u is safe, $\Delta(u) \ge 0$, and multiplying (12) by $\Delta(u)/\Gamma$ gives

$$(1-\rho_s)\Delta(u) = \frac{\Delta(u)}{\Gamma}(1-\rho_s)\Gamma \le \frac{\Delta(u)}{\Gamma}\frac{3(B+1)}{2B-3} M_s(a_s).$$

For the second term, apply the candidate-gap bound and then (13):

$$\begin{aligned}\rho_s \Delta(b_s) &\le (1 + 1/B)\rho_s M_s(b_s) \\ &\le \frac{(B+1)(3B-2)}{B(2B-3)} M_s(a_s).\end{aligned}$$

Adding the two estimates gives (20). For the simpler constants, observe that for $B \ge 2$,

$$1 + \frac{1}{B} \le \frac{3}{2}, \qquad \frac{3B-2}{2B-3} = \frac{3}{2} + \frac{5}{2(2B-3)} \le 4, \qquad \frac{3(B+1)}{2B-3} = \frac{3}{2} + \frac{15}{2(2B-3)} \le 9.$$

The first two bounds give the coefficient $6 = \frac{3}{2} \cdot 4$ of the candidate term, and the third gives the coefficient 9 of the baseline term. Letting $B \to \infty$ in the exact coefficients gives the stated limit. □

B.3.6 Stage count

Let $d_{\mathcal{A}} := \dim \operatorname{span}(\mathcal{A}) \le d$. If $d_{\mathcal{A}} = 0$, regret vanishes; assume $d_{\mathcal{A}} > 0$ below.

Lemma 10 (Stage count via log-determinant growth)**.** *Assume the stage rule $n_s = \lceil c_{\mathrm{len}} / \|a_s\|_{V_s^{-1}} \rceil$ and the update $V_{s+1} = V_s + n_s^2 a_s a_s^\top$. Then after S completed stages,*

$$S \;\le\; \frac{d_{\mathcal{A}}}{\log(1+c_{\mathrm{len}}^2)}\;\log\Bigl(1+\frac{1}{d_{\mathcal{A}}\lambda}\sum_{s=1}^{S} n_s^2\Bigr) \;\le\; \frac{d_{\mathcal{A}}}{\log(1+c_{\mathrm{len}}^2)}\;\log\Bigl(1+\frac{1}{d_{\mathcal{A}}\lambda}\Bigl(\sum_{s=1}^{S} n_s\Bigr)^2\Bigr).$$

Proof. Write $c = c_{\mathrm{len}}$. By the matrix determinant lemma and the ceiling in the stage rule, each completed stage multiplies the determinant by at least $1 + c^2$:

$$\begin{aligned}\frac{\det V_{s+1}}{\det V_s} &= 1 + n_s^2 \|a_s\|_{V_s^{-1}}^2 \\ &\ge 1 + c^2.\end{aligned}$$

Multiplying these ratios over the completed stages gives a lower bound of $(1 + c^2)^S$ on $\det V_{S+1} / \det V_1$.

For the upper bound, let $D = V_{S+1} - \lambda I$, which vanishes on the orthogonal complement of $\operatorname{span}(\mathcal{A})$, and let $\gamma_1, \ldots, \gamma_{d_{\mathcal{A}}}$ be its eigenvalues on $\operatorname{span}(\mathcal{A})$, including any zeros. Since $V_1 = \lambda I$, diagonalization and the arithmetic–geometric mean inequality give

$$\begin{aligned}\frac{\det V_{S+1}}{\det V_1} &= \prod_{i=1}^{d_{\mathcal{A}}} (1 + \gamma_i/\lambda) \\ &\le \left(1 + \frac{\operatorname{tr}(D)}{d_{\mathcal{A}}\lambda}\right)^{d_{\mathcal{A}}}.\end{aligned}$$

Here $\sum_i \gamma_i = \operatorname{tr}(D)$. Taking traces in the stage updates and using $\|a_s\|_2 \le 1$ yields

$$\operatorname{tr}(D) = \sum_{s=1}^{S} n_s^2 \|a_s\|_2^2 \le \sum_{s=1}^{S} n_s^2.$$

Taking logarithms of the two determinant bounds and dividing by $\log(1 + c^2)$ gives the first claim. Expanding the square of $\sum_s n_s$ adds only nonnegative cross terms, which gives the second claim. □

B.3.7 Warm-up bound

The algorithm plays a_{safe} whenever $M_s(a_{\mathrm{safe}}) > \Gamma_0/3$. We show this can only happen for a polylogarithmic number of oracle interactions.

Lemma 11 (Warm-up query cost)**.** *Let T_{warm} be the total number of oracle queries spent in stages where QRS-COLTS plays a_{safe} because $M_s(a_{\mathrm{safe}}) > \Gamma_0/3$. Assume the QME subroutine used for stage estimation has query complexity $q_s \le \widetilde{O}(g(m)\, n_s)$ with $g(m) \ge 1$. Then on $\mathrm{Con}(\delta)$,*

$$T_{\mathrm{warm}} = \widetilde{O}\left(g(m)\,\frac{B(\delta)}{\Gamma_0}\,\beta_S\right),$$

where S is the number of stages begun by the horizon.

Proof. Let $s_1 < \cdots < s_\ell$ be the warm-up stages begun by the horizon. If $\ell = 0$, the claim is immediate. Otherwise write $u = a_{\mathrm{safe}}$, $c = c_{\mathrm{len}}$, and $z_s = u^\top V_s^{-1} u$. The warm-up test implies $u \ne 0$. Because $V_s \succeq I$ and $\|u\|_2 \le 1$, we have $0 < z_s \le 1$.

At each completed warm-up stage, the update is in the direction u. Applying Sherman–Morrison in this direction gives

$$\begin{aligned}z_{s+1} &= z_s - \frac{n_s^2 z_s^2}{1 + n_s^2 z_s} \\ &= \frac{z_s}{1 + n_s^2 z_s} \le \frac{z_s}{1 + c^2}.\end{aligned}$$

The last inequality uses $n_s\sqrt{z_s} \geq c$. Every intervening update increases V_s in the positive-semidefinite order, so it can only decrease z_s. Every warm-up stage before s_ℓ is completed, so iterating this decrease with $\kappa_{\mathrm{w}} = (1+c^2)^{-1/2} < 1$ gives

$$z_{s_\ell} \leq \kappa_{\mathrm{w}}^{2(\ell-j)} z_{s_j}, \qquad 1 \leq j \leq \ell,$$
$$\frac{1}{\sqrt{z_{s_j}}} \leq \frac{\kappa_{\mathrm{w}}^{\ell-j}}{\sqrt{z_{s_\ell}}}.$$

Since $n_s \leq c/\sqrt{z_s} + 1 \leq (c+1)/\sqrt{z_s}$, summing over the warm-up stages and then the geometric series gives

$$\begin{aligned}\sum_{j=1}^{\ell} n_{s_j} &\leq (c+1)\sum_{j=1}^{\ell} \frac{1}{\sqrt{z_{s_j}}} \\ &\leq \frac{c+1}{(1-\kappa_{\mathrm{w}})\sqrt{z_{s_\ell}}}.\end{aligned}$$

