Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Polynomial Time Quantum Approximation Schemes for Constrained Optimization

Chinonso Onah

conah@atom-computing.com

Volkswagen Group (Germany)

Kristel Michielsen

Forschungszentrum Jülich

Article

Keywords:

Posted Date: October 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11280943/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Polynomial Time Quantum Approximation Schemes for Constrained Optimization

Chinonso Onah[1,2] and Kristel Michielsen[3,4]

[1]Volkswagen AG, Berliner Ring 2, 38440 Wolfsburg, Germany

[2]Department of Physics, RWTH Aachen University, 52056 Aachen, Germany

[3]Forschungszentrum Jülich, Germany

[4]Universität zu Köln, 50923 Köln, Germany

When does a noisy quantum sampler enable an end-to-end polynomial-time algorithm with performance guarantees? Building on the instance-dependent finite-depth, finite-shot guarantees in the Constraint-Enhanced QAOA (CE–QAOA) studied in Ref. [1], we answer this question for a complete quantum–classical pipeline executed on near-term quantum devices. The analysis of Ref. [1] provides conditions under which the ideal CE–QAOA sampler can assign inverse-polynomial probability $p_0 = \Omega(n^{-k})$ to the set of optimal solutions. Here, we extend this analysis and show that, with independent-shot amplification, polynomial-time feasibility repair, and exact scoring, the resulting pipeline recovers an optimum with high probability in polynomial runtime. It therefore gives a quantum-powered fully polynomial randomized approximation scheme, which we call an $\mathbf{FPRAS}_q$.

We next show that this guarantee can survive device noise within an instance-dependent window. For effective circuit depth $G(n,p) = \Theta(pn)$, retaining a $G(n,p)^{-1}$ fraction of the ideal optimal mass is sufficient to recover an optimum with probability at least $1-\delta$ using

$$\mathcal{O}\Big(pn^{k+1}\log(1/\delta)\Big)$$

shots. Outside this exact-hit window, deterministic repair continues to guarantee feasibility and yields an instance-dependent $(1+\varepsilon)$-approximation scheme whenever the repair-induced objective inflation is controlled. We term this the **ε-Polynomial-Time Quantum Approximation Scheme ($\varepsilon-$PTQAS)**.

We also place our hybrid algorithm within the Chen–Cotler–Huang–Li oracle model [2] and locate the source of the quantum–classical separation. On any NP-hard kernel-admissible promise family, no polynomial-time classical sampler can reproduce the inverse-polynomial optimal overlap of NP-HQ unless NP $\subseteq$ BPP, even when granted the same classical repair and perfect access to the constraint structure. The separation therefore lies in generating the sampling distribution, not in classical processing

Finally, we introduce **Heavy-Hitter QAOA (HH-QAOA)**, a geometry-informed filtering refinement that preserves these conditional guarantees while reducing the retained candidate set from $\mathcal{O}(n^{k+1})$ to $\mathcal{O}(n^k)$ and the classical post-processing cost by one power of n. We demonstrate the complete pipeline on 127-qubit IBM Eagle-r3 processors using QOptlib [3]

Chinonso Onah: Corr. Author. conah@atom-computing.com

travelling-salesman instances with up to 100 logical variables, matching or improving the published reference tours in every tested case.

1 Introduction

Constrained combinatorial optimization covers a broad class of scientific and engineering problems in which the objective is evaluated over configurations satisfying hard constraints[4, 5]. Representative examples include the Travelling Salesman Problem (TSP/ATSP) [6, 7], the Quadratic Assignment Problem (QAP) [8, 9], capacitated vehicle routing [10], generalized assignment and multiple-knapsack problems [11, 12], k-dimensional matching for $k \geq 3$ [13, 14], shared-transportation problems [15, 16, 17], and job-shop and flow-shop scheduling problems [18, 19, 16]. Their admissible configurations are organized by assignment, matching, routing, capacity, ordering, and scheduling constraints[5]. The same structural motifs appear in polynomial-time solvable problems. Linear assignment is solved exactly by the Hungarian algorithm [20, 21], and maximum matching admits polynomial-time algorithms based on Edmonds' blossom construction [22]. Assignment matrices, matching relations, and one-hot or k-hot cardinality constraints therefore provide a common structural language across tractable and NP-hard optimization problems. One-hot encodings represent cardinality constraints by assigning one excitation to each logical block while relations among blocks enforce the remaining local and global constraints. The corresponding feasible configurations form a subset of the already cardinality-constrained computational basis, and the quantum sampler must place sufficient probability on this subset to produce candidates for downstream processing.

However, concentration of probability mass on the feasible subspace is only the first step in constrained quantum optimization. To obtain an end-to-end algorithm, one must also identify and score candidate solutions efficiently from the obtained samples. On real devices, quantum noise degrades feasibility further through noise-induced constraint violations, so one must additionally repair corrupted samples before scoring. This paper develops that full algorithmic layer around the recently introduced Onah–Firt–Michielsen (OFM) kernel of Ref. [23, 24] and shows how to exploit the same kernel structure classically, through deterministic checking, repair, and scoring, so that the quantum and classical stages operate as a single optimization pipeline. The CE-QAOA sampler already exploits these structures and symmetries in the quantum dynamics, leading to several guarantees studied in Refs. [1, 24, 25, 26] and reviewed in Appendix A.1. The associated polynomial-time hybrid quantum–classical (PHQC) algorithm was introduced in Ref. [23]. In that construction, quantum samples generated by CE–QAOA are passed to a classical checker for feasibility verification and scoring, while infeasible samples are discarded. Here we propose a noise-aware extension called the Noisy Polytime Hybrid Quantum-Classical (NP-HQ) algorithm. The post-processor first checks feasibility, applies a problem-structured repair map when noise corrupts the one-hot patterns, evaluates the objective, and returns the best repaired candidate.

Our main results arise by exploiting the structures and symmetries of the OFM kernel at both stages of the pipeline: through the symmetry-powered quantum dynamics of CE–QAOA and through problem-dependent repair maps applied classically to quantum-generated samples. We show that, in the low-noise regime where the optimal solution retains inverse-polynomial sampling weight, NP-HQ is an exact-hit fully polynomial randomized approximation scheme, FPRAS_q with polynomial shot complexity and polynomial

classical runtime. Beyond that regime, deterministic repair still supports an instance-dependent $(1+\varepsilon)$-approximation scheme whenever its objective inflation is controlled. A second contribution is a noise-to-runtime translation. We exploit a Lipschitz-type total-variation bound to convert device-level noise directly into a lower bound on post-noise optimal mass and hence into an explicit polynomial shot budget.

A further contribution is a conditional quantum–classical separation for the resulting sampling pipeline. We show that the complete NP-HQ pipeline fits the Chen–Cotler–Huang–Li NISQ oracle model[2]. We compare NP-HQ with classical polynomial-time strategies allowed to sample arbitrary efficiently samplable distributions, apply the same classical repair map, and score every repaired candidate. If such a classical strategy achieved inverse-polynomial optimal overlap uniformly on an NP-hard kernel-admissible promise family, standard success amplification would place the corresponding promise search problem in BPP. This conclusion remains unchanged when the classical sampler is granted perfect access to row-one-hot strings or uniformly random permutations. The computational distinction is therefore in the generation of the optimum-biased sampling distribution. Finally, we introduce Heavy-Hitter QAOA (HH-QAOA), a refined post-processing strategy that retains only those bit-strings whose empirical frequencies exceed a prescribed threshold. Focusing on a small thresholded candidate set reduces post-processing cost substantially, and in the strongest regimes where additional problem structure present in the quantum-generated samples is utilized, the downstream checking, repair, and scoring cost depends on the size of the retained heavy-hitter set, leading to orders-of-magnitude reductions in the classical postprocessing time. We demonstrate the complete NP-HQ pipeline on 127-qubit IBM Eagle-r3 processors using QOptlib[3] travelling-salesman instances with up to 100 logical variables, matching or improving reference tours in every tested case.

1.1 Related Work

Post-selection and filtering heuristics appear widely in variational and sampling-based quantum algorithms. Common strategies include discarding low-quality samples, reweighting outcomes, and applying symmetry checks to improve empirical performance [27, 28]. A substantial body of NISQ work also addresses device noise through zero-noise extrapolation, symmetry verification, readout mitigation, and regression-based error mitigation [29, 30, 27, 31, 32]. These methods primarily target expectation-value estimation, whereas the present work concerns the finite-shot recovery of optimal bitstrings, for which existing error-mitigation tools are of limited utility.

The *quantum alternating-operator* paradigm uses mixing operators that preserve hard constraints, thereby confining the evolution to structured subspaces and yielding several reported performance gains [33, 34, 35, 36, 23, 25, 1, 24]. In implementation, however, constraint-preserving mixer Hamiltonians may contain noncommuting terms, so their Trotterized realizations need not reproduce the ideal subspace-preserving dynamics exactly [37, 24]. Device noise can further induce leakage from the encoded subspace, motivating explicit noise-handling mechanisms for constrained quantum optimization. Recent hardware studies with XY-type mixers nevertheless show that constrained dynamics can produce significant gains on near-term devices when these implementation and noise obstacles are addressed [38, 39, 40]. A related line of work exploits quantum noise using physical insights from gauge transformations to improve the performance of hybrid quantum algorithms [39, 40, 41].

Classical post-processing has also been incorporated directly into quantum optimization pipelines. Greedy post-processing has been shown to increase the approximation quality of QAOA samples [42], while post-processing variational algorithms can repair infeasible

outputs produced by annealers or gate-based devices [43]. These works establish the algorithmic value of classical processing after quantum sampling. The present construction goes further by incorporating deterministic checking, feasibility restoration, and exact scoring into a finite-shot complexity analysis that tracks the optimal sampling mass through device noise and converts it into an explicit end-to-end success probability and runtime.

Heavy-hitter and frequent-item extraction originated in the classical data-stream literature, where thresholded frequency summaries identify items carrying a prescribed fraction of the observed mass [44, 45]. We apply this principle to finite-shot quantum output histograms. The classical heavy-hitter bound supplies a distribution-independent candidate-set reduction, while the CE–QAOA optimal-mass bound and its device-noise window permit a sharper quantum-informed threshold. The resulting refinement is therefore tied to the structure and noise stability of the quantum sampling distribution rather than generic frequency estimation.

Classical approximation schemes for structured metric optimization exploit substantial geometric or combinatorial structure [46]. In the same spirit, we repurpose polynomial-time assignment and matching algorithms [20, 22] as deterministic feasibility-restoration primitives for measured quantum samples. From a complexity-theoretic perspective, our construction fits within the hybrid model of noisy quantum computation studied by Chen, Cotler, Huang, and Li [2]. To the best of our knowledge, no previous work combines finite-shot optimal recovery, deterministic feasibility restoration, an explicit finite-noise lower bound on the recovered optimal mass, polynomial shot complexity, and polynomial post-processing runtime into a single end-to-end optimization guarantee under device noise. The hardware experiments in Section 4 instantiate the complete sampling–checking–repair–scoring pipeline on contemporary superconducting processors.

2 Polynomial-Time Quantum Approximation Schemes

The approximation schemes developed below inherit their quantum sampling guarantees from the CE–QAOA analyses of Refs. [1, 24, 25]. To make the assumptions entering those guarantees explicit, we first state the Onah–Firt–Michielsen (OFM) kernel in full. The kernel provides the common interface between the quantum sampler and the classical checker–repair–scoring stage: it fixes the encoded manifold, the residual constraint Hamiltonians, their symmetries, the compatible mixer, and the initial state. The subsequent complexity results apply to problem families admitting this interface together with the stated instance-dependent lower bound on optimal sampling mass.

2.1 The OFM kernel and its assignment specialization

Definition 1 (The Onah–Firt–Michielsen (OFM) kernel[23, 24])**.** *An optimization instance belongs to the Onah–Firt–Michielsen kernel if there exist $m, n_{\mathrm{loc}} \in \mathbb{N}$ and a one-hot encoder such that the quantum dynamics is initialized in the fixed-Hamming-weight space*

$$\mathcal{H}_{\mathrm{OH}} := (\mathcal{H}_1)^{\otimes m}, \qquad \mathcal{H}_1 := \mathrm{span}\{|e_1\rangle, \ldots, |e_{n_{\mathrm{loc}}}\rangle\},$$

with one excitation in each block. Its computational-basis states are identified with words

$$x = (x_1, \ldots, x_m) \in X := [n_{\mathrm{loc}}]^m, \qquad |x\rangle = |e_{x_1}\rangle \otimes \cdots \otimes |e_{x_m}\rangle .$$

Thus,

$$|X| = \dim(\mathcal{H}_{\mathrm{OH}}) = n_{\mathrm{loc}}^m.$$

In this encoded basis, the diagonal problem Hamiltonian decomposes as

$$H_C = H_{\mathrm{pen}} + H_{\mathrm{obj}} + H'_{\mathrm{pen}},$$

where H_{obj} encodes the objective, H_{pen} contains the symmetry-preserving residual constraint penalties, and H'_{pen} contains additional diagonal penalties that need not preserve the full kernel symmetry. The following requirements hold.

(a) Penalty structure. *The Hamiltonian H_{pen} is a sum of squared affine one-hot, degree, or capacity residuals, optionally together with linear forbids, with integer coefficients bounded polynomially in the encoded instance size. Consequently,*

$$\mathrm{spec}(H_{\mathrm{pen}}) \subseteq \{0, 1, \ldots, t_{\max}\}, \qquad t_{\max} = \mathrm{poly}(m, n_{\mathrm{loc}}).$$

(b) Pattern symmetry. *The symmetry-preserving penalty H_{pen} is invariant under block permutations S_m and global symbol relabelings $S_{n_{\mathrm{loc}}}$. It therefore organizes the encoded space into structured penalty levels*

$$L_t := \{x \in X : \langle x| \, H_{\mathrm{pen}} \, |x\rangle = t\}.$$

The term H'_{pen} collects problem-dependent penalty contributions that break some or all of these symmetries.

(c) Mixer and initial state. *The canonical one-hot mixer is the normalized block-local XY Hamiltonian*

$$\widetilde{H}_{XY}^{(b)} = \frac{1}{n_{\mathrm{loc}} - 1} \sum_{1 \le u < v \le n_{\mathrm{loc}}} \left(X_u^{(b)} X_v^{(b)} + Y_u^{(b)} Y_v^{(b)} \right), \tag{1}$$

which preserves $\mathcal{H}_1$ and satisfies

$$\left\| \widetilde{H}_{XY}^{(b)} \right\| = O(1)$$

on the one-excitation sector. The corresponding initial state is

$$|s_0\rangle = |s_{\mathrm{blk}}\rangle^{\otimes m}, \qquad |s_{\mathrm{blk}}\rangle = \frac{1}{\sqrt{n_{\mathrm{loc}}}} \sum_{j=1}^{n_{\mathrm{loc}}} |e_j\rangle.$$

Definition 1 provides a common interface from which problem-structure-aware operations can be systematically derived for many well-studied problems, including the TSP, QAP, CVRP, shared transportation problems, generalized assignment and multiple-knapsack problems, and k-dimensional matching for $k \geq 3$ [15, 17, 26, 5, 7, 13, 23].

The kernel separates two forms of constraint handling. Local cardinality constraints are enforced by the encoder and are preserved by the mixer, so the ideal quantum dynamics remains inside X. Residual global constraints are retained in the diagonal penalty Hamiltonians and define a generally smaller globally feasible set

$$\mathcal{F} \subseteq X.$$

The same residual constraint structure is subsequently available to the classical checker and repair map. For the assignment and travelling-salesman specialization,

$$m = n_{\mathrm{loc}} = n, \qquad X = [n]^n, \qquad |X| = n^n,$$

and the encoding uses n^2 physical qubits. Each word

$$x = (x_1, \ldots, x_n) \in [n]^n$$

selects one symbol in every one-hot block. Global feasibility further requires every symbol to occur exactly once, so that

$$\mathcal{F} = \{x \in [n]^n : x_a \neq x_b \text{ for all } a \neq b\} \cong S_n.$$

Thus the mixer explores the full encoded assignment space X, while the residual penalties distinguish the permutation sector $\mathcal{F} \subset X$.

2.2 CE–QAOA sampling distribution

Define the total mixer Hamiltonian and its unitary by

$$H_M := \sum_{b=1}^{m} \widetilde{H}_{XY}^{(b)}, \qquad U_M(\beta) := e^{-i\beta H_M} = \bigotimes_{b=1}^{m} \exp\left(-i\beta \widetilde{H}_{XY}^{(b)}\right).$$

Because the block terms act on disjoint registers, $U_M(\beta)$ preserves $\mathcal{H}_{\mathrm{OH}}$. Since H_C is diagonal in the computational basis, the problem unitary

$$U_C(\gamma) := e^{-i\gamma H_C}$$

also preserves $\mathcal{H}_{\mathrm{OH}}$. A depth-$p$ CE–QAOA circuit therefore prepares

$$|\psi_p(\boldsymbol{\beta}, \boldsymbol{\gamma})\rangle = \left(\prod_{r=1}^{p} U_M(\beta_r) U_C(\gamma_r)\right) |s_0\rangle \in \mathcal{H}_{\mathrm{OH}}. \tag{2}$$

Its ideal output distribution is

$$q_0(x) := |\langle x|\psi_p(\boldsymbol{\beta}, \boldsymbol{\gamma})\rangle|^2, \qquad x \in X.$$

Let

$$C(x) := \langle x| H_{\mathrm{obj}} |x\rangle$$

denote the true objective evaluated on a feasible configuration, and define

$$C^\star := \min_{x \in \mathcal{F}} C(x), \qquad \Omega^\star := \{x \in \mathcal{F} : C(x) = C^\star\}.$$

The ideal optimal sampling mass is then

$$q_0 := q_0(\Omega^\star) = \sum_{x \in \Omega^\star} q_0(x).$$

The exact-hit and approximation guarantees developed below are conditioned on the instance-dependent regime

$$q_0 \geq c\, n^{-k}$$

for fixed constants $c > 0$ and $k \geq 0$. The remaining analysis is to determine when this optimal mass survives device noise, how many circuit repetitions are then sufficient for recovery, and how deterministic checking, repair, and scoring complete the polynomial-time hybrid pipeline.