At the start of s_ℓ the warm-up test still holds. Rearranging it and using the monotonicity of β_s gives

$$\frac{1}{\sqrt{z_{s_\ell}}} < \frac{3B(\delta)\beta_{s_\ell}}{\Gamma_0} \leq \frac{3B(\delta)\beta_S}{\Gamma_0}.$$

For $n_s \leq T$, a stage uses at most $g(m)n_s$ queries times a common polylogarithmic factor in $m, T, 1/\delta$. If $n_s > T$, the stage can use at most $T < n_s$ actual queries. Summing these costs proves the claim for fixed c. □

B.3.8 Main regret theorem

We now state the main regret guarantee. The theorem applies to a generic stage estimator with query cost $q_s = \widetilde{O}(g(m)n_s)$. The two instantiations are (i) coordinate-wise QMC (Lemma 1), with $g(m) = m+1$, and (ii) the multivariate estimator (Theorem 3), with $g(m) = \sqrt{m+1}$ up to logarithmic factors.

Theorem 5 (High-probability safety and regret)**.** *Fix $\delta \in (0,1)$ and run QRS-COLTS for a total of T oracle queries. Assume: (i) $\|\theta^\star\|_2 \leq 1$ and $\|\phi_i^\star\|_2 \leq 1$, (ii) there exists a strictly safe baseline a_{safe} with margin $\Gamma_0 > 0$, and (iii) for each stage s, the QME subroutine outputs $\widehat{\mu}_s$ satisfying $\|\widehat{\mu}_s - \mu(a_s)\|_\infty \leq 1/n_s$ with probability at least $1-\delta_s$, using a* deterministic *number q_s of oracle queries satisfying $q_s = \widetilde{\Theta}(g(m)\, n_s)$ for $g(m) \geq 1$.*

Let $B := B(\delta)$ be as in Lemma 3, and define

$$C_\Delta(B) := \frac{(B+1)(3B-2)}{B(2B-3)} + \frac{3(B+1)}{2B-3} \cdot \frac{\Delta(a_{\mathrm{safe}})}{\Gamma_0}.$$

Then with probability at least $1-3\delta$ the algorithm is safe at every round and

$$\mathrm{Reg}(T) \;=\; \widetilde{O}\Big(g(m)\, B\, (c_{\mathrm{len}}+1)\, C_\Delta(B)\, S^{3/2}\Big) \;+\; \widetilde{O}(T_{\mathrm{warm}} \cdot \Delta(a_{\mathrm{safe}}))\,,$$

where S is the number of stages begun by time T and T_{warm} is bounded by Lemma 11. Moreover, with $d_{\mathcal{A}} = \dim \mathrm{span}(\mathcal{A}) \leq d$,

$$S - 1 \leq \widetilde{O}\Big(\frac{d_{\mathcal{A}}}{\log(1+c_{\mathrm{len}}^2)}\, \log\Big(1 + \frac{T^2}{g(m)^2\, d_{\mathcal{A}}\, \lambda}\Big)\Big) \qquad \textit{(by Lemma 10).}$$

With the deterministic Gaussian draw budget from Lemma 3, $B(\delta) = O(\sqrt{d} + \sqrt{\log(2S_{\max}/\delta)})$. When $d \gtrsim \log(2S_{\max}/\delta)$, this gives $B(\delta) = \Theta(\sqrt{d})$ and the preceding bounds imply

$$\mathrm{Reg}(T) = \widetilde{O}\Big(g(m) d^2 \log^{3/2}(2T/\delta)\, \Big(1 + \frac{\Delta(a_{\mathrm{safe}})}{\Gamma_0}\Big)\Big)\,.$$

Proof. By Lemmas 2, 3 and 5, the respective failure probabilities of confidence, bounded Gaussian norms, and optimism at stages with valid confidence bounds are at most δ, $\delta/2$, and δ. By a union bound, their intersection has probability at least $1 - 5\delta/2 \geq 1 - 3\delta$. On this event, Lemma 6 gives feasibility of every played action, including the action of a partly executed stage.

Lemmas 7 and 9 apply at every stage that reaches the subproblems. They need $B \geq 2$, which holds because $d \geq 1$, $N_{\mathrm{Gauss}} \geq 1$, and $\delta < 1$ give

$$B(\delta) = \sqrt{d} + \sqrt{2\log(2N_{\mathrm{Gauss}}/\delta)} \geq 1 + \sqrt{2\log 2} > 2.$$

They also need a candidate and a step: the successful draw of Lemma 5 has $a^\star$ as a feasible point, so its subproblem has a solution, and by (14) the step $\rho = 0$ passes (1). Hence, on this event, every stage that passes the warm-up test plays a scaled candidate.

Write Q_s for the number of queries actually executed in stage s. A completed stage has $Q_s = q_s$, while the last stage can have $Q_s < q_s$. All queries of a stage use the same action, so if $\mathcal{I}$ indexes the stages that pass the warm-up test, then

$$\mathrm{Reg}(T) = T_{\mathrm{warm}}\Delta(a_{\mathrm{safe}}) + \sum_{s\in\mathcal{I}} Q_s\Delta(a_s).$$

For $s \in \mathcal{I}$, Lemma 9 bounds the gap by $C_\Delta(B)M_s(a_s)$, since the baseline slack is at least Γ_0. In particular the gap is zero when $a_s = 0$. For a nonzero action, rounding the stage length up gives

$$n_s \left\|a_s\right\|_{V_s^{-1}} \leq c_{\mathrm{len}} + \left\|a_s\right\|_{V_s^{-1}} \leq c_{\mathrm{len}} + 1,$$

where the last step uses $V_s \succeq I$ and $\left\|a_s\right\|_2 \leq 1$. Substituting the definition of M_s and applying this ceiling bound yields

$$\begin{aligned} Q_s\Delta(a_s) &\leq C_\Delta(B)B\beta_s Q_s \left\|a_s\right\|_{V_s^{-1}} \\ &\leq C_\Delta(B)B(c_{\mathrm{len}}+1)\beta_s\frac{Q_s}{n_s}. \end{aligned}$$

When $n_s \leq T$, assumption (iii) bounds Q_s/n_s by $\widetilde{O}(g(m))$ with logarithmic factors uniform up to the horizon. When $n_s > T$, we have $Q_s/n_s \leq T/n_s < 1$. Using $g(m) \geq 1$, summing over optimistic stages, and enlarging the sum to all stages gives

$$\begin{aligned} \sum_{s\in\mathcal{I}} Q_s\Delta(a_s) &\leq \widetilde{O}(g(m)C_\Delta(B)B(c_{\mathrm{len}}+1))\sum_{s=1}^{S}\beta_s, \\ \sum_{s=1}^{S}\beta_s &= S\sqrt{\lambda} + \sum_{j=0}^{S-1}\sqrt{j} \\ &\leq S\sqrt{\lambda} + \int_0^S \sqrt{x}\,dx = S\sqrt{\lambda} + \tfrac{2}{3}S^{3/2}. \end{aligned}$$

The equality uses the definition of β_s, and the integral bounds the sum because $\sqrt{x}$ is increasing. With $\lambda = 1$ this gives the claimed optimistic-stage cost, and adding the warm-up contribution proves the first regret bound.

Finally, each of the first $S-1$ stages has used its full budget. Applying the lower query bound in assumption (iii) and then the horizon constraint gives

$$\sum_{s=1}^{S-1} n_s \leq \widetilde{O}\left(\frac{1}{g(m)}\sum_{s=1}^{S-1} Q_s\right) \leq \widetilde{O}\left(\frac{T}{g(m)}\right).$$

For the two concrete estimators, the padded budgets satisfy $q_s \geq g(m)n_s$, so $n_s \leq T$ in every completed stage. For a generic budget satisfying assumption (iii), the same comparison holds up to logarithmic factors. To see this, use $q_s \leq T$, $s \leq T$, and $g(m) \geq 1$ in the lower bound. For fixed constants C, k, it follows that

$$n_s \leq CT\left[\log(2 + m + n_s + T + 1/\delta)\right]^k.$$

If $n_s \le 2 + m + T + 1/\delta$, its logarithm is already bounded. Otherwise the logarithm on the right is at most $\log(2n_s)$. Since $[\log(2n_s)]^k \le \sqrt{n_s}$ for all sufficiently large n_s, the last display gives $n_s \le C^2T^2$ in this case. Hence the logarithms are uniform over all completed stages. Applying Lemma 10 to this completed prefix gives the stated bound on $S-1$. For the dimension-dependent rate, use $d_{\mathcal{A}} \le d$, $S_{\max} = T$, and the simplified coefficients from Lemma 9. The warm-up bound of Lemma 11 contributes only the smaller stage factor $\beta_S = O(\sqrt{S})$. □

Corollary 1 (Optimizing the stage-length parameter)**.** *For fixed $c = c_{\mathrm{len}} > 0$, substituting the stage-count bound into the leading $S^{3/2}$ term in Theorem 5 gives the explicit prefactor*

$$F(c) = \frac{c+1}{[\log(1+c^2)]^{3/2}}.$$

This factor is minimized at $c \approx 5.48$, where $F(c) \approx 1.02$.

Proof. The numerator comes from rounding the stage length up, and the denominator from the power $3/2$ of the stage count, which is proportional to $1/\log(1+c^2)$. Write $\Lambda = \log(1+c^2)$. Logarithmic differentiation gives

$$\frac{F'(c)}{F(c)} = \frac{1}{1+c} - \frac{3c}{(1+c^2)\Lambda}.$$

Since the denominators are positive, the sign of $F'(c)$ equals the sign of $H(c) = (1+c^2)\Lambda - 3c(1+c)$, whose first two derivatives are

$$H'(c) = 2c\Lambda - 4c - 3,$$
$$H''(c) = 2\Lambda + \frac{4c^2}{1+c^2} - 4.$$

Both nonconstant terms of H'' increase, so H'' changes sign once, from negative to positive. Hence H' decreases from $H'(0) = -3$ and then increases to infinity, so it also changes sign once, and H decreases from $H(0) = 0$ and then increases to infinity. The unique positive zero of H therefore minimizes F, and solving $H(c) = 0$ numerically gives the values in the statement. □

B.3.9 Instantiations and quantum advantages

We now instantiate the generic regret theorem (Theorem 5) with concrete quantum mean-estimation routines. In both cases, the proof is an application of Theorem 5 once we verify its stage-estimator assumption (iii) and identify the corresponding function $g(m)$.

Corollary 2 (Coordinate-wise QMC instantiation)**.** *Using coordinate-wise QMC (Lemma 1) on all $m+1$ coordinates in each stage yields $g(m) = m+1$ and thus*

$$\mathrm{Reg}(T) = \widetilde{O}\left((m+1)d^2 \log^{3/2}(T/\delta)\left(1 + \frac{\Delta(a_{\mathrm{safe}})}{\Gamma_0}\right)\right).$$

Proof. Condition on the history at the start of stage s, including the choice of a_s. Write $M = m+1$ and let $\mu_{s,j}$ denote the mean of feedback coordinate j. For each coordinate, use the scalar estimator with accuracy $1/n_s$ and failure probability δ_s/M. By Lemma 1,

$$\mathbb{P}(|\widehat{\mu}_{s,j} - \mu_{s,j}| > 1/n_s) \le \delta_s/M.$$

The vector estimate fails exactly when some coordinate fails, so a union bound gives

$$\begin{aligned}\mathbb{P}(\|\widehat{\mu}_s - \mu(a_s)\|_\infty > 1/n_s) &\le \sum_{j=1}^{M} \mathbb{P}(|\widehat{\mu}_{s,j} - \mu_{s,j}| > 1/n_s) \\ &\le M\,(\delta_s/M) = \delta_s.\end{aligned}$$

Each scalar routine uses at most $C_{\mathrm{qmc}} n_s \log(2M/\delta_s)$ queries. Summing the M budgets gives

$$q_s = O(M n_s \log(2M/\delta_s)).$$

We fix a deterministic integer budget of this order and at least Mn_s, and spend any unused calls on a fresh register at the same action. This verifies Assumption 1 with $g(m) = m+1$, and Theorem 5 gives the stated regret bound. □

Corollary 3 (Bounded multivariate QME instantiation)**.** *Using the bounded multivariate QME routine of Theorem 3 yields $g(m) = \sqrt{m+1}$ up to additional polylogarithmic factors, and thus*

$$\mathrm{Reg}(T) = \widetilde{O}\Big(\sqrt{m+1}\; d^2 \log^{3/2}(T/\delta)\Big(1 + \frac{\Delta(a_{\mathrm{safe}})}{\Gamma_0}\Big)\Big).$$

Proof. We prove Theorem 3 by applying QBOUNDED [CHJ22, Theorem 3.3] to a rescaled copy of Z. In the oracle model of Definition 4, fix a and regard the basis pair $(\omega, Z(\omega))$ as the sample label, so that $\mathcal{O}_a$ and $\mathcal{O}_a^\dagger$ implement their quantum experiment. Their binary value oracle reads the stored value and XORs a scaled copy into a fresh register by reversible arithmetic, with no further feedback queries, also in inverse and controlled form. A controlled reflection about the prepared state, $\mathcal{O}_a(I - 2\,|0\rangle\langle 0|)\mathcal{O}_a^\dagger$, needs control only on the middle reflection, so it too uses ordinary calls to $\mathcal{O}_a$ and $\mathcal{O}_a^\dagger$.