2.3 Background to Performance Guarantees

The problem-dependent assumption rests on the lattice normalization of the problem Hamiltonian which proceeds as follows: Fix a base angle γ and define the *wrapped phase* of each basis string:

$$\theta(z) \;:=\; \gamma\, E(z) \pmod{2\pi}. \tag{3}$$

Let $H_C |x\rangle = E(x) |x\rangle, \qquad x \in X$ so that $\theta^\star := \gamma\, E^\star \pmod{2\pi}$ is the common wrapped phase of the optimal energy level. Define the *wrapped phase separation*:

$$\delta \;:=\; \min_{y \notin \Omega^\star} \mathrm{dist}_{\mathbb{T}}(\theta(y),\, \theta^\star), \tag{4}$$

where $\mathrm{dist}_{\mathbb{T}}(\phi, \varphi) := \min_{k \in \mathbb{Z}} |\phi - \varphi + 2\pi k| \in [0, \pi]$. The condition $\delta > 0$ means that the optimal phase is separated from all suboptimal phases on the circle.

Remark 2 (When $\delta > 0$ holds). The wrapped phase separation δ depends on the cost spectrum $\{E(z)\}$ and the angle γ. For generic γ (i.e., outside a measure-zero set), distinct energy levels map to distinct phases, so $\delta > 0$. Moreover, if the energy spectrum has integer or rational-multiple structure (as is common in combinatorial problems with integer-valued costs), then a suitable γ can be chosen to maximize δ. As discussed in Ref. [1], one may

also relax the exact phase-gap assumption by averaging over a grid of γ values when several near-optimal levels cluster in phase; the resulting bound degrades gracefully.

The finite-depth, finite-shot guarantees on which the remainder of the analysis rests arise from three analytic ingredients, whose interplay is reviewed in detail in Appendix A. Here we recall the key ingredients.

Ingredient 1. Mixer envelope. The block-local XY mixer, viewed through a cost-basis dephasing channel, induces a doubly stochastic Markov chain on the diagonal of the density operator. Starting from the uniform one-hot product state $|s_0\rangle$, the iterated transition gives a nonnegative, normalized mixer envelope

$$W_p(z;\boldsymbol{\beta}) \;=\; [M_{\beta_p}\cdots M_{\beta_1}\, v^{(0)}](z), \qquad v^{(0)}(z) = |\,\langle z|s_0\rangle\,|^2, \tag{5}$$

where each $M_\beta(z|y) = |\,\langle z|\,U_M(\beta)\,|y\rangle\,|^2$ is unistochastic (see Eq. (27) in Appendix A). This envelope quantifies how the mixer spreads probability across the encoded manifold. The key quantity for the success bound is the optimal-set envelope weight

$$C_\beta \;=\; \sum_{x\in\Omega^\star} W_p(x;\boldsymbol{\beta}), \tag{6}$$

the total mixer-envelope mass on the set $\Omega^\star$ of optimal feasible strings.

Ingredient 2. Fejér phase filter. The diagonal cost operator H_C assigns a wrapped phase $\theta(z) = \gamma E(z) \pmod{2\pi}$ to each basis string, where γ is the cost-layer angle. The Fejér kernel of order p is the nonnegative trigonometric polynomial

$$F_p(\vartheta) \;=\; \frac{1}{p+1}\left|\sum_{r=0}^{p} e^{ir\vartheta}\right|^2 \;=\; \frac{1}{p+1}\left(\frac{\sin(\frac{(p+1)\vartheta}{2})}{\sin(\frac{\vartheta}{2})}\right)^2, \tag{7}$$

which peaks at $\vartheta = 0$ with value $F_p(0) = p+1$ and decays away from the peak. Strings whose cost phase lies close to the optimal phase $\theta^\star$ receive amplification of order $p+1$, while strings separated by a phase gap $\delta > 0$ are suppressed by $M_p(\delta) \le [(p+1)\sin^2(\delta/2)]^{-1}$.

Ingredient 3. Factorized reference distribution. In a dephased reference model, a normalized factorized reference distribution emerges as

$$\Pr_p^{\text{ref}}[z] \;=\; \frac{W_p(z;\boldsymbol{\beta})\; F_p(\theta(z)-\theta^\star)}{\sum_{y\in X} W_p(y;\boldsymbol{\beta})\; F_p(\theta(y)-\theta^\star)}. \tag{8}$$

This factorized distribution leads to a dimension-free lower bound on the optimal-set probability by combining Eq. 6 and Eq. 7(Theorem 36)

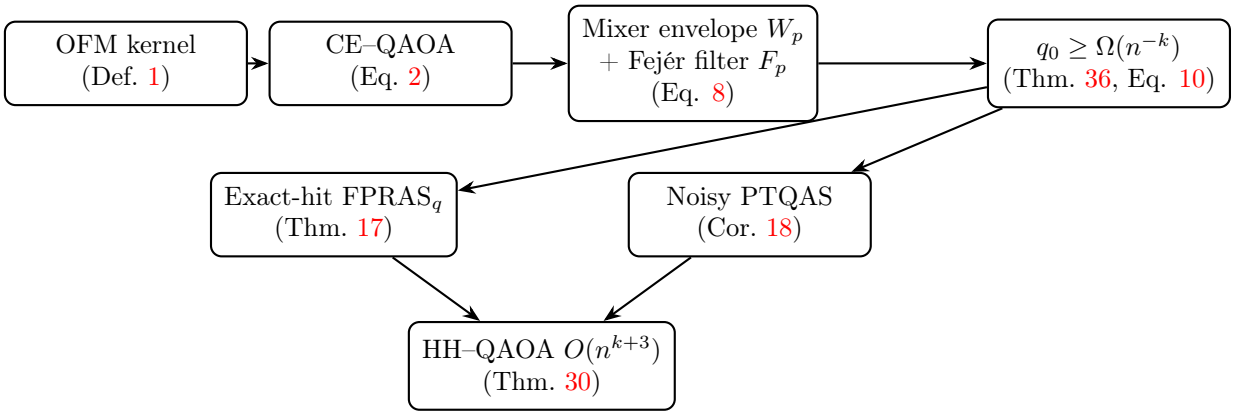
$$q_0 \;\ge\; \frac{(p+1)\,C_\beta}{(p+1)\,C_\beta \;+\; M_p(\delta)\,(1-C_\beta)}, \tag{9}$$

which depends only on the filter order p, the phase gap δ, and the optimal-set envelope weight C_β. The detailed derivation is given in Appendix A. When evaluated on a problem family for which the instance-dependent envelope and phase conditions detailed in Appendix A.5 hold, this joint success probability law yields the inverse-polynomial optimal-mass premise used throughout the downstream analysis:

$$q_0 \;\ge\; \frac{(p+1)c_0 n^{-a}}{(p+1)c_0 n^{-a} + \big[(p+1)c_1\big]^{-1}} \;=\; \Omega(n^{-a}). \tag{10}$$

The load-bearing geometric quantity in the finite-depth Fejér mechanism of Ref. [1] is the optimal-set mixer-envelope weight C_β defined in Eq. (6). The recent geometry–interference analysis of Ref. [24] suggests that the corresponding optimal-set transport

weight need not be small under coherent implementation. It also establishes a complementary logarithmic-depth route to inverse-polynomial success guarantees under the same normalized OFM transport structure and lattice normalization. No general pointwise ordering between coherent and dephased output probabilities is assumed here; the dephased reference law is compared with the coherent circuit only under the phase-alignment conditions of Ref. [24]. The logical flow from kernel structure to end-to-end hybrid guarantees is visualized in the diagram below.



2.4 The hybrid quantum–classical (PHQC) framework

To operationalize the quantum ansatze, Algorithm 1, the polynomial-time hybrid quantum–classical (PHQC) algorithm was introduced in Ref. [23] where the quantum samples generated by CE-QAOA are handed over to the classical checker for feasibility. We further extend the pipeline to include a coarse parameter-grid search. A single appearance is sufficient for exact recovery. Let p be the circuit depth and let

$$\mathcal{G} = \{(\beta_a, \gamma_b) : 0 \leq a \leq N_\beta,\ 0 \leq b \leq N_\gamma\} \subset [0, \pi] \times [0, \pi]$$

be a rectangular coarse grid, for example

$$\beta_a = a\,\Delta_\beta, \qquad \gamma_b = b\,\Delta_\gamma, \qquad \Delta_\beta = \frac{\pi}{N_\beta}, \quad \Delta_\gamma = \frac{\pi}{N_\gamma}.$$

For each $(\beta, \gamma) \in \mathcal{G}$, PHQC prepares the CE–QAOA state

$$|\psi_p(\beta, \gamma)\rangle = U_M(\beta)\, e^{-i\gamma H_C}\, |s_0\rangle$$

(for $p = 1$; or depth-p generalization in Eq. 2), samples it S times, and feeds each measured bitstring into the deterministic feasibility oracle.

Algorithm 1 PHQC — CE–QAOA grid search with feasibility filtering and scoring[23]

Require: depth p; coarse grid $\mathcal{G} \subset [0,\pi]^2$; cost Hamiltonian H_C; objective Hamiltonian H_{obj}; mixer U_M; initial state $|s_0\rangle$; shots S per grid point.
Ensure: best feasible sampled bitstring $b^\star$ and its score $E^\star$.

$b^\star \leftarrow$ `null`, $E^\star \leftarrow +\infty$
for each $(\beta,\gamma) \in \mathcal{G}$ **do**
  for $r = 1,\dots,S$ **do**
    prepare $|s_0\rangle$
    apply the CE–QAOA circuit for (β,γ)
    measure a bitstring b
    if FeasibleCheck(b) **then**
      compute $E(b) = \langle b \mid H_{\text{obj}} \mid b\rangle$
      if $E(b) < E^\star$ **then**
        $b^\star \leftarrow b$, $E^\star \leftarrow E(b)$
      end if
    end if
  end for
end for
return $(b^\star, E^\star)$

Lemma 3 (No-hit bound at a good grid point[23])**.** *Assume there exists a parameter pair $(\beta^\sharp,\gamma^\sharp) \in \mathcal{G}$ such that*

$$p_\star^\sharp := p_\star(\beta^\sharp,\gamma^\sharp) > 0.$$

If PHQC draws S independent samples at each grid point, then the probability of observing no optimal feasible bitstring at the good grid point satisfies

$$\Pr\Big[\textit{no hit at } (\beta^\sharp,\gamma^\sharp)\Big] = (1-p_\star^\sharp)^S \le e^{-p_\star^\sharp S}.$$

Hence

$$S \ge \Big\lceil \frac{\ln(1/\delta)}{p_\star^\sharp} \Big\rceil \quad\Longrightarrow\quad \Pr\Big[\textit{at least one optimal feasible hit at } (\beta^\sharp,\gamma^\sharp)\Big] \ge 1-\delta.$$

Theorem 4 (Exact-recovery guarantee for PHQC[23])**.** *Assume the coarse grid $\mathcal{G}$ contains a parameter pair $(\beta^\sharp,\gamma^\sharp)$ with optimal feasible probability mass $p_\star^\sharp > 0$. Run PHQC with S shots at each grid point and let the deterministic checker return the minimum-cost feasible bitstring over all samples. Then*

$$S \ge \Big\lceil \frac{\ln(1/\delta)}{p_\star^\sharp} \Big\rceil \quad\Longrightarrow\quad \Pr[\textit{PHQC returns a globally optimal feasible bitstring}] \ge 1-\delta.$$

In particular, if $p_\star^\sharp \ge n^{-k}$ for some constant $k > 0$, then

$$S = O\big(n^k \log(1/\delta)\big),$$

and if moreover $|\mathcal{G}| = \text{poly}(n)$, the full quantum-plus-classical pipeline has polynomial shot complexity and polynomial total runtime.

Algorithm 2 NP-HQ Algorithm — depth-p CE–QAOA + deterministic checker/repair

Require: problem size n; depth p; angles $\{(\beta_\ell, \gamma_\ell)\}_{\ell=1}^{p}$; penalty Hamiltonian H_{pen}; mixer Hamiltonian H_M; objective Hamiltonian H_C; shot budget S.

Ensure: best feasible bit-string $b^\star$ and objective value $c^\star$.

Prepare $|\psi\rangle \leftarrow |s_0\rangle = |s_{\text{blk}}\rangle^{\otimes m}$.
for $\ell = 1$ **to** p **do**
    Apply $U_{\text{C}}(\gamma_\ell) = e^{-i\gamma_\ell H_{\text{C}}}$ to $|\psi\rangle$.
    Apply $U_M(\beta_\ell) = e^{-i\beta_\ell H_M}$ to $|\psi\rangle$.
end for
Measure $|\psi\rangle$ for S shots; store multiset $\mathcal{S}$.
$(b^\star, c^\star) \leftarrow (\text{null}, +\infty)$.
for all $b \in \mathcal{S}$ **do**
    if FEASIBLE(b) **then**
        $b_{\text{ok}} \leftarrow b$
    else
        $b_{\text{ok}} \leftarrow$ REPAIR(b)
    end if
    $c \leftarrow \langle b_{\text{ok}}|H_C|b_{\text{ok}}\rangle$.
    if $c < c^\star$ **then**
        $(b^\star, c^\star) \leftarrow (b_{\text{ok}}, c)$
    end if
end for
return $(b^\star, c^\star)$.

2.5 Noisy Polytime Hybrid Quantum–Classical Algorithm (NP-HQ)

Here we study the **NP-HQ** Algorithm, a noise adapated version of Alg 1 endowed with classical repairs based on problem structure. The classical repairs are based on strongly polynomial classical matching algorithms like Hungarian[20] and Blossom[22] algorithms. After p CE–QAOA layers acting on $|s_0\rangle$, and measuring the circuit $S = \text{poly}(n)$ times on a noisy quantum device; a deterministic post-processing performs: (i) feasibility check against one-hot constraints ($O(n^2)$); (ii) if needed, a polynomial-time *repair* in $O(n^3)$; (iii) exact scoring $b \mapsto \langle b|H_C|b\rangle$ in $O(n^2)$. Total classical work $O(S\,n^3)$. The key insight is that every corrupt block can be projected back to the nearest feasible block in polynomial time. Likewise, the full sample can be projected to a feasible bitstring. We outline the full NP-HQ algorithm in Alg. 2 and return to the processing of noisy samples next.

2.5.1 Post Processing of Noisy Quantum Samples

For a constrained optimization problem, the CE-QAOA circuit run on a quantum device returns a multiset of measurement outcomes, represented as a map $C : \{0,1\}^L \to \mathbb{N}$ where $L = n^2$ for an $n \times n$ assignment. A classical routine must (i) discard infeasible bit-strings, (ii) decode every feasible string, (iii) evaluate the true cost, and (iv) pick the best candidate(s) for reporting. This is the PHQC algorithm outlined in Alg. 1.

To expose the runtime features, let $K = |\text{keys}(C)| \le \min(S, 2^L)$ be the number of unique bit-strings obtained from S shots. Feasibility checking inspects $L = \Theta(n^2)$ bits per key. Decoding and summing its edge costs each take $O(n^2)$ operations. Hence the loop runs in $O(K\,n^2)$. Sorting at most K feasible samples adds $O(K \log K)$, so the total time is

$$T(K, n) \;=\; O(K\,n^2 + K \log K).$$

In typical fixed-shot experiments, the full routine is therefore *effectively quadratic* in the problem size n.

In presence of quantum noise, the original blockwise one–hot and k-hot constraints themselves begin to fail and the obtained bitstrings may violate the global constraints. Thus, we introduce a *deterministic repair* based on the classical matching algorithms [20, 22]. This converts noisy samples into *feasible* ones while preserving polynomial–time guarantees since these classical interventions remain polynomial.

2.5.2 Complexity Bounds in the NP-HQ Algorithm

Let $G(n,p)$ denote the effective depth of the p-layer CE-QAOA circuit. Under parallel scheduling, the state-preparation circuit, the block-local mixer, and the problem unitary each admit $O(n)$-depth implementations[23]. Hence

$$G(n,p) = O((p+1)n) = O(pn), \qquad p \geq 1. \tag{11}$$

The effective depth should be distinguished from the total two-qubit gate count. Example, for the dense TSP construction, the problem Hamiltonian contains $\Theta(n^3)$ two-qubit interaction terms per CE-QAOA layer as noted in Remark 32. Here, we model the accumulated noise at each effective circuit layer by a channel $\mathcal{N}_\epsilon$ satisfying

$$\|\mathcal{N}_\epsilon - \mathcal{I}\|_\diamond \leq \epsilon. \tag{12}$$

Here ϵ is the effective error accumulated within one noisy depth layer, including the combined contribution of gates executed in parallel. For example, a primitive two-qubit depolarising channel

$$\Lambda_\eta(\rho) = (1-\eta)\rho + \eta\frac{\mathbb{I}_4}{4} \tag{13}$$

satisfies [47]

$$\frac{1}{2}\|\Lambda_\eta - \mathcal{I}\|_\diamond = \frac{15}{16}\eta \leq \eta. \tag{14}$$

The effective layer parameter ϵ absorbs the accumulated contribution of all such local channels acting within the same circuit layer.

Let P_0 and P_ϵ denote the ideal and noisy output distributions. Repeated application of the trace-distance triangle inequality [47, Lemma 9.1.2], together with contractivity under quantum channels and measurement, gives

$$\|P_\epsilon - P_0\|_{\mathrm{TV}} \leq \frac{1}{2}G(n,p)\epsilon. \tag{15}$$

The complete telescoping argument and the resulting Lipschitz continuity of the measured output distribution are given in Appendix A.6.

Let $\mathcal{X}^\star$ denote the set of globally optimal feasible strings and define

$$p_0 := P_0(\mathcal{X}^\star), \qquad p_{\min}(\epsilon) := P_\epsilon(\mathcal{X}^\star). \tag{16}$$

The total-variation bound implies

$$p_{\min}(\epsilon) \geq p_0 - \frac{1}{2}G(n,p)\epsilon. \tag{17}$$

Suppose that the ideal optimal mass satisfies

$$p_0 \geq cn^{-k} \tag{18}$$

for fixed constants $c > 0$ and $k \geq 0$, then the following theorem holds.

Theorem 5 (Noise-robust exact-hit complexity)**.** *Let $\mathcal{X}^\star$ denote the set of globally optimal feasible strings, and define*

$$p_0 := P_0(\mathcal{X}^\star), \qquad p_{\min}(\epsilon) := P_\epsilon(\mathcal{X}^\star).$$

Suppose that the ideal optimal mass satisfies

$$p_0 \geq cn^{-k}$$

for fixed constants $c > 0$ and $k \geq 0$. Let $G(n,p) > 1$ denote the effective depth of the p-layer circuit, and assume that the effective error per noisy depth layer satisfies

$$\epsilon \leq \frac{2p_0}{G(n,p)}\left(1 - \frac{1}{G(n,p)}\right). \tag{19}$$

Then the noisy optimal mass satisfies

$$p_{\min}(\epsilon) \geq \frac{p_0}{G(n,p)} \geq \frac{c}{G(n,p)}n^{-k}. \tag{20}$$

Consequently, a globally optimal feasible string is observed with probability at least $1-\delta$ after

$$S = \left\lceil \frac{G(n,p)}{c}\, n^k \log\left(\frac{1}{\delta}\right) \right\rceil \tag{21}$$

independent circuit executions. If $G(n,p) = \Theta(pn)$, then

$$S = \mathcal{O}\left(pn^{k+1}\log(1/\delta)\right). \tag{22}$$

Thus accumulated circuit noise increases the shot budget by one factor in n.

Proof. The total-variation continuity bound gives Eq. 17. To retain at least a $G(n,p)^{-1}$ fraction of this optimal probability mass, it is sufficient that the effective per-layer error in Eq. 17 obeys

$$\epsilon \leq \frac{2p_0}{G(n,p)}\left(1 - \frac{1}{G(n,p)}\right). \tag{23}$$

That is Eq. 19. Substituting (19) back into (17) yields

$$p_{\min}(\epsilon) \geq \frac{p_0}{G(n,p)} \geq \frac{c}{G(n,p)}n^{-k}.$$

This proves Eq. 20. By Lemma 3, the probability that none of S independent measurements belongs to $\mathcal{X}^\star$ is $\leq \exp[-Sp_{\min}(\epsilon)]$. Solving for S one obtains (21). Finally, substituting $G(n,p) = \Theta(pn)$ gives the stated noisy shot-complexity bounds. □

3 Quantum randomized exact algorithm (FPRASq)

Definition 6 (Polynomial-Time Quantum Approximation Scheme (PTQAS))**.** *Fix a family of constrained minimization problems with optimum value $C^\star$. A (hybrid) algorithm family $\{\mathcal{A}_{n,\varepsilon,\delta}\}$ is a* Polynomial-Time Quantum Approximation Scheme (PTQAS) *if for every accuracy $\varepsilon \in (0,1]$ and confidence $\delta \in (0,1)$, on input size n the algorithm outputs a feasible solution X such that*

$$\Pr[C(X) \leq (1+\varepsilon)\, C^\star] \;\geq\; 1-\delta,$$

and the total runtime (quantum circuit calls + classical post-processing) is

$$\mathrm{poly}(n,\, 1/\varepsilon,\, \log(1/\delta)).$$

We say the PTQAS is strong *if its runtime is* $\mathrm{poly}(n,\, \log(1/\delta))$ *(degree 0 in $1/\varepsilon$).*

Definition 7 (FPRAS_q (hybrid))**.** *A hybrid quantum–classical algorithm $\mathcal{A}_{n,\varepsilon,\delta}$ is an $FPRAS_q$ if for all $\varepsilon \in (0,1]$ and $\delta \in (0,1)$ it outputs a feasible solution X such that*

$$\Pr[C(X) \leq (1+\varepsilon)C^\star] \geq 1-\delta,$$

and the total runtime (including circuit calls and classical post-processing) is $\mathrm{poly}(n, 1/\varepsilon, \log(1/\delta))$. *If in addition $\Pr[C(X) = C^\star] \geq 1-\delta$, we call it an* exact-hit *$FPRAS_q$.*

Definition 8 (Promised bounded-error hybrid approximation scheme)**.** *Fix $\rho \geq 1$. A hybrid algorithm is a* promised ρ-approximation scheme *if it always outputs a feasible solution X satisfying*

$$C(X) \leq \rho\, C^\star$$

for all instances in a stated promise class (e.g. bounded aspect ratio), with runtime $\mathrm{poly}(n)$.

In the noiseless setting, and under limited gate noise satisfying $\epsilon \leq p_0/G(n,p)$, NP-HQ is an *exact-hit FPRAS$_q$* in the sense of Definition 7: with $S = \mathrm{poly}(n, \log(1/\delta))$ circuit calls it returns a global optimum with probability at least $1-\delta$. Beyond this Lipschitz/TV regime, exact hits need not occur with inverse-polynomial probability; nevertheless NP-HQ remains feasibility-first via deterministic repair and admits a *promised bounded-error* guarantee (Definition 8) whenever the instance class ensures an additive repair loss $\Delta(n) = O(nw_{\max})$, yielding a constant-factor floor $C_{\mathrm{out}} \leq (1 + c_1\kappa)C^\star$ on misses. By Theorem 17 and Corollary 18, NP-HQ constitutes a *strong PTQAS* in the sense of Definition 6 whenever the (post-repair) optimal mass is inverse–polynomial in n.

3.1 Feasibility repair via nearest–permutation projection

Let $\Pi_n \subset \{0,1\}^{n\times n}$ be the set of permutation matrices. Given a measured bit–matrix $Y \in \{0,1\}^{n\times n}$, we *repair* it by projecting to the nearest permutation in Hamming distance:

$$R(Y) \;:=\; \arg\min_{P\in\Pi_n} d_H(Y,P), \qquad d_H(Y,P) = \sum_{i,j} |Y_{ij} - P_{ij}|. \tag{24}$$

This coincides with a maximum–weight assignment with binary weights.

Lemma 9 (Equivalence to maximum–weight matching; complexity)**.** *For fixed $Y \in \{0,1\}^{n\times n}$,*

$$\arg\min_{P\in\Pi_n} d_H(Y,P) \;=\; \arg\max_{P\in\Pi_n} \sum_{i=1}^{n} Y_{i,\,P(i)}.$$

Consequently, $R(Y)$ can be computed in $O(n^3)$ time by the Hungarian algorithm.

Proof. Expand the Hamming distance for any $P \in \Pi_n$:

$$d_H(Y,P) = \sum_{i,j}(Y_{ij} + P_{ij} - 2Y_{ij}P_{ij}) = \underbrace{\sum_{i,j} Y_{ij}}_{\text{const in } P} + \underbrace{\sum_{i,j} P_{ij}}_{=n} - 2\sum_{i=1}^{n} Y_{i,\,P(i)}.$$

Thus $\arg\min_P d_H(Y,P) = \arg\max_P \sum_i Y_{i,P(i)}$. The latter is the maximum–weight perfect matching on the complete bipartite graph with weights $w_{ij} = Y_{ij} \in \{0,1\}$, solvable in $O(n^3)$ time by the Hungarian method. □

Lemma 10 (Idempotency)**.** *For all Y, $R(Y) \in \Pi_n$. Moreover, R is idempotent and fixes feasible inputs: if $Y \in \Pi_n$ then $R(Y) = Y$, and $R(R(Y)) = R(Y)$ for all Y.*

Proof. By definition $R(Y) \in \Pi_n$. If $Y \in \Pi_n$, then $d_H(Y,Y) = 0 \leq d_H(Y,P)$ for all $P \in \Pi_n$, so $R(Y) = Y$. Finally, since $R(Y) \in \Pi_n$, applying R again leaves it unchanged. □

Lemma 11 (Repair monotonicity on feasible subsets)**.** *Let Ω be the feasible set and $R : \{0,1\}^{n^2} \to \Omega$ be deterministic with $R(x) = x$ for all $x \in \Omega$. Then for every $A \subseteq \Omega$ and any random outcome Y,*

$$\Pr[R(Y) \in A] \;\geq\; \Pr[Y \in A].$$

In particular, for the optimal set $\mathcal{X}^\star \subseteq \Omega$,

$$p_{\min}^{\mathrm{rep}} := \Pr[R(Y) \in \mathcal{X}^\star] \;\geq\; p_{\min}^{\mathrm{raw}} := \Pr[Y \in \mathcal{X}^\star].$$

Proof. If $Y \in A \subseteq \Omega$, then $R(Y) = Y \in A$. Hence $\{Y \in A\} \subseteq \{R(Y) \in A\}$ and the probability inequality follows. □

Lemma 12 (Total–variation contraction under deterministic repair)**.** *Let $R : \{0,1\}^{n\times n} \to \Pi_n$ be deterministic and let P, Q be probability distributions on $\{0,1\}^{n\times n}$. Denote by $R_{\#}P$ and $R_{\#}Q$ the push–forwards of P and Q under R. Then*

$$\|R_{\#}P - R_{\#}Q\|_{\mathrm{TV}} \;\le\; \|P - Q\|_{\mathrm{TV}}.$$

Proof. For any $A \subseteq \Pi_n$ we have

$$(R_{\#}P)(A) - (R_{\#}Q)(A) = P(R^{-1}(A)) - Q(R^{-1}(A)).$$

Taking absolute values and then the supremum over $A \subseteq \Pi_n$ yields $\|R_{\#}P - R_{\#}Q\|_{\mathrm{TV}} \le \|P - Q\|_{\mathrm{TV}}$. □

Lemma 13 (Repair cannot decrease mass on feasible subsets)**.** *Assume $R(x) = x$ for all $x \in \Pi_n$. Then for any $A \subseteq \Pi_n$ and any distribution P on $\{0,1\}^{n\times n}$,*

$$(R_{\#}P)(A) \;\ge\; P(A).$$

In particular, $(R_{\#}P)(\mathcal{X}^\star) \ge P(\mathcal{X}^\star)$.

Proof. If $Y \in A \subseteq \Pi_n$ then $R(Y) = Y \in A$, hence $\{Y \in A\} \subseteq \{R(Y) \in A\}$. □

Remark 14. The bound on $p_{\min}(\epsilon)$ uses the telescoping triangle inequality with $\frac{1}{2}\|\Lambda_\epsilon - \mathcal{I}\|_\diamond$ per two–qubit gate; it does not require gate–independent noise, only a uniform per–gate bound. The instance–dependent condition $\epsilon \le p_0/G(n,p)$ ensures $p_{\min}(\epsilon) \ge \frac{1}{2}p_0$.

Corollary 15 (No loss of mass on any fixed optimum)**.** *Fix $X^\star \in \Pi_n$. Then*

$$\Pr[R(Y) = X^\star] \;\ge\; \Pr[Y = X^\star].$$

Proof. The event $\{Y = X^\star\}$ is contained in $\{R(Y) = X^\star\}$ by Lemma 10. □

3.2 Feasibility via Swap–and–Repair and Hungarian Assignment

The NP-HQ post-processor includes scans for any *feasible* bit-string $b \in \Omega$ and application of *deterministic repair* to transform invalid outcomes to valid bitstrings in $\widetilde{b} \in \Omega$. A known algorithm to realise such repair in practice is the swap–based patching used in the Euclidean TSP PTAS [46].

Lemma 16 (Column-swap repair[46])**.** *For any bit-string $x \in \{0,1\}^{n^2}$ that has **exactly one** '1' in every row (but possibly multiple in a column) the column-swap heuristic returns $\widetilde{x} \in \Omega$ in $O(n)$ time with*

$$C(\widetilde{x}) \;\le\; C(x) \;+\; 3n\, w_{\max}, \qquad w_{\max} := \max_{i,j} M_{ij}.$$

Consequently $\Delta(n) := C(\widetilde{x}) - C^\star = O(n\, w_{\max})$ for every metric cost matrix.

In the foregoing analyses where we have focused on extracting valid permutations from one-hot encoded initial states, in the noise regime that satisfies $\epsilon \le \frac{p_0}{G(n,p)}$ the greedy SwapRepair routine (see Lemma 16) suffices. However, it assumes that each row already contains at least one "donor" 1–bit that can be moved into a vacant column, so feasibility is restored with $\le 3n$ local swaps. (c.f: the $3n$-swap fixer of Arora's PTAS for Euclidean TSP [46]). This is always possible if the quantum noise leaves one-hot structure mostly intact while feasible solutions fail to appear from it.

A **strong depolarising tail** appears once the two–qubit error rate approaches the Lipschitz threshold of Theorem 5. In that regime the raw QAOA samples may violate *both*

column and row one–hot constraints, signalling significant leakage outside the intended subspace, causing the greedy procedure to fail. Rather than increasing the shot budget (which would only linearly suppress the leakage probability) we could repair every infeasible output *deterministically* via the Hungarian assignment algorithm [20], as depicted in Algorithm 3. This would always return a valid permutation closest in Hamming distance to the noisy bit-string while still keeping the overall complexity polynomial. The procedure is to reshape the length-n^2 bitstring b into an $n \times n$ binary matrix

$$M_{ij} := \begin{cases} 1 & \text{if the qubit in row } i \text{ and column } j \text{ is 1,} \\ 0 & \text{otherwise.} \end{cases}$$

Algorithm 3 HungarianRepair

Require: measured bit–string $b \in \{0,1\}^{n^2}$
Ensure: feasible permutation b_{ok}
1: reshape b into $M \in \{0,1\}^{n \times n}$
2: $C \leftarrow -M$ ▷ cost matrix for max-overlap assignment
3: $(rows, cols) \leftarrow \text{HUNGARIAN}(C)$
4: $P \leftarrow 0^{n \times n}$; $P_{rows[k],cols[k]} \leftarrow 1$ for all k
5: flatten P to $b_{\text{ok}} \in \{0,1\}^{n^2}$
6: **return** b_{ok}

Thus, the Hungarian algorithm solves the linear assignment problem $P^\star = \arg\min_{P \in \Pi_n} \langle C, P \rangle$, $C := -M$, $\Pi_n := \{P \in \{0,1\}^{n \times n} \mid P\mathbf{1} = \mathbf{1},\ P^\mathsf{T}\mathbf{1} = \mathbf{1}\}$, where $\langle A, B \rangle = \sum_{ij} A_{ij} B_{ij}$ is the Frobenius inner-product. Because $C = -M$, minimising $\langle C, P \rangle$ is equivalent to maximising the Hamming overlap $\langle M, P \rangle$; thus $P^\star$ is the permutation matrix closest to M. Each invocation of HungarianRepair costs $O(n^3)$ classical operations.

3.3 Guarantees and Approximation Schemes.

Theorem 17 (NP-HQ is an FPRAS$_q$ in the exact–hit regime)**.** *Fix any accuracy $\varepsilon \in (0,1]$ and confidence $\delta \in (0,1)$. Suppose the p-layer CE–QAOA circuit assigns total optimal mass $p_{\min} \geq c n^{-k}$ to $\mathcal{X}^\star$, for fixed constants $c > 0$ and $k \geq 0$, and run Algorithm 2 with*

$$S \geq \left\lceil \frac{\ln(1/\delta)}{p_{\min}} \right\rceil$$

shots in total, applying the deterministic repair R to each outcome and then selecting the minimum measured cost. Then:

(a) *With probability at least $1 - \delta$, at least one repaired sample equals an optimal tour $X^\star$ and the algorithm outputs $X^\star$.*

(b) *The output therefore satisfies $C_{\text{out}} \leq (1 + \varepsilon)\, C^\star$ with probability at least $1 - \delta$ (indeed, equality $C_{\text{out}} = C^\star$).*

(c) *The total runtime is $\text{poly}(n, 1/\varepsilon, \log(1/\delta))$, with the $1/\varepsilon$–dependence being benign (degree 0).*

(d) *The quantum depth is $O(pn)$, and the number of circuit calls is $S = \text{poly}(n, \log(1/\delta))$.*

Hence NP-HQ is a fully–polynomial randomized approximation scheme in the hybrid sense (FPRAS$_q$) whenever $p_{\min}$ is inverse–polynomial in n.