Specifically, let $D = 2^{\lceil \log_2 \sqrt{M} \rceil}$ and $Z' = Z/D$; dividing a binary value by this power of two is a bit shift. Since $\|Z\|_2 \le \sqrt{M} \le D$, we have $\|Z'\|_2 \le 1$ and $\mathbb{E}\|Z'\|_2 \le 1$. We can thus take $L_2 = 1$ in the cited theorem. Choose

$$n = \lceil D \log(2M/\delta)/\varepsilon \rceil .$$

With probability at least $1 - \delta$, the estimate satisfies

$$\begin{aligned} \|D\widehat{\mu}_{Z'} - \mathbb{E}Z\|_\infty &= D\|\widehat{\mu}_{Z'} - \mathbb{E}Z'\|_\infty \\ &\le \frac{D \log(M/\delta)}{n} \le \varepsilon. \end{aligned}$$

The equality rescales the mean. The first inequality applies the simultaneous coordinate-wise guarantee of QBOUNDED, and the last uses the choice of n. The algorithm uses $\widetilde{O}(n)$ calls to the quantum experiment and its binary value oracle. Only the former require calls to the feedback sampler, giving $\widetilde{O}(\sqrt{M}/\varepsilon)$ sampler queries since $D < 2\sqrt{M}$.

Finally, condition on the stage history to fix a_s and apply the result to $Z = X(a_s)$ with $M = m + 1$, $\varepsilon = 1/n_s$, and $\delta = \delta_s$. The accuracy guarantee and the stage budget become

$$\begin{aligned} \mathbb{P}(\|\widehat{\mu}_s - \mu(a_s)\|_\infty > 1/n_s) \le \delta_s, \\ q_s = \sqrt{m+1}\, n_s \;\mathrm{polylog}(2 + m + n_s + 1/\delta_s). \end{aligned}$$

As for the scalar estimators, we fix the budget before the stage and pad at a_s, taking $q_s \ge \sqrt{m+1}\, n_s$. This verifies Assumption 1 with $g(m) = \sqrt{m+1}$. For a completed stage $n_s \le q_s \le T$ and $1/\delta_s = \pi^2 s^2/(6\delta)$, so only logarithms in $m, T, 1/\delta$ are suppressed in Theorem 5. Substituting this $g(m)$ gives the claimed regret. □

C Query-weighted safe-action bomb model: algorithm and proofs

This appendix proves the bomb-certified safe-optimization result stated informally in Theorem 2. The learner can query actions in superposition and receives a reward qsample and the bomb flag. The learner then need not estimate the constraint matrix; in exchange, regret must be query-weighted and the explosion probability controlled.

C.1 Overview and main theorem

We next consider the coherent-action model of Definition 3, in which a query may place amplitude on many actions and the oracle flags unsafe actions exactly. Regret is therefore query-weighted (Definition 7), and the explosion probability is controlled separately.

The bomb flag gives membership access to $\mathcal{S}$, so the learner estimates only the reward vector. In each stage it maximizes K_s perturbed linear objectives over $\mathcal{S}$, splitting its explosion and failure budgets equally among the calls, and keeps the candidate with the largest perturbed value. The subsequent reward queries at this safe candidate have zero explosion probability.

Theorem 6 (Bomb-certified reward-only QLinTS, informal)**.** *Suppose $\|\theta^\star\|_2 \le 1$ and $\|a\|_2 \le 1$ for $a \in \mathcal{A}$, and that $\mathcal{S}$ satisfies Assumption 2 with $\mathbb{B}(a_{\mathrm{safe}}, r) \subseteq \mathcal{S} \subseteq \mathbb{B}(a_{\mathrm{safe}}, R)$. Then BCO-QLinTS has explosion probability at most ε, and with probability at least $1-\delta$ its regret up to termination satisfies*

$$R_{\mathrm{w}}(T) \le \widetilde{O}\Big(d^2 \log^{3/2}(2T/\delta)\Big) + \widetilde{O}\left(\frac{d^4 \log^2(2T)}{\varepsilon} \operatorname{polylog}(2 + R/r, 2/\delta)\right).$$

The first term is the reward-learning term; the second is the bomb-safe optimization overhead. The formal statement is Theorem 8. The constraints enter the bound only through R/r, because the bomb flag reports feasibility directly.

While QRS-COLTS estimates $\Phi^\star$ from numerical constraint feedback, BCO-QLinTS optimizes directly over $\mathcal{S}$ using the safety flag. Note, however, that because the two models evaluate performance under different loss functions, their regret bounds are not directly comparable.

C.2 Model and query-weighted regret

We use the model of Section A.3: the safe set $\mathcal{S}$, the clipped gap $\Delta_+(a) = (\mu^\star - \langle \theta^\star, a\rangle)_+$, the safe-action bomb oracle of Definition 6, and the query-weighted loss $L(\pi_t)$ and regret $R_{\mathrm{w}}(T)$ of Definition 7. A query with action marginal π_t explodes with probability $\pi_t(\mathcal{A} \setminus \mathcal{S})$, which we track separately from regret.

C.3 Bomb-certified safe optimization

BCO-QLinTS optimizes over the true safe set with the bomb flag, which requires the safe set to be well rounded.

Assumption 2 (Well-rounded safe set)**.** *The safe set is a closed convex body. There are known radii $0 < r \le R$ and a known safe point a_{safe} such that*

$$\mathbb{B}(a_{\mathrm{safe}}, r) \subseteq \mathcal{S} \subseteq \mathbb{B}(a_{\mathrm{safe}}, R).$$

If the ambient optimization grid includes labels outside the known action domain $\mathcal{A}$, we assume that membership in $\mathcal{A}$ can be computed reversibly on these labels.

We use the following standard quantum convex-optimization primitive, stated in a form suitable for our application. It follows from quantum convex optimization using membership oracles [vAGGdW20].

Lemma 12 (Quantum convex optimization with membership queries)**.** *Under Assumption 2, suppose the learner has coherent membership access to $\mathcal{S}$, i.e.,*

$$|a\rangle\,|b\rangle \mapsto |a\rangle\,|b \oplus \mathbf{1}\{a \notin \mathcal{S}\}\rangle\,.$$

For any linear objective $c \in \mathbb{R}^d$, accuracy $\gamma \in (0,1)$, and failure probability $\delta \in (0,1)$, there is a quantum algorithm that outputs $\widehat{a} \in \mathcal{S}$ satisfying

$$c^\top \widehat{a} \ge \max_{a \in \mathcal{S}} c^\top a - \gamma$$

with probability at least $1-\delta$, using

$$M_{\mathrm{qco}}(d, R/r, \gamma, \delta) = \widetilde{O}\left(d \operatorname{polylog}\left(\frac{R}{r}, \frac{1}{\gamma}, \frac{1}{\delta}\right)\right)$$

coherent membership queries. The $\widetilde{O}$ hides logarithmic factors in d and $2 + R\|c\|_2/\gamma$.