Proof. Let

$$p_{\min} \;:=\; P(\mathcal{X}^\star),$$

the total probability mass assigned by p CE-QAOA layers to the optimal set $\mathcal{X}^\star$ before repair. Because the repair map R fixes every feasible bit-string, and every $x \in \mathcal{X}^\star$ is feasible, we have

$$\{Y \in \mathcal{X}^\star\} \subseteq \{R(Y) \in \mathcal{X}^\star\},$$

and therefore

$$P(R(Y) \in \mathcal{X}^\star) \;\geq\; P(Y \in \mathcal{X}^\star) \;=\; p_{\min}.$$

Now take S independent shots $Y_1, \ldots, Y_S$. The probability that *none* of the repaired samples lies in $\mathcal{X}^\star$ is

$$P(\forall s,\; R(Y_s) \notin \mathcal{X}^\star) = \prod_{s=1}^{S} P(R(Y_s) \notin \mathcal{X}^\star) \leq (1 - p_{\min})^S \leq e^{-p_{\min} S}.$$

Hence, if

$$S \;\geq\; \left\lceil \frac{\ln(1/\delta)}{p_{\min}} \right\rceil,$$

then

$$e^{-p_{\min} S} \leq \delta,$$

so with probability at least $1 - \delta$ there exists at least one repaired sample $R(Y_s) \in \mathcal{X}^\star$. This proves (a).

On that event, at least one repaired sample has objective value exactly $C^\star$. Since Algorithm 2 returns the repaired sample of minimum cost among all S repaired candidates, the output must also have cost $C^\star$. Therefore

$$C_{\text{out}} = C^\star,$$

and in particular

$$C_{\text{out}} \leq (1 + \varepsilon) C^\star \qquad \text{for every } \varepsilon \in (0, 1].$$

This proves (b).

For (c), each shot consists of one depth-p quantum circuit execution with p layers, followed by one deterministic repair and one cost evaluation. With Hungarian repair, the classical work per shot is polynomial in n (e.g. $O(n^3)$), so the total runtime is

$$\text{poly}(S, n).$$

In the exact-hit regime of the theorem, when $p_{\min}^{-1}$ is polynomial in n, the chosen shot count

$$S = O\Big(p_{\min}^{-1} \log(1/\delta)\Big)$$

is polynomial in n and $\log(1/\delta)$. Since there is no actual dependence on $1/\varepsilon$, the runtime is

$$\text{poly}(n,\, 1/\varepsilon,\, \log(1/\delta))$$

with degree 0 in $1/\varepsilon$. This proves (c).

Finally, (d) follows because the number of CE–QAOA layers is p. For fixed p, the effective quantum depth is $O(pn)$, and the number of circuit executions is exactly S which is polynomial in n and $\log(1/\delta)$ in the exact-hit regime. Therefore NP-HQ is an exact-hit FPRAS_q whenever $p_{\min}$ is inverse-polynomial in n. □

Corollary 18 (FPRAS_q under noise)**.** *Under the hypothesis of Theorem 5, we have* $p_{\min}(\epsilon) = \Omega(n^{-k-1})$. *Choosing* $S \geq \lceil \ln(1/\delta)/p_{\min}(\epsilon) \rceil = O(n^{k+1} \ln(1/\delta))$ *in Algorithm 2 yields the same FPRAS$_q$ guarantee as Theorem 17 on a noisy device.*

Proposition 19 (Promised multiplicative fallback via repair)**.** *Assume the metric satisfies a positive lower bound* $w_{\min} > 0$ *and let* $\kappa := w_{\max}/w_{\min}$*. If* $\varepsilon \geq c\,\kappa$ *for a fixed constant* $c > 0$*, then the repair additive gap* $\Delta(n) = O(n\, w_{\max})$ *implies* $C_{\text{rep}} \leq (1+\varepsilon)\, C^\star$ *whenever the optimum is missed. Thus with* S *as in Theorem 17, NP-HQ attains a* $(1+\varepsilon)$ *approximation with probability at least* $1-\delta$ *even conditioned on miss, under this promise.*

Because NP-HQ processes $S = \text{poly}(n)$ shots, the post–processing stage scales as $O(S\, n^3)$, leaving the overall runtime polynomial in $(n, 1/\varepsilon, \log 1/\delta)$ fully consistent with Theorem 17. HungarianRepair therefore provides a robust, parameter–free fallback for the high–noise regime, extending NP-HQ's applicability.

By optimality of the Hungarian assignment, $P^\star$ maximizes the Hamming–overlap $\langle M, P\rangle$ among all permutations, making HungarianRepair the canonical "least–invasive" fix in the Hamming metric. While the analysis is couched in the language of metric TSP, nothing in the probabilistic argument relies on subtleties specific to tours. Instead, the arguments rely on the existence of (i) a clearly *specifiable target structure* that a classical post-processing step can recognise (e.g., a valid schedule, matching, partition, etc.), and (ii) a repair routine that converts any raw bit-string into a feasible one at a known additive (or multiplicative) cost (e.g., swap-repair, projection, greedy repair). Whenever a problem class admits such a structure–checker–and–repair trio, the same feasibility-first reasoning applies, and analogous promised approximation guarantees follow once the relevant additive or multiplicative repair bound is established.

3.4 Quantum Advantage: Quantum–classical separation

This section formalizes a separation between the hybrid CE–QAOA/NP-HQ pipeline and *purely classical* "sample + Hungarian" strategies. The key resource is the inverse–polynomial *optimal* mass guaranteed by the CE kernel on the *kernel-admissible* (promise) family of instances, together with standard success amplification. Throughout, complexity consequences are conditional on the standard assumption $\mathbf{NP} \not\subseteq \mathbf{BPP}$.

A clarifying remark is in order. The classical baseline in Theorems 21 and 22 is already equipped with the *same* Hungarian repair map $R(\cdot)$ used by NP-HQ. The separation therefore does not arise from any asymmetry in the classical post-processing stage. Rather, it is located entirely in the *sampling distribution*: the CE–QAOA kernel biases amplitude toward the global optimum so that it receives inverse-polynomial mass $p_{\min}(I) = \Omega(n^{-k})$, whereas no efficiently samplable classical distribution can achieve inverse-polynomial overlap with the (a priori unknown) global optimum uniformly over instances in $\mathcal{K}_n$, unless $\mathbf{NP} \subseteq \mathbf{BPP}$. In other words, if a classical algorithm could reproduce the CE–QAOA output distribution, it would immediately inherit the same FPRAS_q guarantee via identical post-processing. The computational hardness lies in generating that distribution classically, not in exploiting it once it is available.

Classical baseline: sample + Hungarian. We compare NP-HQ against any classical randomized algorithm that may: (i) generate $\text{poly}(n)$ samples from *any* efficiently samplable distribution on $\{0,1\}^{n\times n}$ (optionally already restricted to row–one–hot or permutations), (ii) apply the Hungarian projection $R(\cdot)$ (nearest-permutation projection in Hamming distance) to each sample in $O(n^3)$ time, and (iii) compute the objective cost $C(\cdot)$ exactly and output the best repaired tour. This is the strongest natural "purely classical" analogue of NP-HQ if one insists on the same repair primitive.

Chen–Cotler–Huang–Li oracle model. Definition 2.1 of [2] calls a promise problem *NISQ-solvable* if there exists a classical probabilistic polynomial-time algorithm that may query polynomially many times an oracle quantum device which

(i) initialises $\mathrm{poly}(n)$ noisy qubits in $|0\rangle^{\otimes N}$,

(ii) applies a polynomial-length circuit of one- and two-qubit gates, each followed by an independent constant-rate depolarising channel, and

(iii) measures all qubits in the computational basis.

Our CE–QAOA layer uses only one- and two-qubit gates; NP-HQ queries the device $S = \mathrm{poly}(n)$ times and post-processes each outcome in $\mathrm{poly}(n)$ time. Hence NP-HQ fits the NISQ oracle model as a *model-of-computation* statement. The stronger FPRAS_q guarantees established in Sec. 2.5 additionally require a *low-noise* regime.

Corollary 20 (NP-HQ is NISQ-solvable)**.** *For every instance size n and constant depth p, the CE–QAOA/NP-HQ pipeline can be implemented by a polynomial-time classical algorithm with oracle access to a noisy quantum device of the type specified in [2].*

Theorem 21 (No classical "sample + Hungarian" strategy matches NP-HQ on $\mathcal{K}_n$ unless $\mathbf{NP} \subseteq \mathbf{BPP}$)**.** *Fix any constant $k \geq 0$ and constant depth p. Consider the promise family $\mathcal{K}_n$ of kernel-admissible instances. Suppose there exists a* purely classical *probabilistic polynomial-time algorithm $\mathcal{C}$ that, for every $I \in \mathcal{K}_n$, outputs an* ***optimal*** *tour with probability $\Omega(n^{-k})$, while only using:*

(i) $\mathrm{poly}(n)$ *samples from any efficiently samplable distribution on $\{0,1\}^{n\times n}$,*

(ii) Hungarian projection $R(\cdot)$ on each sample in $O(n^3)$ time, and

(iii) total runtime $\mathrm{poly}(n)$*.*

Then the corresponding promise *search problem on $\mathcal{K}_n$ lies in* **BPP***. In particular, if $\mathcal{K}_n$ contains an NP-hard subfamily under polynomial-time reductions, then* $\mathbf{NP} \subseteq \mathbf{BPP}$*.*

Consequently, assuming $\mathbf{NP} \not\subseteq \mathbf{BPP}$*, no such classical sampler exists on any NP-hard kernel-admissible family, whereas NP-HQ achieves inverse–polynomial optimal-hit probability on $\mathcal{K}_n$ in the low-noise regime of Sec. 2.5.2 and Theorem 5, and therefore constitutes an exact-hit FPRAS_q there.*

Proof sketch. Assume $\mathcal{C}$ succeeds with probability at least $c\,n^{-k}$ on every $I \in \mathcal{K}_n$. Repeat $\mathcal{C}$ independently $T = \Theta(n^k \log n)$ times and return the best tour found. Then the probability of missing the optimum in all trials is at most $(1 - c\,n^{-k})^T \leq n^{-\Omega(1)}$, so we obtain a polynomial-time randomized algorithm that solves the *promise* search problem on $\mathcal{K}_n$ with high probability. If $\mathcal{K}_n$ is NP-hard under reductions (or contains an NP-hard subfamily), this implies $\mathbf{NP} \subseteq \mathbf{BPP}$. □

3.5 Limits of Classical Samplers with perfect structural oracles

We now strengthen the classical side by granting the sampler *perfect access* to the constraint structure. The goal is to rule out the objection that NP-HQ's advantage comes merely from sampling outside the feasible region and then repairing.

Classical oracle models. Let $L = n^2$ and view samples as bit-matrices in $\{0,1\}^{n\times n}$. We allow the classical sampler one of the following increasingly generous oracles:

(A) **Unconstrained sampler:** outputs arbitrary $Y \in \{0,1\}^{n\times n}$.

(B) **Row–one–hot sampler:** outputs only matrices with exactly one 1 in each row (no column guarantee), i.e. $Y \in \mathcal{S}_{\text{row}}$ with $|\mathcal{S}_{\text{row}}| = n^n$.

(C) **Uniform permutation oracle:** outputs a uniformly random permutation matrix $Y \in \mathcal{S}_{\text{perm}} = \Pi_n$ with $|\Pi_n| = n!$.

In all three cases, the algorithm may post-process every sample via the Hungarian projection $R(\cdot)$ in $O(n^3)$ time and may repeat the entire procedure $\text{poly}(n)$ times.

Theorem 22 (Perfect structural oracles do not yield inverse–polynomial optimal overlap)**.** *Fix any constant k and any promise family $\mathcal{K}_n$. If a classical randomized algorithm—equipped with any oracle (A)–(C) above and allowed* $\text{poly}(n)$ *total time—outputs an* exact *optimal tour with probability $\Omega(n^{-k-1})$ on* every *instance $I \in \mathcal{K}_n$, then the promise search problem on $\mathcal{K}_n$ lies in* **BPP***; in particular, if $\mathcal{K}_n$ contains an NP-hard subfamily under reductions then* **NP** $\subseteq$ **BPP**.

Proof. The proof is identical to Theorem 21: inverse–polynomial success amplifies to constant success probability under polynomially many repetitions, yielding a **BPP** algorithm for the promise search problem. The argument never uses which oracle (A)–(C) generated the samples; hence even a perfect permutation oracle cannot bypass it. □

Connection to NP-HQ. Oracles (B) and (C) are not assumed hard to implement. In fact, sampling a uniform permutation takes $O(n \log n)$ random bits and time [48]. The bottleneck is that achieving *inverse–polynomial overlap with the (a priori unknown) global optimum uniformly over instances* is tantamount to solving the underlying search problem on the family. In contrast, on the kernel-admissible family $\mathcal{K}_n$ the CE layer biases amplitude so that the *optimal* tour receives inverse–polynomial mass (with suitable instance-dependent angles), and deterministic Hungarian projection guarantees feasibility of every reported output. In the low-noise regime where the optimal mass remains inverse–polynomial (Sec. 2.5), success amplification yields an exact-hit FPRAS_q for NP-HQ on $\mathcal{K}_n$. Importantly, Hungarian repair is used here as a feasibility projection; no cost-preservation property is assumed of Hungarian projection beyond feasibility and its total-variation contraction property (Lemma 12).

In the noiseless setting and under the limited-noise condition of Theorem 5 ensuring inverse–polynomial optimal mass $p_{\min}(\epsilon) = \Omega(n^{-k-1})$, NP-HQ is an *exact-hit FPRAS$_q$*. With $S = \Theta(\ln(1/\delta)/p_{\min}(\epsilon)) = \text{poly}(n), \ln(1/\delta)$ device queries it returns a global optimum with probability at least $1-\delta$. Outside this low-noise regime, exact hits may disappear; nevertheless NP-HQ remains a polynomial-time *quantum approximation scheme* in the pragmatic sense that it outputs a feasible solution together with a provable instance-class-dependent approximation floor whenever the repair pipeline admits an additive (or multiplicative) performance bound.

4 Demonstration on Superconducting Quantum Processors

Combined with Hungarian repair (Sec. 3.2), the NP-HQ pipeline returns a *feasible* candidate deterministically, with classical post-processing cost $O(Sn^3)$ when Hungarian repair

is applied to all samples. The practical question addressed in this section is whether the underlying CE–QAOA sampler remains informative on present-day superconducting hardware at problem sizes approaching 10^2 qubits. To test this, we run single-layer CE–QAOA circuits on 127-qubit IBM Eagle-r3 processors for a family of QOptlib TSP subset instances and apply the same deterministic repair routine to all measured outcomes. The hardware study is meant to demonstrate that shallow CE–QAOA circuits still produce useful sampling bias in a strongly noisy regime and that deterministic repair restores feasibility reliably with high quality objective costs on the repaired samples.

Recall that the prolem Hamiltonian is diagonal and contains only Z and ZZ terms. Accordingly, the dominant hardware overhead comes from scheduling commuting two-qubit phase interactions under the device connectivity constraints, while the mixer acts through native or near-native two-qubit excitation-preserving gates. This makes the present implementation a natural testbed for the constraint-preserving CE–QAOA architecture on current superconducting devices.

All kernel assumptions of Def. 1 continue to hold for the benchmark family considered here, with $t_{\max} = O(n)$, so the hardware study can be viewed as a direct implementation of NP-HQ with CE–QAOA as the front-end sampler.

4.1 Benchmark set and warm-start parameters

We consider seven QOptlib TSP subset instances with $n = 4, \ldots, 10$ cities, corresponding to $n^2 = 16$ to 100 logical qubits [3]. To avoid instance-by-instance retuning on hardware, we reuse $p = 1$ angles obtained from generic-QAOA optimization and feed them directly into CE–QAOA. This choice is motivated by the parameter-transfer picture developed earlier in the manuscript: rather than re-optimizing on noisy hardware, we test whether CE–QAOA can extract useful performance from previously optimized shallow parameters.

Table 1 lists the generic-QAOA warm-start angles used throughout this section. No further parameter tuning was performed on the quantum hardware.

Instance	γ^{gen}	β^{gen}	Best energy (generic)
wi4	0.233	2.025	-4.36×10^4
wi5	1.307	2.387	-6.11×10^4
wi6	0.748	0.741	-4.14×10^4
wi7	2.172	0.577	-7.66×10^4
dj8	0.239	0.460	-3.18×10^4
dj9	0.002	2.507	-1.72×10^6
dj10	1.835	2.399	-1.68×10^4

Table 1: Generic-QAOA warm-start angles at $p = 1$, reused verbatim in the CE–QAOA hardware runs.

4.2 Experimental protocol

All circuits were implemented in Python and transpiled with Qiskit 1.1.2 [49] using device-aware coupling maps and the native two-qubit `ECR`/`iSWAP` basis where applicable. For each (instance, chip) pair we executed a single CE–QAOA layer ($p = 1$) with the warm-start angles

$$(\gamma, \beta) = (\gamma^{\text{gen}}, \beta^{\text{gen}}),$$

and then applied Hungarian repair (Alg. 3) to every measured bit-string.

Measurement-error mitigation was enabled, and dynamical decoupling and Pauli twirling were activated when supported by the execution workflow. For each run we recorded: the transpiled depth and two-qubit gate count, the hardware calibration snapshot, feasible-sample fraction before and after repair, best repaired objective value $C_{\min}$, and end-to-end turnaround time.

4.3 `ibm_brisbane` (127 qubits, Eagle r3)

Device snapshot (21 June 2025 09:30 UTC).

Metric	Median value	Best
Qubits	127	–
Two-qubit ECR error ϵ_{2Q}	6.99×10^{-3}	2.67×10^{-3}
Readout error	1.61×10^{-2}	5.1×10^{-3}
T_1, T_2 (μs)	235.7, 122.4	–

The median two-qubit error rate lies in the 10^{-3} to 10^{-2} regime, so these runs probe CE–QAOA well outside an idealized noiseless setting. The purpose of the experiment is therefore not to test perfect coherent evolution, but to assess whether shallow constrained sampling combined with deterministic repair remains useful under realistic device noise.

Transpilation statistics. Table 2 reports the transpiled depth, two-qubit gate count, and turnaround time obtained on `ibm_brisbane`. Dynamical decoupling and Pauli twirling were enabled in the reported transpilation pipeline, and the resulting circuits reflect the cost of embedding the encoded CE–QAOA structure into the heavy-hex connectivity.

Table 2: Logical qubit count (n^2), two-qubit gates, physical depth, and end-to-end turnaround per instance on `ibm_brisbane`. Wall time includes queue and execution overhead.

Instance	Qubits	2Q gates	Depth	Wall [s]
wi4	16	509	771	761
wi5	25	1 387	1 787	80
wi6	36	2 834	3 060	137
wi7	49	5 141	4 706	228
dj8	64	8 295	7 221	317
dj9	81	12 642	10 695	499
dj10	100	18 770	12 962	719

4.4 Hardware Discussion.

- The transpiled circuits are deep and strongly noise-affected, especially for the `dj` instances.
- Hungarian repair restores feasibility deterministically for every sampled outcome, so the post-processed output is always a valid permutation.
- The local classical post-processing cost is negligible compared with the queue and hardware execution overhead, even though every sample is repaired separately.
- For the benchmark set considered here, the best repaired tours are competitive with the published QOPTLIB references and, for the harder instances, improve on them. The raw and post-repair feasible fractions for `ibm_brisbane` are shown in Fig. 1.

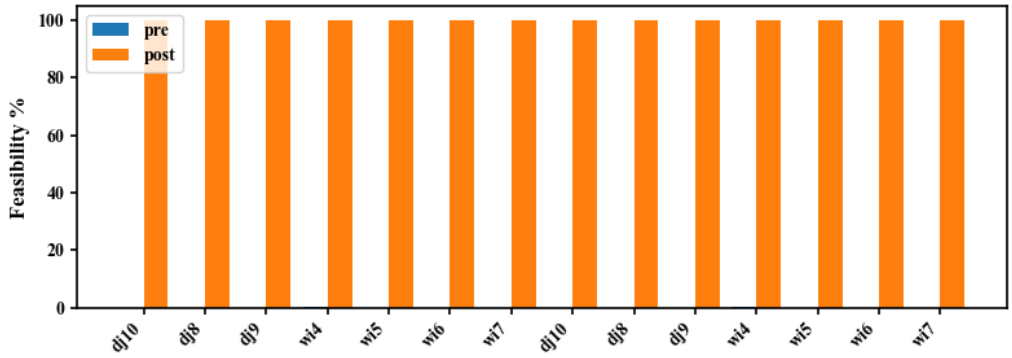


Figure 1: Raw versus post-repair feasible fraction for the QOPTLIB benchmark instances executed on `ibm_brisbane`. Hungarian repair maps all samples back into the feasible permutation set.

4.4.1 Comparison with published QOPTLIB tours

Table 3 compares the repaired CE–QAOA output on `ibm_brisbane` against the reference tours distributed with QOPTLIB [3]. The smaller `wi` instances are matched exactly, while the harder `dj` subset shows modest to double-digit improvements according to the reported fixed-shot comparison.

Table 3: Repaired CE–QAOA output on `ibm_brisbane` versus the reference tours distributed with QOPTLIB. Positive % improvement means a shorter tour than the published reference.

Instance	QOPTLib cost	Our cost	Δ	Improvement [%]
wi4	6 700	6 700	0	0.00
wi5	6 786	6 786	0	0.00
wi6	9 815	9 815	0	0.00
wi7	7 245	7 245	0	0.00
dj8	2 794	2 762	−32	**1.15**
dj9	2 438	2 134	−304	**12.5**
dj10	3 155	2 822	−333	**10.55**

Taken at face value, the standardized comparison indicates that even in a strongly noisy regime the repaired CE–QAOA samples remain competitive with the reference benchmark tours. Figure 2 shows the corresponding raw and post-repair feasible fractions for `ibm_sherbrooke`. The transpilation statistics for `ibm_sherbrooke`, including two-qubit gate counts, physical depth, and turnaround times, are reported in Table 4.

4.5 `ibm_sherbrooke` (127 qubits, Eagle r3)

Device snapshot (21 June 2025, 07:10 UTC).

Metric	**Median**	**Best**
Two-qubit ECR error ϵ_{2Q}	7.13×10^{-3}	2.55×10^{-3}
Read-out error	2.08×10^{-2}	3.9×10^{-3}
Single-qubit SX error	2.27×10^{-4}	6.8×10^{-5}
T_1, T_2 coherence (µs)	262, 212	–
Calib. version	1.6.112 (Eagle r3)	

Table 4: Logical-qubit count (n^2), two-qubit ECR gates, physical depth, and end-to-end turnaround on `ibm_sherbrooke`.

Instance	Qubits	2Q gates	Depth	Wall [s]
wi4	16	547	907	181
wi5	25	1 477	1 964	74
wi6	36	2 876	3 031	150
wi7	49	5 060	4 842	689
dj8	64	8 507	6 896	551
dj9	81	12 827	9 524	574
dj10	100	18 768	11 698	728

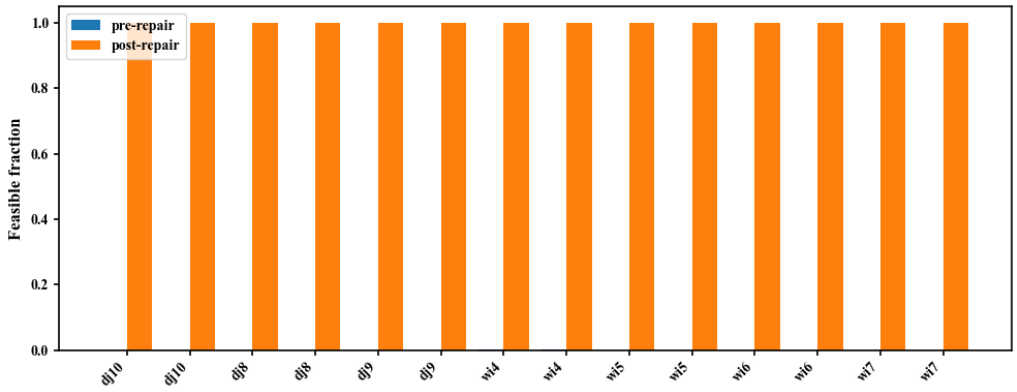


Figure 2: Raw versus post-repair feasible fraction for `ibm_sherbrooke`. As on `ibm_brisbane`, raw feasibility is very small and deterministic repair restores all outputs to feasibility.

Table 5: Best repaired tours obtained on `ibm_sherbrooke` in the standardized comparison, together with the QOPTLIB reference tours [3].

Instance	Qubits	Our cost	QOPTLib cost	Δ	Improvement [%]
dj10	100	2 909	3 155	−246	**7.8**
dj9	81	2 134	2 438	−304	**12.5**
dj8	64	2 762	2 794	−32	**1.1**
wi7	49	7 245	7 245	0	0.0
wi6	36	9 815	9 815	0	0.0
wi5	25	6 786	6 786	0	0.0
wi4	16	6 700	6 700	0	0.0

Transpilation statistics. The `ibm_sherbrooke` runs are qualitatively similar to the `ibm_brisbane` results. The transpiled circuits are again deep and strongly noise-affected; raw feasibility is extremely small; and deterministic Hungarian repair maps every sample back to a valid permutation. Despite this, the repaired samples remain competitive with the benchmark references, matching the known `wi` solutions and improving several `dj` instances in the standardized comparison. Table 5 compares the best repaired tours obtained on `ibm_sherbrooke` with the corresponding QOptlib reference tours.

5 HH-QAOA: Geometry Reduced Shot Analyses Preserving Guarantees

We propose Heavy-Hitter QAOA, a Heavy-Hitter theory that leverages geometry and quantum information to equip the NP-HQ pipeline with a memory-efficient and time-efficient classical post-processing scheme. The key idea is that the classical checker can focus on the small set of bit-strings that received the highest sampling bias from the quantum core. Under purely classical heavy-hitter bounds the retained candidate set has size $O(n^{k+1})$, independent of the raw shot count, leading to post-processing cost $O(n^{k+4})$. By exploiting the inverse-polynomial optimum mass guaranteed by the CE–QAOA kernel, the admissible threshold can be raised by one factor of n, reducing the candidate set to $O(n^k)$ and the post-processing cost to $O(n^{k+3})$. This reduction comes without weakening the approximation guarantees established in the preceding sections.

Definition 23 (Heavy hitter)**.** *Fix integers $n, k \geq 1$ and let $\tau = \Theta(n^{-(k+1)})$. For a probability distribution P on $\{0,1\}^n$, a bit-string $y \in \{0,1\}^n$ is called a* heavy hitter *if its probability mass satisfies $P(y) \geq \tau$. Empirically, given S samples $\{y^{(s)}\}_{s=1}^{S}$, y is a heavy hitter when its observed frequency obeys $\widehat{P}_S(y) \geq \tau$.*

5.1 Speeding up the Analyses of Quantum-Generated Samples

Relation to the exact-hit regime. We propose heavy-hitter refinement as a post-processing refinement of NP-HQ algorithm. Indeed, when the optimal mass satisfies $p_{\min} = \Theta(n^{-k})$, the exact-hit shot budget is already $S = \Theta(n^k \log(1/\delta))$. In that regime, heavy-hitter thresholding does not change the fundamental success exponent controlling the total number of required samples. Its role is instead to organize the retained candidate set more efficiently and to expose additional practical reductions in memory and classical workload, especially in oversampled and hardware-scale settings where the raw histogram is much larger than the minimal exact-hit budget. Accordingly, the theorem-level guarantees in this section should be read as refinements of post-processing, while the stronger reductions seen under block-prefiltering and top-L compression are empirical or instance-dependent. Yet, as we argue below, most of that support is irrelevant for post-processing efforts to identify the global optimum because a set of *heavy hitters* captures essentially all the useful weight containing it. First we invoke classical heavy-hitter theory to quantify this intuition for arbitrary distributions. Then in CE-QAOA we leverage additional algebraic, geometric and information-theoretic structure to shrink the heavy-hitter list even further, culminating in what we call a *Quantum Heavy-Hitter Theory* (Sec. 5.3). The effect of circuit-depth compression on the heavy-hitter threshold and retained candidate set is illustrated in Fig. 3. Figure 4 summarizes the reduction in candidate-set size and classical post-processing complexity.

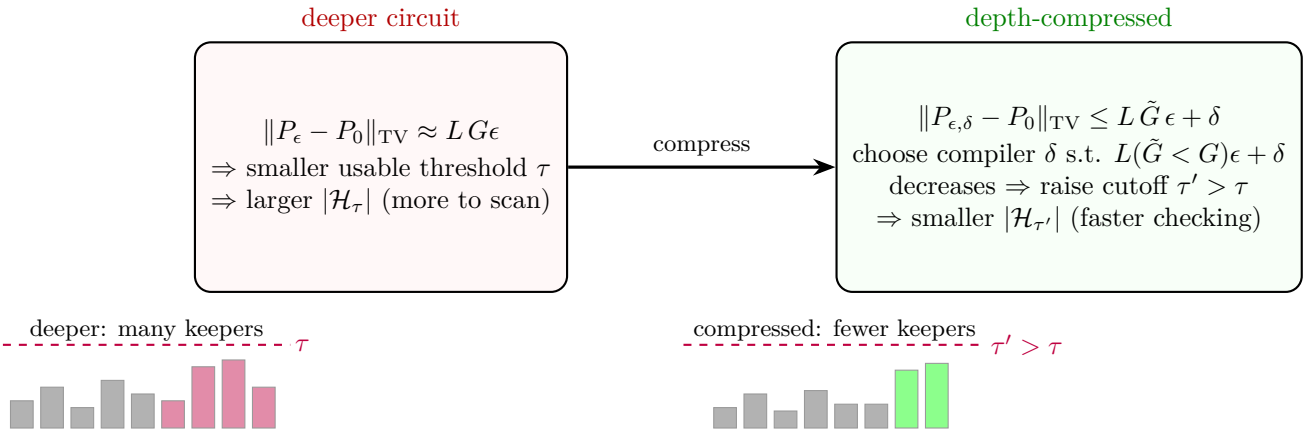


Figure 3: Depth-compression effect. Trading a tiny compiler error δ for a large gate-count reduction $\tilde{G} < G$ reduces total TV error, lets you raise τ, and shrinks the heavy-hitter set. The associated circuit compression is discussed in App. A.8

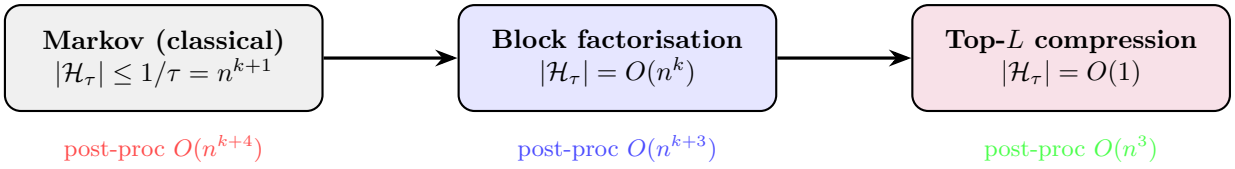


Figure 4: Shrinking the scan set. Structure in CE–QAOA collapses the heavy-hitter list from n^{k+1} to $O(1)$, cutting classical time from $O(n^{k+4})$ to $O(n^3)$.

5.2 Classical Heavy-Hitter Theory

Setup. Let p be any probability distribution supported on a finite feasible set Ω of size D. For a threshold $\tau > 0$, define the heavy-hitter set

$$\mathcal{H}_\tau \;:=\; \{x \in \Omega : p(x) \ge \tau\}, \qquad H_\tau \;:=\; |\mathcal{H}_\tau|.$$

Definition 24 (Heavy-hitter set)**.** *For a distribution p on Ω and threshold $\tau > 0$, the* heavy-hitter set *is*

$$\mathcal{H}_\tau \;:=\; \{x \in \Omega : p(x) \ge \tau\}.$$

Lemma 25 (Cardinality bound)**.** *For every distribution p and every threshold $\tau > 0$,*

$$H_\tau \;\le\; \frac{1}{\tau}.$$

Proof. By definition, each $x \in \mathcal{H}_\tau$ contributes at least τ mass, so

$$1 \;=\; \sum_{x\in\Omega} p(x) \;\ge\; \sum_{x\in\mathcal{H}_\tau} p(x) \;\ge\; H_\tau\,\tau.$$

Hence $H_\tau \le 1/\tau$. □

Lemma 26 (Empirical retention of heavy hitters)**.** *Let $X_1, \ldots, X_S$ be i.i.d. samples from p, and let*

$$\widehat{p}_S(x) \;:=\; \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\{X_s = x\}$$

be the empirical frequency. Define the empirical candidate set

$$\widehat{\mathcal{H}}_{\tau/2} \;:=\; \{x \in \Omega : \widehat{p}_S(x) \ge \tau/2\}.$$

Then the following hold.