Proof sketch. Write $u = a_{\mathrm{safe}}$; if $c = 0$, returning u suffices. Otherwise let $v = c/\|c\|_2$ and, for $0 < \tau \le r/2$, run the weak optimizer of [vAGGdW20, Definition 5] with tolerance $\tau/2$ in the original coordinates; its inner comparison set contains u and is therefore nonempty. On success, its output y_0 lies within $\tau/2$ of $\mathcal{S}$. Rounding y_0 to a grid point y with Euclidean error at most $\tau/4$ gives, by the triangle inequality and $\|v\|_2 = 1$,

$$\begin{aligned} &\operatorname{dist}(y, \mathcal{S}) \le \tau/2 + \tau/4 \le \tau, \\ &v^\top x - v^\top y \le \tau/2 + \tau/4 \le \tau \qquad \text{if } \mathbb{B}(x, \tau) \subseteq \mathcal{S}. \end{aligned}$$

The second line uses the weak objective guarantee at y_0, since each such x also has its $\tau/2$-ball in $\mathcal{S}$; the tolerance $\tau/2$ leaves room for the rounding error $\tau/4$ in both lines.

To obtain an exactly feasible output, take $z \in \mathcal{S}$ with $\|y - z\|_2 \le \tau$, and let $\sigma = r/(r + 2\tau)$ and $p = \sigma y + (1 - \sigma)u$. For every $\|h\|_2 \le \sigma\tau$,

$$p + h = \sigma z + (1 - \sigma)\left(u + \frac{\sigma(y - z) + h}{1 - \sigma}\right),$$
$$\left\|\frac{\sigma(y - z) + h}{1 - \sigma}\right\|_2 \le \frac{2\sigma\tau}{1 - \sigma} = r.$$

The point in parentheses lies in the inner ball $\mathbb{B}(u, r)$, so convexity gives $\mathbb{B}(p, \sigma\tau) \subseteq \mathcal{S}$. Since $\sigma \ge 1/2$, rounding p to a representable $\widehat{a}$ within $\tau/4 < \sigma\tau$ preserves feasibility. The point z is used only in the analysis.

For the objective comparison, let $x^\star$ maximize $v^\top x$ on $\mathcal{S}$ and put $x' = (1 - \tau/r)x^\star + (\tau/r)u$. Convexity gives $\mathbb{B}(x', \tau) \subseteq \mathcal{S}$ by mixing $x^\star$ with every point of $\mathbb{B}(u, r)$. Hence

$$v^\top(x^\star - \widehat{a}) \le \|x^\star - x'\|_2 + v^\top(x' - y) + \|y - p\|_2 + \|p - \widehat{a}\|_2$$
$$\le \frac{R\tau}{r} + \tau + \frac{2\tau(R + \tau)}{r + 2\tau} + \frac{\tau}{4} \le 4\tau(1 + R/r).$$

The four terms come from contracting the comparator, the weak optimization, contracting the output, and the final rounding; the third uses $\|y - u\|_2 \le R + \tau$. Taking

$$\tau = \min\left\{\frac{r}{2}, \frac{\gamma}{4(1 + R/r)\|c\|_2}\right\}$$

gives the claimed objective error after multiplication by $\|c\|_2$. Translating by u and scaling by R reduces the body to outer radius one. The membership-query bound of [vAGGdW20] then depends logarithmically on the inverse normalized tolerance $2R/\tau$ and on the confidence parameter. □

In the bomb model every membership query risks an explosion, so we use the bomb-query framework of Lin and Lin [LL15]. To make the reduction explicit, we first recall their definition and simulation result.

Definition 10 (Bomb query to a Boolean predicate)**.** *Let $h : \mathcal{X} \to \{0, 1\}$ be a Boolean predicate. A coherent query to h would apply*

$$O_h : |x\rangle\,|b\rangle \mapsto |x\rangle\,|b \oplus h(x)\rangle\,.$$

A bomb query is the corresponding dangerous query: the algorithm may put amplitude on inputs x, but whenever the queried branch has $h(x) = 1$, the bomb may explode. In the circuit formulation of Lin and Lin, a bomb query is implemented by a controlled query followed by an immediate measurement of the answer register; outcome 1 is the explosion event. A bomb-query algorithm is required to keep the total probability of explosion below a given budget.

Theorem 7 (Lin–Lin bomb-query simulation)**.** *Let $\mathcal{Q}$ be a bounded-error quantum query algorithm for a Boolean-input problem that uses Q ordinary coherent queries to a Boolean oracle O_h. For any explosion budget $\eta \in (0, 1)$ and target accuracy $\delta \in (0, 1)$, there is a bomb-query algorithm that simulates $\mathcal{Q}$ with explosion probability at most η and error $O(\delta)$ in the surviving subnormalized state, using*

$$\widetilde{O}\left(\frac{Q^2}{\min\{\eta, \delta\}}\right)$$

bomb queries, where the $\widetilde{O}(\cdot)$ hides logarithmic factors in Q, $1/\eta$, and $1/\delta$.

Proof. The query simulation in the proof of [LL15, Theorem 4.2] uses $O(L)$ bomb queries per ordinary query. Its explosion probability and its Euclidean error on the surviving, subnormalized state are each $O(1/L)$ per simulated query. The construction approximates the query operation on arbitrary inputs, so it also applies when the final output encodes a point. A union bound controls explosions, while a telescoping sum controls state error because both the intervening unitaries and the survival projections have operator norm at most one. Consequently,

$$\mathbb{P}(\text{explosion}) = O(Q/L), \qquad \|\psi_{\text{surv}} - \psi_{\text{ideal}}\|_2 = O(Q/L).$$

Taking L to be a sufficiently large constant times $Q/\min\{\eta, \delta\}$ gives both guarantees at the claimed cost. □

To implement this simulation, we extend the unsafe predicate to all grid labels used by the convex optimizer. The sampling oracle is defined only on $\mathcal{A}$, and membership in $\mathcal{A}$ is computable reversibly (Assumption 2), so a membership query on a label outside $\mathcal{A}$ writes the bit 1 locally. Only the predicate $\mathbf{1}\{a \in \mathcal{A} \setminus \mathcal{S}\}$ is unknown, and because the two predicates have disjoint supports, XORing their answers implements the required bit oracle.

The Lin–Lin simulation needs controlled bomb queries only for this unknown predicate, while the oracle of Definition 6 has no control register. We therefore copy the action into a second register on active branches with $a \in \mathcal{A}$ and write a_{safe} into it on all other branches, leaving the original action and control registers unchanged. A forward call at the copied action, followed by measurement of its flag, can explode only on active branches with $a \in \mathcal{A} \setminus \mathcal{S}$. On the surviving branches the copied action is safe, and an inverse call at the same action returns the reward register to $|0\rangle$ with certainty; undoing the copy then restores the original registers. The two calls thus implement the required controlled bomb query, and both count toward the query budget.

Lemma 13 (Bomb-certified safe optimization)**.** *Fix objective c, accuracy γ, failure probability δ, and explosion budget η. There is a bomb-query implementation of the optimizer in Lemma 12 with explosion probability at most η. Every point $\widehat{a}$ it returns lies in $\mathcal{S}$ and, conditional on a return, satisfies the following with probability at least $1-\delta$:*

$$c^\top \widehat{a} \geq \max_{a \in \mathcal{S}} c^\top a - \gamma.$$

It uses

$$B_{\mathrm{opt}}(d, R/r, \gamma, \delta, \eta) = \widetilde{O}\left(\frac{M_{\mathrm{qco}}(d, R/r, \gamma, \delta)^2}{\eta}\right)$$

bomb queries.

Proof. Run the ordinary optimizer of Lemma 12 at failure probability $1/10$, and denote its query count by Q. We simulate this routine in $k = \lceil \log_2(4/\delta) \rceil$ fresh trials, stopping at the first explosion. Each trial uses explosion and state-error parameters $\eta' = c_0\eta/k$ in Theorem 7. Let P project onto outputs encoding feasible, γ-optimal points. Then $\|P\psi_{\mathrm{surv}}\|_2^2$ is the probability of surviving and producing such a point. For a sufficiently small absolute $c_0 \leq 1/4$,

$$\left|\|P\psi_{\mathrm{surv}}\|_2^2 - \|P\psi_{\mathrm{ideal}}\|_2^2\right| \leq 2\|\psi_{\mathrm{surv}} - \psi_{\mathrm{ideal}}\|_2 = O(\eta'),$$

$$\mathbb{P}(\text{survive and produce a feasible, } \gamma\text{-optimal candidate}) \geq 9/10 - O(\eta') \geq 1/2.$$

The first line holds because both states and P have norm at most one, and the second uses the success probability $9/10$ of the ordinary optimizer.

A failed simulation can output an infeasible candidate. We discard invalid encodings and points outside $\mathcal{A}$ locally. For each remaining candidate a, the repeated-rotation bomb test [EV93, LL15] starts a control qubit in $|0\rangle$. In each of $N = \lceil \pi^2 k/(2\eta) \rceil$ rounds, it rotates the qubit by $\vartheta = \pi/(2N)$ and queries the bomb at a on control 1. For safe a, every query survives and the rotations accumulate to $\pi/2$, so the test accepts with certainty. For unsafe a, surviving a round projects the control back to $|0\rangle$. The test therefore never accepts such an action, and its explosion probability satisfies

$$1 - \cos^{2N}\vartheta \leq N(1 - \cos^2\vartheta) \leq N\vartheta^2 = \frac{\pi^2}{4N} \leq \frac{\eta}{2k}.$$

The first inequality is a union bound over the rounds, and the next uses $|\sin\vartheta| \leq \vartheta$.

After the trials, the routine returns the action with the largest value $c^\top a$ among the verified candidates and a_{safe}, so every return is safe. Given any history that reaches a trial, the trial produces and verifies a feasible, γ-optimal candidate with probability at least $1/2$, and an explosion ends the routine without a return. Induction over the trials and a union bound over their explosion budgets give

$$\mathbb{P}(\text{return a point with objective gap greater than } \gamma) \leq 2^{-k} \leq \delta/4,$$

$$\mathbb{P}(\text{explode}) \leq k\left(\eta' + \frac{\eta}{2k}\right) \leq 3\eta/4.$$

Since $\eta \leq 1$, the routine returns with probability at least $1/4$. Dividing the first probability by this lower bound shows that the conditional probability of a gap greater than γ given a return is at most δ.

The k simulations each cost $\widetilde{O}(Q^2/\eta')$ queries, and the k tests each cost $O(N)$ queries. Their sum is $\widetilde{O}((Q^2+1)k^2/\eta)$. Absorbing the logarithms in $1/\delta$ yields the stated bound. If $c = 0$, the baseline suffices without any queries. □

Algorithm 2 BCO-QLinTS with resampled reward directions

Input: Query budget T, failure probability δ, explosion budget ε, known safe point a_{safe} and radii r, R.

Initialize $V_1 = I$ and $s = 1$. Use $S_{\max}$ from (21) below. Stop immediately on explosion or budget exhaustion, including during any subroutine.

while queries remain **do**

  Set $\widehat{\theta}_s = V_s^{-1} \sum_{j<s} q_j^2 a_j \widehat{r}_j$, $\eta_s = \varepsilon / S_{\max}$, $\delta_s = \delta/(2S_{\max})$, $\gamma_s = T^{-2}$ and K_s as in (4).

  Draw K_s independent standard Gaussian vectors $\zeta_s^{(k)} \in \mathbb{R}^d$ and form $\widetilde{\theta}_s^{(k)} = \widehat{\theta}_s + \beta_s V_s^{-1/2} \zeta_s^{(k)}$.

  For each $k = 1, \dots, K_s$, let $a_s^{(k)}$ be the action returned by the routine of Lemma 13 with objective $\widetilde{\theta}_s^{(k)}$, accuracy γ_s, failure probability δ_s/K_s and explosion budget η_s/K_s.

  Choose $k^\star$ maximizing $\left\langle \widetilde{\theta}_s^{(k)}, a_s^{(k)} \right\rangle$ (ties go to the smallest index). Set $a_s = a_s^{(k^\star)}$, $\widetilde{\theta}_s = \widetilde{\theta}_s^{(k^\star)}$ and $w_s = \|a_s\|_{V_s^{-1}}$.

  If $w_s = 0$, spend the remaining queries at a_s and stop. Otherwise $q_s \leftarrow \lceil c_{\mathrm{len}}/w_s \rceil$; estimate the reward at a_s with q_s queries (estimator described in the text).

  After all q_s queries, set $V_{s+1} = V_s + q_s^2 a_s a_s^\top$, record the estimate $\widehat{r}_s$, and increment s.

end while

C.4 BCO-QLinTS algorithm

A single Gaussian perturbation is optimistic at $a^\star$ only with constant probability, so each stage draws K_s perturbed reward vectors and compares the values $\langle \widetilde{\theta}_s^{(k)}, a_s^{(k)} \rangle$ at the actions returned for them. The proof uses only the largest of these values; the perturbation that attains it need not be optimistic at $a^\star$. Fix $c_{\mathrm{len}} > 0$ and $T \geq 2$ (for $T = 1$, one query at a_{safe} has regret at most one), and define

$$\nu_s = \delta/(\pi^2 s^2), \qquad L_s = \max\{1, C_{\mathrm{qmc}}\} \log(2/\nu_s), \qquad \beta_s = 1 + L_s \sqrt{s-1}.$$

Here C_{qmc} is the query constant in Lemma 1.

Write $\mu_s = \langle \theta^\star, a_s \rangle$. If $q_s > L_s$, the reward estimator runs the scalar QME of Lemma 1 with accuracy L_s/q_s and failure probability ν_s, which needs at most

$$C_{\mathrm{qmc}} \frac{q_s}{L_s} \log(2/\nu_s) \leq q_s$$

queries; it clips the estimate to $[0, 1]$ and spends any unused calls at a_s. If $q_s \leq L_s$, it averages q_s measured rewards, whose error is at most $1 \leq L_s/q_s$ because the estimate and μ_s both lie in $[0, 1]$. In both cases every completed reward stage satisfies

$$\mathbb{P}(|\widehat{r}_s - \mu_s| > L_s/q_s \mid \text{history before reward estimation}) \leq \nu_s.$$

A partial reward stage contributes no regression observation.