(i) *For every* $x \in \mathcal{H}_\tau$,

$$\Pr\left[\widehat{p}_S(x) < \frac{\tau}{2}\right] \;\le\; \exp\left(-\frac{\tau S}{8}\right).$$

(ii) *Consequently,*

$$\Pr\left[\mathcal{H}_\tau \subseteq \widehat{\mathcal{H}}_{\tau/2}\right] \;\ge\; 1 - \frac{1}{\tau}\exp\left(-\frac{\tau S}{8}\right).$$

(iii) *In every sample realization,*

$$|\widehat{\mathcal{H}}_{\tau/2}| \;\le\; \frac{2}{\tau}.$$

In particular, if

$$S \;\ge\; \frac{8}{\tau}\log\frac{1}{\tau\delta},$$

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

$$\mathcal{H}_\tau \subseteq \widehat{\mathcal{H}}_{\tau/2} \quad \textit{and} \quad |\widehat{\mathcal{H}}_{\tau/2}| \le \frac{2}{\tau}.$$

Proof. For $x \in \mathcal{H}_\tau$, the count $N_x := \sum_{s=1}^{S} \mathbf{1}\{X_s = x\}$ is binomial with mean $\mathbb{E}[N_x] = S\,p(x) \ge S\tau$. By the multiplicative Chernoff bound,

$$\Pr\left[N_x < \frac{1}{2}\mathbb{E}[N_x]\right] \;\le\; \exp\left(-\frac{\mathbb{E}[N_x]}{8}\right) \;\le\; \exp\left(-\frac{\tau S}{8}\right),$$

which proves (i). Since $|\mathcal{H}_\tau| \le 1/\tau$ by Lemma 25, a union bound gives (ii).

For (iii), every $x \in \widehat{\mathcal{H}}_{\tau/2}$ contributes at least $\tau/2$ empirical mass, so

$$1 \;=\; \sum_{x\in\Omega} \widehat{p}_S(x) \;\ge\; \sum_{x\in\widehat{\mathcal{H}}_{\tau/2}} \widehat{p}_S(x) \;\ge\; |\widehat{\mathcal{H}}_{\tau/2}|\,\frac{\tau}{2}.$$

Hence $|\widehat{\mathcal{H}}_{\tau/2}| \le 2/\tau$. The final statement follows by choosing S so that $\tau^{-1}e^{-\tau S/8} \le \delta$. □

Remark 27 (Exact recovery requires a margin). Without an explicit separation gap around the threshold τ, one should not claim exact recovery of $\mathcal{H}_\tau$ from finitely many samples. What Lemma 26 guarantees is the more robust statement that all true τ-heavy strings are retained inside a slightly relaxed empirical set of size $O(1/\tau)$. Exact identification of the thresholded set requires an additional margin assumption, for example

$$p(x) \ge \tau + \eta \quad \text{for } x \in \mathcal{H}_\tau, \qquad p(x) \le \tau - \eta \quad \text{for } x \notin \mathcal{H}_\tau,$$

together with a Hoeffding/union-bound argument over Ω.

Corollary 28 (Classical heavy-hitter post-processing cost)**.** *Suppose the deterministic checker/repair/scoring stage costs* $O(n^3)$ *per candidate. If the empirical candidate set is chosen as* $\widehat{\mathcal{H}}_{\tau/2}$*, then with probability at least* $1-\delta$ *all true* τ*-heavy strings are retained and the number of candidates is at most* $2/\tau$*. Hence the total post-processing cost is*

$$O\left(\frac{n^3}{\tau}\right).$$

In particular, for the choice

$$\tau = n^{-(k+1)},$$

the retained set has size $O(n^{k+1})$ *and the post-processing cost is*

$$O(n^{k+4}).$$

Proposition 29 (Worst-case sharpness of the cardinality bound)**.** *The bound $H_\tau \le 1/\tau$ is worst-case sharp up to rounding. More precisely, for every $\tau \in (0,1]$, there exists a distribution p such that*

$$H_\tau = \left\lfloor \frac{1}{\tau} \right\rfloor .$$

Proof. Let $N = \lfloor 1/\tau \rfloor$ and define p to be uniform on N points:

$$p(x) = \frac{1}{N} \qquad \text{for } x \in \{x_1, \dots, x_N\}.$$

Since $N \le 1/\tau$, $1/N \ge \tau$, so all N support points are τ-heavy. Thus $H_\tau = N = \lfloor 1/\tau \rfloor$. □

5.3 Quantum Heavy-Hitter Refinements

The classical heavy-hitter bound of Sec. 5.2 applies to *any* output distribution and therefore ignores the additional structure present in CE–QAOA. In the present setting, a quantum-informed threshold uses the inverse-polynomial optimum weight proved for the CE kernel together with the TV-noise bound to raise the heavy-hitter cutoff from the generic scale $n^{-(k+1)}$ to the sharper scale n^{-k}. So the main refinement comes when one connects the lower bound on the ideal optimum weight and a total-variation stability bound under noise. These two ingredients allow one to choose a *larger* heavy-hitter threshold than in the generic classical analysis. For this refinement, we impose the stronger condition that at least one optimal string carries inverse-polynomial probability mass and state the result as follows.

Theorem 30 (Quantum-informed noisy heavy hitters)**.** *Let $x^\star$ denote the optimal feasible bit-string. Assume the ideal CE–QAOA output distribution p_0 satisfies*

$$p_0(x^\star) \;\ge\; c\, n^{-k}$$

for fixed constants $c > 0$ and $k \ge 0$, and let p_ϵ be the noisy output distribution. Suppose further that

$$\|p_\epsilon - p_0\|_{\mathrm{TV}} \;\le\; \Delta_n, \qquad \Delta_n \le \frac{c}{4}\, n^{-k}.$$

Define the quantum-informed threshold $\tau_q := \frac{c}{4}\, n^{-k}$. Then:

(i) *the optimal string remains heavy under noise, $p_\epsilon(x^\star) \;\ge\; \tau_q$;*

(ii) *the noisy heavy-hitter set at threshold τ_q, $\mathcal{H}_{\tau_q}(p_\epsilon) := \{x : p_\epsilon(x) \ge \tau_q\}$, has cardinality $|\mathcal{H}_{\tau_q}(p_\epsilon)| \;\le\; \frac{1}{\tau_q} \;=\; \frac{4}{c}\, n^k$.*

Proof. For the singleton event $\{x^\star\}$, total variation distance gives

$$|p_\epsilon(x^\star) - p_0(x^\star)| \;\le\; \|p_\epsilon - p_0\|_{\mathrm{TV}} \;\le\; \Delta_n.$$

Hence

$$p_\epsilon(x^\star) \;\ge\; p_0(x^\star) - \Delta_n \;\ge\; c\, n^{-k} - \frac{c}{4}\, n^{-k} \;=\; \frac{3c}{4}\, n^{-k} \;\ge\; \tau_q,$$

which proves (i). Part (ii) is Lemma 25 applied with $\tau = \tau_q$. □

Corollary 31 (Empirical candidate list at quantum-informed threshold)**.** *Under the hypotheses of Theorem 30, define the empirical candidate set*

$$\widehat{\mathcal{H}}_{\tau_q/2} \;:=\; \{x : \widehat{p}_S(x) \ge \tau_q/2\}.$$

If $S \;\ge\; \frac{8}{\tau_q} \log \frac{1}{\tau_q \delta} \;=\; O\big(n^k \log(n/\delta)\big)$, then with probability at least $1-\delta$,

$$x^\star \in \widehat{\mathcal{H}}_{\tau_q/2} \qquad \text{and} \qquad |\widehat{\mathcal{H}}_{\tau_q/2}| \;\le\; \frac{2}{\tau_q} \;=\; \frac{8}{c}\, n^k.$$

Proof. Apply Lemma 26 from the classical heavy-hitter subsection with $\tau = \tau_q$, together with Theorem 30, which ensures that $x^\star \in \mathcal{H}_{\tau_q}(p_\epsilon)$. □

The generic classical heavy-hitter analysis used a threshold of order $n^{-(k+1)}$, leading to a retained set of size $O(n^{k+1})$. Theorem 30 uses the kernel guarantee to obtain the sharper threshold $\tau_q = \Theta(n^{-k})$. This immediately improves the retained-set size by one factor of n, from $O(n^{k+1})$ to $O(n^k)$, while preserving the optimum with high probability after polynomially many shots.

5.4 Heavy-Hitter Extraction and Runtime Gains

For the empirical evaluation, we supplement the theorem-level heavy-hitter threshold with two frequency-ranked compression steps. After forming the global empirical heavy-hitter set $\mathcal{C}_{\mathrm{HH}}$, we apply an empirical *block filter* by retaining its n^k most frequent elements:

$$\mathcal{C}_{\mathrm{blk}} := \mathrm{Top}_{n^k}(\mathcal{C}_{\mathrm{HH}}).$$

We may then retain only the L most frequent elements of the block-filtered set:

$$\mathcal{C}_L := \mathrm{Top}_L(\mathcal{C}_{\mathrm{blk}}).$$

If deterministic checking, repair, and scoring cost $O(n^3)$ per retained candidate, then

$$T_{\mathrm{post}} = O(|\mathcal{C}|n^3),$$

where $\mathcal{C}$ is the retained candidate set. The first pipeline uses only the generic classical heavy-hitter bound of Sec. 5.2. The second uses the sharper *quantum-informed* threshold from Theorem 30.

5.4.1 Empirical Footprints of Heavy-Hitter Filtering

We now examine the empirical footprint of the three post-processing stages introduced in Sec. 5.4: the *classical heavy-hitter filter*, the *block filtering stage*, and the optional *top-L* empirical compression. The purpose of this section is to quantify the practical reduction in candidate-set size observed on representative instances. For each run, we report the retained candidate count after the classical thresholding stage, the count after block-prefiltering, and the count after the optional top-L compression step. We also record the percentage cost gap relative to the best solution obtained from the full retained histogram.

Table 6: Empirical candidate-set footprints and percentage cost gaps for $k = 3$ with top-L window $L = 25k = 75$. Here $|\mathcal{C}_{\mathrm{HH}}|$ is the retained set after classical thresholding, $|\mathcal{C}_{\mathrm{blk}}|$ is the retained set after block-prefiltering, and $|\mathcal{C}_L|$ is the final top-L compressed set.

Instance	Shots	$\lvert\mathcal{C}_{\mathrm{HH}}\rvert$	$\lvert\mathcal{C}_{\mathrm{blk}}\rvert$	$\lvert\mathcal{C}_L\rvert$	Gap HH [%]	Gap blk [%]	Gap top-L [%]
eil10	5000	5000	5000	75	0.00	0.00	0.00
dj8	10000	10000	512	75	0.00	0.07	0.07
dj9	10000	10000	729	75	0.00	0.19	0.19
wi4	10000	128	64	64	0.00	0.00	0.00
wi4	100000	234	64	64	0.00	0.00	0.00
wi5	10000	6	6	6	0.52	0.52	0.52
wi5	100000	2449	125	75	0.00	0.00	0.00
wi6	10000	10000	216	75	0.00	0.00	0.00
wi7	10000	10000	343	75	0.00	0.04	0.04

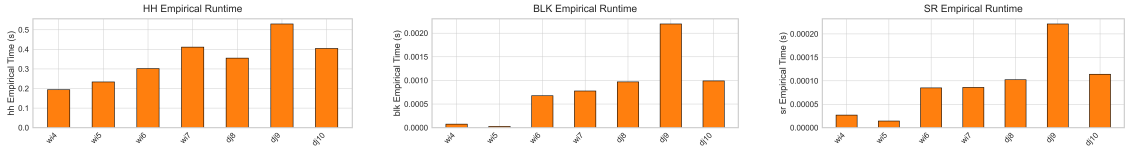
Table 7: Empirical candidate-set footprints and % cost gaps for $k = 4$ with top-L window $L = 25k = 100$.

Instance	Shots	$\lvert\mathcal{C}_{\mathrm{HH}}\rvert$	$\lvert\mathcal{C}_{\mathrm{blk}}\rvert$	$\lvert\mathcal{C}_L\rvert$	Gap HH [%]	Gap blk [%]	Gap top-L [%]
eil10	5000	5000	5000	100	0.00	0.00	0.00
dj8	10000	10000	4096	100	0.00	0.00	0.07
dj9	10000	10000	6561	100	0.00	0.00	0.19
wi4	10000	1398	256	100	0.00	0.00	0.00
wi4	100000	253	253	100	0.00	0.00	0.00
wi5	10000	9994	625	100	0.00	0.00	0.00
wi5	100000	3113	625	100	0.00	0.00	0.00
wi6	10000	10000	1296	100	0.00	0.00	0.00
wi7	10000	10000	2401	100	0.00	0.00	0.04

Tables 6 and 7 show that the block-prefiltering step can reduce the retained set substantially relative to the purely classical thresholding stage, especially on the larger benchmark instances. The optional top-L compression produces a further reduction in footprint, often with negligible change in the best recovered cost. This should be interpreted as an empirical acceleration rather than a formal guarantee: its effectiveness depends on how sharply peaked the filtered histogram is.

To probe robustness more directly, we ran multiple trials on the `wi4`–`wi7` instances with varying shot budgets. In each trial we recorded the retained set size after classical thresholding, after block-prefiltering, and after optional top-L compression, together with the cost gap relative to the best solution obtained from the full retained set.

The main empirical lesson is that the block-prefiltering stage is the most reliable source of compression. It consistently reduces the candidate-set size while preserving the best recovered cost in all reported runs. The additional top-L compression is much more aggressive and usually remains harmless, but it is best viewed as a practical accelerator rather than part of the formal guarantee. The runtime gains are plotted in Fig 5.



(a) Empirical runtime of the classical heavy-hitter filter.

(b) Empirical runtime after block-prefiltering.

(c) Empirical runtime after optional top-L compression.

Figure 5: Empirical CPU runtimes for the three post-processing stages. The classical heavy-hitter stage has the largest footprint, the block-prefiltering stage is substantially cheaper, and the optional top-L compression is cheapest of all.

The empirical results illustrate that although the generic classical heavy-hitter filter already reduces the raw histogram, the quantum-informed threshold sharpens this reduction by one power of n, lowering the worst-case post-processing cost. In addition, the block-prefiltering step often yields a much stronger reduction in practice than is visible from the worst-case bound alone. The optional top-L compression can shrink the candidate set further, but it remains a heuristic empirical accelerator. Crucially, we observe that the empirical quantum-informed refinement inherits the same device-dependent noise sensitivity as the optimal-mass guarantee from which its threshold is derived. We document this

in Appendix A.9.

6 Conclusion and Outlook

We developed a complete hybrid optimization layer around the CE–QAOA sampler, built from structured OFM-kernel elements, and argued that constrained quantum optimization should be viewed as an end-to-end quantum–classical pipeline in which problem structure is exploited throughout both the quantum and classical stages. We subsequently introduced the NP-HQ algorithm as a concrete realization of this paradigm. On the quantum side, the CE–QAOA sampler concentrates probability on low-energy computational-basis states of a diagonal problem Hamiltonian within an encoded constraint-preserving sector. It uses ancilla-free one-hot initialization together with a two-local block-XY mixer. On the classical side, deterministic checking, repair, and scoring map noisy measurement outcomes to feasible candidate solutions and evaluate them in polynomial time.

The performance guarantees we derived rest on the level of bias achievable by the quantum sampler. In the instance-dependent regime where the ideal CE–QAOA distribution assigns inverse-polynomial mass to the optimal feasible set, we showed that the complete pipeline is an exact-hit FPRAS_q. Our total-variation analysis yields a non-asymptotic guarantee and shows that this FPRAS_q guarantee persists within a device-noise window determined by the ideal optimal mass and the instance-dependent effective circuit depth. Concretely, we showed that the recovered optimal mass remains inverse-polynomial, thereby yielding a polynomial shot budget and polynomial classical overhead. Outside the exact-hit regime (whether because the ideal sampler does not assign inverse-polynomial mass to the optimal feasible set or because that mass is no longer guaranteed after noise and repair), our algorithm continues to guarantee feasibility and yields an instance-dependent $(1+\varepsilon)$-approximation scheme whenever the objective inflation induced by repair is controlled.

The complexity analysis also identifies a conditional quantum–classical separation. Even if a classical baseline is granted the same polynomial-time classical projection and scoring routine as NP-HQ, and may sample directly from row-one-hot or feasible spaces, inverse-polynomial overlap with the unknown global optimum uniformly over an NP-hard kernel-admissible promise family would yield a BPP algorithm by standard amplification. Thus, unless $\mathrm{NP} \subseteq \mathrm{BPP}$, such a classical baseline cannot uniformly reproduce the inverse-polynomial optimal-overlap guarantee established here. The distinction, we argue, cannot be attributed merely to constraint handling or repair; it lies in the structured sampling distribution generated by the quantum kernel.

The heavy-hitter analysis sharpens this picture further. Once the quantum kernel places sufficient weight on a small set of important candidates, the classical post-processor need not inspect the full raw output histogram. Thresholding the empirical histogram and checking only the retained candidates can dramatically reduce memory and runtime while preserving the exact-hit or approximation guarantees available in the relevant regime. As documented in Appendix A.9, the empirical quantum-informed refinement inherits this device dependence. Nevertheless, we identify heavy-hitter filtering as a natural interface between constrained quantum sampling and scalable classical post-processing.

Finally, the hardware experiments instantiate the complete pipeline on 127-qubit IBM Eagle-r3 processors using QOptlib travelling-salesman instances with up to 100 logical variables. Across all tested instances, the final outputs match or improve the published reference tours, providing a current-hardware demonstration of the checker–repair–scoring architecture analyzed in this work.