C.5 Regret analysis

Proposition 1 (Reward confidence and optimism under resampling)**.** *Assume $\|\theta^\star\|_2 \leq 1$ and $\|a\|_2 \leq 1$ on $\mathcal{A}$. Write $N = \sum_{s \leq S_{\max}} K_s$ and $B = \sqrt{d} + \sqrt{2 \log(6N/\delta)}$. With probability at least $1 - \delta/2$, every sampled reward vector satisfies*

$$\|\widetilde{\theta}_s^{(k)} - \theta^\star\|_{V_s} \leq (B+1)\beta_s, \qquad \|\widetilde{\theta}_s^{(k)}\|_2 \leq 1 + (B+1)\beta_s.$$

On this same event, in every stage whose K_s calls return γ_s-optimal actions, the selected pair satisfies $\left\langle \widetilde{\theta}_s, a_s \right\rangle \geq \mu^\star - \gamma_s$.

Proof. Conditional on the history before each reward routine, a completed estimate misses its tolerance with probability at most ν_s, because returning an inaccurate estimate by the horizon implies the failure event of the uninterrupted routine on the same randomness. A union bound over stages bounds the total failure probability by $\sum_{s \geq 1} \nu_s = \delta/6$. For the completed stages $i < s$, stack the

rows $q_i a_i^\top$ in X and let $\xi_i = q_i(\widehat{r}_i - \langle \theta^\star, a_i \rangle)$. Then $V_s = I + X^\top X$, and the ridge normal equations, $X^\top X \preceq V_s$, $I \preceq V_s$ and $|\xi_i| \le L_i \le L_s$ give

$$\begin{aligned} V_s(\widehat{\theta}_s - \theta^\star) &= X^\top \xi - \theta^\star, \\ \|\widehat{\theta}_s - \theta^\star\|_{V_s} &\le \|V_s^{-1/2} X^\top \xi\|_2 + \|V_s^{-1/2}\theta^\star\|_2 \\ &\le \|\xi\|_2 + 1 \le L_s\sqrt{s-1} + 1 = \beta_s. \end{aligned}$$

There are at most N draws by the deterministic stage count proved below. Each fresh Gaussian has norm exceeding B with probability at most $\delta/(6N)$, so all norms satisfy the bound except with probability $\delta/6$. Since $\|V_s^{-1/2}\zeta\|_{V_s} = \|\zeta\|_2$, the triangle inequality gives the first claimed norm bound. The second follows from $V_s \succeq I$ and $\|\theta^\star\|_2 \le 1$.

Fix a history at the start of a reached stage for which the ridge bound holds. For $a^\star \neq 0$, put $u = V_s^{-1/2} a^\star / \|a^\star\|_{V_s^{-1}}$. The projection $u^\top \zeta_s^{(k)}$ is standard normal. When this projection is at least one, the ridge bound gives

$$\left\langle \widetilde{\theta}_s^{(k)}, a^\star \right\rangle = \left\langle \widehat{\theta}_s, a^\star \right\rangle + \beta_s \|a^\star\|_{V_s^{-1}} u^\top \zeta_s^{(k)} \ge \langle \theta^\star, a^\star \rangle .$$

For $a^\star = 0$ equality holds for every direction. Independence of the draws bounds the conditional probability of missing optimism by $(1-\pi_{\mathrm{opt}})^{K_s} \le e^{-\pi_{\mathrm{opt}} K_s} \le \nu_s$. Averaging this conditional bound over histories that reach stage s with valid confidence bounds gives probability at most ν_s of reaching such a stage and missing optimism there. Summing over stages gives $\sum_{s \ge 1} \nu_s = \delta/6$. On the intersection of the three events, let k_0 be an optimistic draw. If every call returns a γ_s-optimal action, then the selection rule, the comparison of $a_s^{(k_0)}$ with the feasible point $a^\star$, and the optimism of draw k_0 give

$$\begin{aligned} \left\langle \widetilde{\theta}_s, a_s \right\rangle &\ge \left\langle \widetilde{\theta}_s^{(k_0)}, a_s^{(k_0)} \right\rangle \\ &\ge \left\langle \widetilde{\theta}_s^{(k_0)}, a^\star \right\rangle - \gamma_s \ge \mu^\star - \gamma_s. \end{aligned}$$

□

Lemma 14 (Stage gap bound)**.** *On the event of Proposition 1, if every optimizer call of stage s returns a γ_s-optimal action, then $\Delta_+(a_s) \le (B+1)\beta_s w_s + \gamma_s$.*

Proof. For the selected pair, the proposition and Cauchy–Schwarz in the design norm give

$$\begin{aligned} \mu^\star - \langle \theta^\star, a_s \rangle &\le \gamma_s + \left\langle \widetilde{\theta}_s - \theta^\star, a_s \right\rangle \\ &\le \gamma_s + \|\widetilde{\theta}_s - \theta^\star\|_{V_s} \|a_s\|_{V_s^{-1}} \le \gamma_s + (B+1)\beta_s w_s. \end{aligned}$$

The right-hand side is nonnegative, so it also bounds $\Delta_+(a_s)$. □

Theorem 8 (End-to-end bomb-certified safe optimization bound)**.** *Suppose Assumption 2 holds, $\|a\|_2 \le 1$ on $\mathcal{A}$, and $\|\theta^\star\|_2 \le 1$. Run Algorithm 2 with fixed $c_{\mathrm{len}} > 0$ and*

$$S_{\max} = 1 + \left\lceil \frac{d \log(1 + T^2/d)}{\log(1 + c_{\mathrm{len}}^2)} \right\rceil . \tag{21}$$

The explosion probability is at most ε. With probability at least $1-\delta$, the query-weighted regret up to termination satisfies

$$\begin{aligned} R_{\mathrm{w}}(T) &\le (c_{\mathrm{len}} + 1)(B+1) \sum_{s=1}^{S_{\max}} \beta_s + T^{-1} \\ &\quad + \sum_{s=1}^{S_{\max}} K_s \, B_{\mathrm{opt}}(d, R/r, \gamma_s, \delta_s/K_s, \eta_s/K_s) \\ &\le \widetilde{O}\Big(d^2 \log^{3/2}(2T/\delta)\Big) + \widetilde{O}\left(\frac{d^4 \log^2(2T)}{\varepsilon} \operatorname{polylog}(2 + R/r, 2/\delta) \right). \end{aligned}$$

Here B_{opt} is evaluated at the largest objective norm allowed by Proposition 1. The constraints enter the bound only through R/r.

Proof. If a reward stage updates the design, the determinant lemma and $q_s w_s \geq c_{\mathrm{len}}$ give

$$\det V_{s+1} / \det V_s = 1 + q_s^2 a_s^\top V_s^{-1} a_s \geq 1 + c_{\mathrm{len}}^2.$$

After n updates, $\sum_{s \leq n} q_s \leq T$. Multiplying these ratios from $\det V_1 = 1$, and using $\|a_s\|_2 \leq 1$ and the arithmetic–geometric mean inequality, gives

$$\operatorname{tr} V_{n+1} = d + \sum_{s \leq n} q_s^2 \|a_s\|_2^2 \leq d + T^2,$$

$$n \log(1 + c_{\mathrm{len}}^2) \leq \log \det V_{n+1} \leq d \log(\operatorname{tr} V_{n+1}/d) \leq d \log(1 + T^2/d).$$

Every stage except possibly the last updates the design; the last may end by an explosion, at the horizon, or with $w_s = 0$. Thus (21) bounds the number of stages on every path.

Conditional on the history at the start of each reached optimizer call, Lemma 13 bounds its explosion probability by η_s/K_s and its probability of returning a suboptimality gap greater than γ_s by δ_s/K_s. Cutting off a call at the horizon can only remove subsequent explosions and bad returns. Every returned action is safe, so reward queries cannot explode. Summing these conditional bounds gives

$$\mathbb{P}(\text{an explosion}) \leq \sum_{s \leq S_{\max}} K_s(\eta_s/K_s) = \varepsilon,$$

$$\mathbb{P}(\text{any bad optimizer return}) \leq \sum_{s \leq S_{\max}} K_s(\delta_s/K_s) = \delta/2.$$

Work on the intersection of the event of no bad returns with the event of Proposition 1; it has probability at least $1 - \delta$ and includes paths that stop early. Each optimizer query, including a terminating one, incurs loss at most one, and a stage that ends before all its calls return incurs no other loss. In a stage that reaches reward queries, let ℓ_s be their number. If $w_s > 0$, then $\ell_s \leq q_s$ and, since $V_s \succeq I$, $w_s \leq 1$. The ceiling defining q_s and Lemma 14 then give

$$\ell_s w_s \leq q_s w_s \leq c_{\mathrm{len}} + w_s \leq c_{\mathrm{len}} + 1,$$

$$\ell_s \Delta_+(a_s) \leq (c_{\mathrm{len}} + 1)(B + 1)\beta_s + \ell_s \gamma_s.$$

The second line also holds at $w_s = 0$, where the gap is at most γ_s. Summing over stages and using $\sum_s \ell_s \gamma_s \leq T^{-1}$ proves the first regret inequality.

Write $S = S_{\max}$ and $K = K_S$. The schedules give $S = O(d \log(2T))$, $K = O(\log(2S/\delta))$ and $N \leq SK$. Monotonicity of L_s and an integral bound on $\sum_{j=0}^{S-1} \sqrt{j}$ yield

$$\sum_{s=1}^{S} \beta_s \leq S + \tfrac{2}{3} L_S S^{3/2},$$

$$(B+1) \sum_{s=1}^{S} \beta_s \leq O\Big((1 + \sqrt{d} + \sqrt{\log(6SK/\delta)})(S + L_S S^{3/2})\Big)$$

$$\leq \widetilde{O}\Big(d^2 \log^{3/2}(2T/\delta)\Big).$$

In the last step $L_S = O(\log(2S/\delta))$. The suppressed factors here are logarithmic in $2d$, $2/\delta$, and $\log(2T)$.

For the optimizer cost, we may use outer radius $\min(R, 2)$ because both a_{safe} and every feasible action have norm at most one. Hence $\min(R, 2)\, \|\widetilde{\theta}_s^{(k)}\|_2 \leq 2[1 + (B+1)\beta_S]$ on the same event, and at $\gamma_s = T^{-2}$ Lemma 12 uses $\widetilde{O}(d)$ membership queries, up to the logarithmic factors stated there. Put $H = \operatorname{polylog}(2 + R/r, 2/\delta)$. The inverse explosion-budget dependence in Lemma 13 gives

$$\sum_{s=1}^{S} K_s B_{\mathrm{opt}}(d, R/r, \gamma_s, \delta_s/K_s, \eta_s/K_s) \leq \widetilde{O}\left(d^2 H \sum_{s=1}^{S} K_s^2/\eta_s\right)$$

$$\leq \widetilde{O}\big(d^2 H K^2 S^2/\varepsilon\big) \leq \widetilde{O}\big(d^4 H \log^2(2T)/\varepsilon\big).$$

One factor K_s counts calls and the other comes from their budgets η_s/K_s. The middle inequality substitutes $\eta_s = \varepsilon/S$ and $K_s \leq K$, and the last uses the stage and resampling bounds. Intersecting the regret event with survival yields the full-horizon bound with probability at least $1 - \delta - \varepsilon$. □

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope: quantum feedback models for safe stochastic linear bandits, QRS-COLTS, and the safe-action bomb model. These claims are matched by the formal model definitions, algorithms, and regret/safety theorems in the main text and appendix.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper includes a dedicated discussion of limitations and further directions, including that the guarantees are query-complexity guarantees, do not accelerate classical optimization subroutines, and rely on mean-constraint safety rather than almost-sure sample-wise safety.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The paper is primarily theoretical and states the assumptions, models, algorithms, and theorem conditions explicitly. Complete proofs are provided in the appendix for the QRS-COLTS and bomb-certified safe optimization guarantees.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [N/A]

   Justification: The paper does not report empirical experiments; the main claims are theoretical safety and regret guarantees.

   Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [N/A]

   Justification: The paper does not include experiments requiring code or data to reproduce empirical results. Reproducibility is provided through formal algorithm descriptions and proofs rather than an experimental code release.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [N/A]

   Justification: The paper does not include training or test experiments, datasets, hyperparameters, or optimizer settings.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [N/A]

   Justification: The paper does not include empirical experiments, so error bars or statistical significance tests are not applicable.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [N/A]

   Justification: The paper does not include empirical experiments requiring compute-resource reporting.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work is theoretical and does not involve human subjects, private data, scraped datasets, or deployed systems. To the best of our knowledge, it conforms to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: The introduction motivates potential positive applications of safe learning, such as clinical trials, resource-constrained decision-making, fairness-related allocation, robotics, and control. The paper also clarifies the scope of its safety notion and its limitations; no direct high-risk deployment or released system is proposed.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper does not release data, pretrained models, image generators, language models, or other high-risk assets requiring release safeguards.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [N/A]

Justification: The paper does not use experimental datasets, pretrained models, or code assets as part of an empirical study. Prior work and technical primitives are credited through citations.

Guidelines:

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

13. **New assets**

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

Answer: [N/A]

Justification: The paper does not introduce or release new datasets, code packages, pretrained models, or other external assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval or equivalent review is not applicable.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core method development does not involve LLMs as an important, original, or non-standard component.

Guidelines:

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