A Mechanism behind the inherited Fejér success bound

A.1 Effective Depth of Constraint–Enhanced QAOA Circuits

The Constraint–Enhanced QAOA (CE–QAOA) introduced in Ref. [23] makes problem–algorithm co-design explicit through the so-called OFM kernel, a designed product of blocks of one-hot or k-hot manifolds, and fixes the mixer family and initial state to match the encoding. It further expands the co-design choice to include symmetries derived from the hard cardinality constraints satisfied by valid solutions. These facts lead to simple gate vs depth accounting.

Remark 32 (Gate vs depth). Usually, the problem Hamiltonian contains $\Theta(n^3)$ two-qubit ZZ interaction terms per CE-QAOA layer (the dense TSP construction is a good example) which can be parallelized to $O(n)$ sequential schedules. The complete XY mixer applies one exchange unitary to each pair of qubits within every one-hot block. Since there are n blocks, each containing n qubits, the number of exchange operations per mixing layer is

$$N_{\mathrm{ex}}(n) = n\binom{n}{2} = \Theta(n^3).$$

Exchanges acting on disjoint qubit pairs can be executed in parallel, so an edge-colouring schedule gives $\Theta(n)$ sequential exchange rounds per mixing layer so the effective sequential depth remains $G(n,p) = O(pn)$. Initial state preparation requires depth of $O(n)$ [23].

A.2 Finite Shot Finite Depth Guarantees

This appendix provides a self-contained account of the analytic mechanism underlying the dimension-free success bound inherited from Ref. [1]. The discussion is given in the notation of the present paper and is organized as follows: Sec. A.3 introduces the dephased reference model that exposes the positive trigonometric filter; Sec. A.4 derives the factorized reference distribution; and Sec. A.5 draws out the finite-shot consequences.

A.3 Dephased reference model

The analytic strategy is to introduce an intermediate *dephased* model that retains the positive structure needed for lower bounding while remaining analytically tractable. Define the cost-basis dephasing channel

$$\mathcal{T}(\rho) \;:=\; \int_0^{2\pi} \frac{d\phi}{2\pi}\, e^{-i\phi H_C}\, \rho\, e^{+i\phi H_C}. \tag{25}$$

This channel projects onto the diagonal in the H_C eigenbasis: it removes all coherences between distinct eigenspaces and leaves only the occupation probabilities $\langle z|\,\rho\,|z\rangle$.

Remark 33 (Role of dephasing). The dephasing channel is *not* a physical noise model applied during the computation. It is an analytic device: by inserting $\mathcal{T}$ after each CE–QAOA layer, we obtain a reference model that is easier to bound from below.

After inserting $\mathcal{T}$ between each pair of layers, the evolution of the diagonal is governed by a classical Markov chain. Write $v^{(r)}(z) := \langle z|\,\rho_r\,|z\rangle$ for the diagonal distribution after r dephased layers. The mixer step acts as

$$v^{(r)}(z) \;=\; \sum_{y\in X} M_{\beta_r}(z|y)\, v^{(r-1)}(y), \qquad M_\beta(z|y) \;:=\; \big|\langle z|\,U_M(\beta)\,|y\rangle\big|^2. \tag{26}$$

Each transition matrix M_β is *unistochastic* (derived from the squared moduli of a unitary matrix) and hence doubly stochastic.

Lemma 34 (Properties of the mixer transition matrix)**.** *For every $\beta \in \mathbb{R}$:*

(a) *M_β is doubly stochastic on $\mathcal{H}_{\mathrm{OH}}$: $\sum_z M_\beta(z|y) = 1$ and $\sum_y M_\beta(z|y) = 1$.*

(b) *M_β preserves the uniform distribution: if $v(z) = 1/D$ for all z, then $[M_\beta\, v](z) = 1/D$.*

(c) *At $\beta = 0$, $M_0 = I$ (the identity permutation on distributions).*

(d) *Because $U_M(\beta)$ factorizes over blocks, M_β factorizes as a tensor product of m local transition matrices, each of size $n_{\mathrm{loc}} \times n_{\mathrm{loc}}$.*

Proof. (a) follows from unitarity of $U_M(\beta)$. (b) is immediate from (a). (c) is immediate from $U_M(0) = I$. (d) follows from the tensor-product structure of $U_M(\beta)$. □

Starting from the initial diagonal $v^{(0)}(z) = |\langle z|s_0\rangle|^2 = D^{-1}$ (uniform on $\mathcal{H}_{\mathrm{OH}}$), iterating the Markov chain gives the *mixer envelope*:

$$W_p(z;\boldsymbol{\beta}) \;:=\; [M_{\beta_p} \cdots M_{\beta_1}\, v^{(0)}](z). \tag{27}$$

Because each M_{β_r} is doubly stochastic and $v^{(0)}$ is uniform, W_p is a probability distribution on $\mathcal{H}_{\mathrm{OH}}$. Its role in the success bound is to quantify how the mixer schedule redistributes mass across the encoded manifold, independently of the cost structure.

Lemma 35 (Mixer envelope is a probability distribution)**.** *For every angle schedule $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, $W_p(\cdot;\boldsymbol{\beta}) \geq 0$ and $\sum_{z\in\mathcal{H}_{\mathrm{OH}}} W_p(z;\boldsymbol{\beta}) = 1$.*

Proof. Nonnegativity follows from $M_\beta(z|y) \geq 0$. Normalization follows by induction: $v^{(0)}$ is normalized, and each doubly stochastic step preserves the total mass. □

The *Fejér kernel* of order p is the nonnegative trigonometric polynomial:

$$F_p(\vartheta) \;:=\; \frac{1}{p+1}\left|\sum_{r=0}^{p} e^{ir\vartheta}\right|^2 \;=\; \frac{1}{p+1}\left(\frac{\sin\!\left(\frac{(p+1)\vartheta}{2}\right)}{\sin\!\left(\frac{\vartheta}{2}\right)}\right)^2. \tag{28}$$

For all $|\vartheta| \geq \delta > 0$ [1],

$$F_p(\vartheta) \;\leq\; \frac{1}{(p+1)\sin^2(\delta/2)}. \tag{29}$$

The key feature is the contrast between the on-peak value $F_p(0) = p+1$ and the off-peak bound $M_p(\delta) = O((p+1)^{-1})$: the ratio is $(p+1)^2\sin^2(\delta/2)$, which grows quadratically with filter depth. This is the mechanism through which deeper CE–QAOA alternations can enhance discrimination between optimal and suboptimal cost phases.

A.4 Factorized reference distribution

The dephased reference model produces a distribution on $\mathcal{H}_{\mathrm{OH}}$ that factorizes into two components: the mixer envelope W_p and the Fejér filter F_p[1]. To see this, consider the effect of p dephased layers on the initial state. After the r-th cost layer, the off-diagonal coherences acquire cost-dependent phases. Subsequent mixing and dephasing transfer this phase dependence into the diagonal reference weights. After dephasing, these phases manifest as a multiplicative weight on each basis string. The dephased dynamics determines the mixer envelope $W_p(z;\boldsymbol{\beta})$. Under the harmonic phase schedule $\gamma_r = r\gamma$, the Dirichlet-filtered reference construction of Ref. [1] weights this envelope by the Fejér factor $F_p(\theta(z)-\theta^\star)$. The resulting diagonal weight is therefore proportional to $W_p(z;\boldsymbol{\beta})F_p(\theta(z)-\theta^\star)$.

The normalized *factorized reference distribution* is:

$$\underset{p}{\overset{\mathrm{ref}}{\Pr}}[z] \;:=\; \frac{W_p(z;\boldsymbol{\beta})\; F_p(\theta(z)-\theta^\star)}{\displaystyle\sum_{y\in X} W_p(y;\boldsymbol{\beta})\; F_p(\theta(y)-\theta^\star)}. \tag{30}$$

This factorization isolates the two competing effects:

- The envelope W_p controls how the mixer spreads mass across the encoded manifold. A string that receives little mixer mass will contribute little to the success probability regardless of its cost phase.
- The Fejér factor $F_p(\theta(z) - \theta^\star)$ amplifies strings near the optimal phase and suppresses those far from it. The degree of amplification grows with the filter order p.

Define the *optimal-set envelope weight*: $C_\beta := \sum_{x\in\Omega^\star} W_p(x;\boldsymbol{\beta})$.

Theorem 36 (Reference Fejér factorization and success bound)**.** *Assume that the TSP specialization admits the factorized reference law (8), and assume a positive wrapped phase separation $\delta > 0$ in the sense of (4). Then the probability of sampling an optimal basis string satisfies*

$$q_0 := \Pr_p^{\mathrm{ref}}[\Omega^\star] \;\geq\; \frac{(p+1)\,C_\beta}{(p+1)\,C_\beta \;+\; M_p(\delta)\,(1-C_\beta)}. \tag{31}$$

In particular, the lower bound is dimension-free*: it depends only on the filter order p, the phase gap δ, and the optimal-set envelope weight C_β.*

Proof. Summing (8) over $x \in \Omega^\star$ and using $F_p(0) = p+1$ gives

$$q_0 \;=\; \frac{(p+1)\,C_\beta}{\displaystyle\sum_{y\in X} W_p(y;\boldsymbol{\beta})\,F_p(\theta(y)-\theta^\star)}.$$

Split the denominator into optimal and suboptimal contributions:

$$\mathrm{denom} \;=\; (p+1)\,C_\beta \;+\; \sum_{y\notin\Omega^\star} W_p(y;\boldsymbol{\beta})\,F_p(\theta(y)-\theta^\star).$$

For every $y \notin \Omega^\star$, the phase-gap assumption implies $|\theta(y)-\theta^\star| \geq \delta$ on $\mathbb{T}$, so $F_p(\theta(y)-\theta^\star) \leq M_p(\delta)$ by Eq. 29. Hence the suboptimal sum is at most

$$M_p(\delta) \sum_{y\notin\Omega^\star} W_p(y;\boldsymbol{\beta}) \;=\; M_p(\delta)\,(1-C_\beta),$$

which yields (31). See Ref. [1] for details □

A.5 Consequences for the hybrid pipeline

We now translate the reference-model bound into the guarantees used in the main text.

Corollary 37 (Instance-dependent inverse-polynomial mass from the joint envelope–filter law)**.** *Consider a problem family admitting the factorized reference law (8). Suppose that, for each size-n instance and the chosen mixer schedule, the corresponding envelope and phase parameters satisfy*

$$C_\beta \geq c_0 n^{-a}, \qquad \sin^2(\delta/2) \geq c_1,$$

for constants $c_0, c_1 > 0$ and $a \geq 0$ independent of n. Then, for any fixed depth p, the joint envelope–filter law gives

$$q_0 \geq \frac{(p+1)c_0 n^{-a}}{(p+1)c_0 n^{-a} + \big[(p+1)c_1\big]^{-1}} = \Omega(n^{-a}).$$

Setting $k := a$, this supplies the inverse-polynomial premise

$$q_0 = \Omega(n^{-k})$$

used in Theorem 5 and the FPRAS_q *analysis.*

Proof. By Theorem 36,

$$q_0 \geq \frac{(p+1)C_\beta}{(p+1)C_\beta + M_p(\delta)(1-C_\beta)}.$$

The right-hand side is increasing in C_β and decreasing in $M_p(\delta)$. Moreover,

$$M_p(\delta) \leq \frac{1}{(p+1)\sin^2(\delta/2)} \leq \frac{1}{(p+1)c_1}.$$

Using $C_\beta \geq c_0 n^{-a}$ and $1 - C_\beta \leq 1$ therefore gives

$$q_0 \geq \frac{(p+1)c_0 n^{-a}}{(p+1)c_0 n^{-a} + \left[(p+1)c_1\right]^{-1}}.$$

For fixed p, the numerator is $\Theta(n^{-a})$, while the denominator is $\Theta(1)$. Hence $q_0 = \Omega(n^{-a})$. □

Corollary 38 (Sufficient depth for target success probability[1])**.** *To achieve $q_0 \geq 1 - \varepsilon$ in the reference model, it suffices to choose the filter order p such that*

$$(p+1)^2 \;\geq\; \frac{1-\varepsilon}{\varepsilon} \cdot \frac{1-C_\beta}{C_\beta} \cdot \csc^2(\delta/2). \tag{32}$$

In particular, if $C_\beta = \Omega(n^{-a})$ and $\sin^2(\delta/2) = \Omega(1)$, then $p = O(n^{a/2})$ suffices for any fixed $\varepsilon > 0$.

A.6 Lipschitz Continuity of Noisy Output Distributions

Under the effective-layer noise model of Sec. 2.5.2, the measured output distribution obeys the following Lipschitz bound. For probability distributions P and Q on the same finite sample space Ω, recall that

$$\|P - Q\|_{\mathrm{TV}} = \frac{1}{2}\sum_{x\in\Omega} |P(x) - Q(x)| = \max_{A\subseteq\Omega} |P(A) - Q(A)|.$$

Lemma 39 (Lipschitz continuity under local noise)**.** *Suppose that the noisy implementation $\mathcal{E}_j$ of each of the $G(n)$ sequential noisy locations satisfies*

$$\|\mathcal{E}_j - \mathcal{U}_j\|_\diamond \leq C\epsilon$$

for some constant $C > 0$, where $\mathcal{U}_j$ is the corresponding ideal channel. Then

$$\|P_\epsilon - P_0\|_{\mathrm{TV}} \leq \frac{C}{2} G(n)\epsilon =: LG(n)\epsilon.$$

In particular, when $C = 1$,

$$\|P_\epsilon - P_0\|_{\mathrm{TV}} \leq \frac{1}{2} G(n)\epsilon.$$

Proof. A telescoping expansion of the noisy and ideal channel compositions, together with the triangle inequality and contractivity of the diamond norm under composition with quantum channels, gives

$$\|\mathcal{E}_G \circ \cdots \circ \mathcal{E}_1 - \mathcal{U}_G \circ \cdots \circ \mathcal{U}_1\|_\diamond \leq \sum_{j=1}^{G(n)} \|\mathcal{E}_j - \mathcal{U}_j\|_\diamond \leq CG(n)\epsilon.$$

Consequently, $\|\rho_\epsilon - \rho_0\|_1 \leq CG(n)\epsilon$. Let $\mathcal{M}$ denote computational-basis measurement. By contractivity of trace distance under CPTP maps,

$$\|P_\epsilon - P_0\|_{\mathrm{TV}} = \frac{1}{2}\|\mathcal{M}(\rho_\epsilon) - \mathcal{M}(\rho_0)\|_1 \leq \frac{1}{2}\|\rho_\epsilon - \rho_0\|_1 \leq \frac{C}{2} G(n)\epsilon.$$

See Refs. [50, 47] for the standard channel-norm and trace-distance properties used above. □

A.7 Bound on Success Probability in Presence of Noise

Let $\mathcal{X}^* \subseteq \Omega$ denote the set of optimal strings, and define

$$p_0 = P_0(\mathcal{X}^*), \quad p_{\min}(\epsilon) = P_\epsilon(\mathcal{X}^*).$$

By the defining event bound for total variation,

$$|p_{\min}(\epsilon) - p_0| = |P_\epsilon(X^\star) - P_0(X^\star)| \le \|P_\epsilon - P_0\|_{\mathrm{TV}} \le LG(n,p)\epsilon.$$

Hence, $p_{\min}(\epsilon) \ge p_0 - LG(n,p)\epsilon$.

Treating each sample as an independent Bernoulli trial with success probability at least $p_{\min}(\epsilon)$, the probability of not seeing an optimal string in S shots is

$$(1 - p_{\min}(\epsilon))^S \le \exp(-S\, p_{\min}(\epsilon)).$$

To ensure an overall failure probability at most δ, it is enough to choose

$$S \ge \left\lceil \frac{\ln(1/\delta)}{p_0 - LG(n,p)\epsilon} \right\rceil,$$

provided $p_0 - LG(n,p)\epsilon > 0$. A sufficient condition for $p_{\min}(\epsilon) \ge n^{-k-1}$ is

$$\epsilon \le \frac{p_0 - n^{-k-1}}{LG(n,p)} =: \epsilon_{\mathrm{th}}.$$

Below this threshold, the solver succeeds exactly with $\mathrm{poly}(n)$ shots.

A.8 Heavy Hitters From Depth-Compressed Noisy Quantum Circuits

Theorem 40 (Combined TV-distance bound on circuit approximation)**.** *Let*

$$U_{\mathrm{PC}}, \qquad U_{\mathrm{APC}},$$

be the ideal unitaries of the original circuit ***PC*** *and its compressed approximation* ***APC****, respectively, acting on n qubits. Define their ideal output distributions by*

$$P_{\mathrm{PC}}(x) \equiv P_0(x) := \left|\langle x|U_{\mathrm{PC}}|0^n\rangle\right|^2, \qquad P_{\mathrm{APC}}(x) := \left|\langle x|U_{\mathrm{APC}}|0^n\rangle\right|^2,$$

for $x \in \{0,1\}^n$.

Suppose the circuit approximation satisfies

$$\left\|U_{\mathrm{PC}} - U_{\mathrm{APC}}\right\|_{\mathrm{tr}} \le \delta,$$

and consequently

$$\left\|P_{\mathrm{APC}} - P_0\right\|_{\mathrm{TV}} \le \delta.$$

Now consider hardware noise of strength ϵ. Let $P_{\epsilon,\delta}$ denote the noisy output distribution obtained by running ***APC****, and let $G_{\mathrm{APC}}(n)$ denote its effective noisy depth. Applying the Lipschitz bound of Appendix A.6 to the approximate circuit gives*

$$\left\|P_{\epsilon,\delta} - P_{\mathrm{APC}}\right\|_{\mathrm{TV}} \le L\, G_{\mathrm{APC}}(n)\, \epsilon, \qquad L > 0, \quad G_{\mathrm{APC}}(n) \ge 1.$$

Then

$$\left\|P_{\epsilon,\delta} - P_0\right\|_{\mathrm{TV}} \le \underbrace{L\, G_{\mathrm{APC}}(n)\, \epsilon}_{\textit{noise error}} + \underbrace{\delta}_{\textit{approximation error}}.$$

Let $x^\star$ be an optimal string and suppose its ideal probability satisfies

$$p_0^\star := P_0(x^\star) \ge \frac{c}{n^k}$$

for fixed constants $c > 0$ and $k \ge 0$. Its probability under the noisy approximate circuit then obeys

$$p^\star_{\epsilon,\delta} := P_{\epsilon,\delta}(x^\star) \ge p_0^\star - L\, G_{\mathrm{APC}}(n)\, \epsilon - \delta.$$

For the heavy-hitter threshold $1/H$, define

$$\mathcal{H}_{1/H} := \left\{ x : P_{\epsilon,\delta}(x) \ge \frac{1}{H} \right\}.$$

If

$$p_0^\star - L\, G_{\mathrm{APC}}(n)\, \epsilon - \delta \geq \frac{1}{H},$$

then

$$x^\star \in \mathcal{H}_{1/H}, \qquad \left|\mathcal{H}_{1/H}\right| \leq H.$$

Sketch. (a) By the triangle inequality for total variation,

$$\begin{aligned} \left\|P_{\epsilon,\delta} - P_0\right\|_{\mathrm{TV}} &\leq \left\|P_{\epsilon,\delta} - P_{\mathrm{APC}}\right\|_{\mathrm{TV}} + \left\|P_{\mathrm{APC}} - P_0\right\|_{\mathrm{TV}} \\ &\leq L\, G_{\mathrm{APC}}(n)\, \epsilon + \delta. \end{aligned}$$

(b) Applying the defining event bound for total variation to the singleton event $\{x^\star\}$ gives

$$\begin{aligned} p^\star_{\epsilon,\delta} &= P_{\epsilon,\delta}(x^\star) \\ &\geq P_0(x^\star) - \left\|P_{\epsilon,\delta} - P_0\right\|_{\mathrm{TV}} \\ &\geq p_0^\star - L\, G_{\mathrm{APC}}(n)\, \epsilon - \delta. \end{aligned}$$

(c) If

$$p_0^\star - L\, G_{\mathrm{APC}}(n)\, \epsilon - \delta \geq \frac{1}{H},$$

then $P_{\epsilon,\delta}(x^\star) \geq 1/H$, and hence $x^\star \in \mathcal{H}_{1/H}$.

Finally, normalization gives

$$1 \geq \sum_{x \in \mathcal{H}_{1/H}} P_{\epsilon,\delta}(x) \geq \frac{\left|\mathcal{H}_{1/H}\right|}{H},$$

so

$$\left|\mathcal{H}_{1/H}\right| \leq H.$$

□

Corollary 41 (Threshold-vs-Depth Trade-off)**.** *Let D be the original two-qubit–layer depth and ϵ_L be the per-layer error rate. Let $\alpha \in (0, 1]$ be the depth-reduction factor (so the approximate circuit has depth αD), and δ satisfy $\left\|P_{\mathrm{APC}} - P_{\mathrm{PC}}\right\|_{\mathrm{TV}} \leq \delta$. Then under the same Lipschitz-in-TV noise bound above in Theorem 40 $\left\|P_\epsilon - P_0\right\|_{\mathrm{TV}} \leq D\, \epsilon_L$, the new heavy-hitter threshold*

$$\theta_{\mathrm{new}} \;\leq\; p_0 \;-\; (\alpha D\, \epsilon_L) \;-\; \delta$$

guarantees the optimum remains in $\{x : P_{\epsilon,\mathrm{apx}}(x) \geq \theta_{\mathrm{new}}\}$. Moreover, comparing to the old threshold $\theta_{\mathrm{old}} = p_0 - D\, \epsilon_L$, the shift $\Delta\theta = \theta_{\mathrm{new}} - \theta_{\mathrm{old}} = (1 - \alpha)\, D\, \epsilon_L - \delta$ implies you only need to lower your cutoff *iff $\delta \;>\; (1 - \alpha)\, D\, \epsilon_L$.*

Algorithm 4 Depth-Compressed Heavy-Hitter Threshold Adjustment

Require: Original two-qubit-layer depth D, Per-layer error rate ϵ_L, Depth-reduction factor $\alpha \in [0,1]$, Approximation error $\delta \in [0,1]$ in total variation distance and Ideal estimated success probability p_0

Ensure: New heavy-hitter cutoff θ_{new}, action $\in \{\textsc{Raise}, \textsc{Lower}, \textsc{Unchanged}\}$, Adjustment magnitude $|\Delta\theta|$

$\theta_{\text{old}} \leftarrow p_0 - D\epsilon_L$ ▷ Original cutoff
$\theta_{\text{new}} \leftarrow p_0 - \alpha D\epsilon_L - \delta$ ▷ Compressed cutoff
$\Delta\theta \leftarrow \theta_{\text{new}} - \theta_{\text{old}}$
if $\Delta\theta > 0$ **then**
    action $\leftarrow$ Raise
else if $\Delta\theta < 0$ **then**
    action $\leftarrow$ Lower
    $\Delta\theta \leftarrow -\Delta\theta$ ▷ Take absolute value
else
    action $\leftarrow$ Unchanged
end if
return $(\theta_{\text{new}}, \text{action}, \Delta\theta)$

We can visualize the effect of Algorithm 4 on $\Delta\theta(\alpha,\delta)$ for fixed D, ϵ_L, p_0 from the derived relation $\Delta\theta(\alpha,\delta) = (1-\alpha)\, D\, \epsilon_L - \delta$ in Figure 6 where it shows a two-dimensional contour map of the maximum permissible heavy-hitter threshold

$$\theta_{\max}(\alpha,\delta) = p_0 \; - \; \alpha\, D\, \epsilon_L \; - \; \delta,$$

plotted over $0 \le \alpha \le 1$, $0 \le \delta \le 0.30$, for the illustrative parameters $p_0 = 0.2$, $D = 100$, $\epsilon_L = 4.55 \times 10^{-3}$.

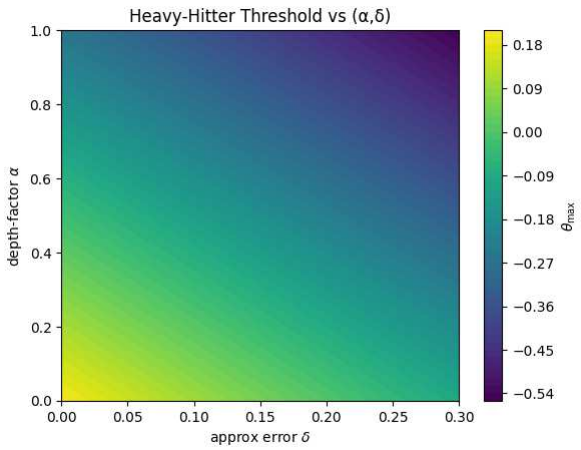


Figure 6: Maximum heavy-hitter threshold $\theta_{\max}(\alpha,\delta)$ as a function of the depth factor α and approximation error δ. At fixed δ, depth compression increases $\theta_{\max}$ linearly; at fixed α, increasing δ decreases it one-to-one. The bottom-left region shows that aggressive depth reduction and negligible approximation error permit a cutoff near 0.18. Moving upward or rightward lowers the threshold through the combined error budget $\alpha D\epsilon_L + \delta$. In the darkest region, this budget exceeds the ideal success probability $p_0 = 0.2$, giving $\theta_{\max} < 0$.

In practice, one chooses a desired heavy-hitter cutoff θ (e.g. based on the mean amplitude) and then reads off from this plot the allowable combinations of (α,δ) that satisfy $\theta \le \theta_{\max}(\alpha,\delta)$. Shot-Budget shifts visibly under the (α,δ) trade-off. From the heavy–hitter probability bound proved above, $p_{\min}(\alpha,\delta) \ge p_0 - \alpha D\epsilon_L - \delta$, feeds directly into the Chernoff–style shot formula (Theorem 4).

A.9 Noise sensitivity of the quantum-informed heavy-hitter refinement

The quantum-informed heavy-hitter refinement inherits the device-dependent character of the optimal-mass guarantee on which its threshold is based. In particular, the refined

candidate set is guaranteed to retain the optimum only while the noisy output distribution remains inside the total-variation window assumed in Theorem 30. Tables 8 and 9 report the full empirical data for $k = 3$ and $k = 4$, respectively. Here $|H_{\mathrm{HH}}|$ is the candidate count after ordinary heavy-hitter filtering, $|H_{\mathrm{blk}}|$ is the count after block filtering, and $|H_{\mathrm{SR}}|$ is the count after the final structure-refined quantum filtering step. Each reported gap is the percentage increase in the best recovered cost relative to exhaustive analysis of the corresponding raw sample histogram. Repeated rows for an instance correspond to independent hardware runs or different shot budgets.

The ordinary heavy-hitter stage reproduces the exhaustive best cost in almost every finite case. Additional structure-aware compression usually preserves that cost while substantially reducing the candidate set, but a minority of runs exhibit a nonzero gap. These departures are consistent with the device-dependent scope of the quantum-informed threshold. Once noise depletes or redistributes the mass supporting the best candidate beyond the admissible window, aggressive filtering need not retain it. This effect is more visible for $k = 3$, where several runs produce nonzero gaps and some noisy histograms yield an empty retained set. The corresponding entries are reported as ∞, because no finite recovered cost exists from which to form a percentage gap. Increasing to $k = 4$ enlarges the retained sets and eliminates empty-set failures in the reported data, although isolated departures remain after the more aggressive refinement.

Table 8: Full heavy-hitter data for $k = 3$. The three gap columns give the percentage increase in the best recovered cost relative to exhaustive analysis of the raw sample histogram. An entry of ∞ indicates that the corresponding retained set was empty.

Instance	Shots	$\lvert H_{\mathrm{HH}}\rvert$	$\lvert H_{\mathrm{blk}}\rvert$	$\lvert H_{\mathrm{SR}}\rvert$	Gap HH [%]	Gap blk [%]	Gap SR [%]
dj10	50000	50000	1000	1000	0.00	6.25	6.25
dj10	5000	5000	1000	1000	0.00	7.99	7.99
dj10	4096	4096	1000	1000	0.00	4.89	4.89
dj8	32000	32000	512	512	0.00	0.07	0.07
dj8	32000	32000	512	512	0.00	0.00	0.00
dj8	32000	32000	512	512	0.00	0.00	0.00
dj8	4096	4096	512	512	0.00	0.00	0.00
dj9	40500	40500	729	729	0.00	0.14	0.14
dj9	40500	40500	729	729	0.00	0.61	0.61
dj9	4050	4050	729	729	0.00	3.80	3.80
dj9	4096	4096	729	729	0.00	4.97	4.97
eil10	50000	50000	1000	1000	0.00	17.75	17.75
eil10	5000	5000	1000	1000	0.00	0.00	0.00
eil10	4096	4096	1000	1000	0.00	4.23	4.23
eil11	60500	60500	1331	1331	0.00	16.17	16.17
eil11	6050	6050	1331	1331	0.00	3.47	3.47
eil11	4096	4096	1331	1331	0.00	0.00	0.00
wi4	8000	1	1	1	0.00	0.00	0.00
wi4	8000	7	7	7	6.70	6.70	6.70
wi4	800	793	64	64	0.00	0.00	0.00
wi4	10000	230	64	64	0.00	0.00	0.00
wi4	100000	234	64	64	0.00	0.00	0.00
wi4	100000	216	64	64	0.00	0.00	0.00
wi4	4096	449	64	64	0.00	0.00	0.00
wi5	12500	0	0	0	∞	∞	∞
wi5	12500	0	0	0	∞	∞	∞
wi5	1250	1250	125	125	0.00	0.00	0.00
wi5	4096	4095	125	125	0.00	0.00	0.00
wi6	18000	0	0	0	∞	∞	∞
wi6	18000	0	0	0	∞	∞	∞
wi6	1800	1800	216	216	0.00	0.00	0.00
wi6	4096	4096	216	216	0.00	0.00	0.00
wi7	24500	0	0	0	∞	∞	∞
wi7	24500	0	0	0	∞	∞	∞
wi7	2450	2450	343	343	0.00	0.00	0.00
wi7	4096	4096	343	343	0.00	0.00	0.00

Table 9: Full heavy-hitter data for $k = 4$. The three gap columns give the percentage increase in the best recovered cost relative to exhaustive analysis of the raw sample histogram.

Instance	Shots	$\lvert H_{\mathrm{HH}}\rvert$	$\lvert H_{\mathrm{blk}}\rvert$	$\lvert H_{\mathrm{SR}}\rvert$	Gap HH [%]	Gap blk [%]	Gap SR [%]
dj10	50000	50000	10000	2000	0.00	0.98	6.25
dj10	5000	5000	5000	2000	0.00	0.00	0.00
dj10	4096	4096	4096	2000	0.00	0.00	0.00
dj8	32000	32000	4096	2000	0.00	0.00	0.00
dj8	32000	32000	4096	2000	0.00	0.00	0.00
dj8	32000	32000	3200	2000	0.00	0.00	0.00
dj8	4096	4096	4096	2000	0.00	0.00	0.00
dj9	40500	40500	6561	2000	0.00	0.14	0.14
dj9	40500	40500	6561	2000	0.00	0.00	0.14
dj9	4050	4050	4050	2000	0.00	0.00	0.00
dj9	4096	4096	4096	2000	0.00	0.00	1.22
eil10	50000	50000	10000	2000	0.00	4.14	17.16
eil10	5000	5000	5000	2000	0.00	0.00	0.00
eil10	4096	4096	4096	2000	0.00	0.00	0.00
eil11	60500	60500	14641	2000	0.00	10.78	10.78
eil11	6050	6050	6050	2000	0.00	0.00	3.47
eil11	4096	4096	4096	2000	0.00	0.00	0.00
wi4	8000	7543	256	256	0.00	0.00	0.00
wi4	8000	7307	256	256	0.00	0.00	0.00
wi4	800	793	256	256	0.00	0.00	0.00
wi4	10000	242	242	242	0.00	0.00	0.00
wi4	100000	253	253	253	0.00	0.00	0.00
wi4	100000	246	246	246	0.00	0.00	0.00
wi4	4096	3495	256	256	0.00	0.00	0.00
wi5	12500	12500	625	625	0.00	0.00	0.00
wi5	12500	12499	625	625	0.00	0.00	0.00
wi5	1250	1250	625	625	0.00	0.00	0.00
wi5	4096	4095	625	625	0.00	0.00	0.00
wi6	18000	18000	1296	1296	0.00	0.00	0.00
wi6	18000	18000	1296	1296	0.00	0.00	0.00
wi6	1800	1800	1296	1296	0.00	0.00	0.00
wi6	4096	4096	1296	1296	0.00	0.00	0.00
wi7	24500	24500	2401	2000	0.00	0.00	0.00
wi7	24500	24500	2401	2000	0.00	0.00	0.00
wi7	2450	2450	2401	2000	0.00	0.00	0.00
wi7	4096	4096	2401	2000	0.00	0.00	0.00

References

[1] Chinonso Onah and Kristel Michielsen. *Finite-Depth, Finite-Shot Guarantees for Constrained Quantum Optimization via Fejér Filtering*. 2026. arXiv: 2603.01809 [quant-ph]. URL: https://arxiv.org/abs/2603.01809.

[2] Sitan Chen et al. "The complexity of NISQ". In: *Nature Communications* 14 (2023), p. 6001. DOI: 10.1038/s41467-023-41217-6.

[3] Eneko Osaba and Erika Villar-Rodríguez. "QOPTLib: a Quantum-Computing-Oriented Benchmark for Combinatorial Optimisation Problems". In: *Proc. Quantum Tech. 2024*. 2024.

[4] Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice Hall, 1982.

[5] Andrew Lucas. "Ising formulations of many NP problems". In: *Frontiers in Physics* 2 (2014), p. 5. DOI: 10.3389/fphy.2014.00005.

[6] Eugene L. Lawler et al. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley, 1985.

[7] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

[8] Tjalling C. Koopmans and Martin Beckmann. "Assignment Problems and the Location of Economic Activities". In: *Econometrica* 25.1 (1957), pp. 53–76. DOI: 10.2307/1907742.

[9] E. M. Loiola et al. "A Survey for the Quadratic Assignment Problem". In: *European Journal of Operational Research* 176.2 (2007), pp. 657–690. DOI: 10.1016/j.ejor.2005.09.032.

[10] Paolo Toth and Daniele Vigo. *Vehicle Routing: Problems, Methods, and Applications*. 2nd ed. MOS-SIAM Series on Optimization. SIAM, 2014. DOI: 10.1137/1.9781611973594.

[11] Sartaj Sahni and Teofilo Gonzalez. "P-Complete Approximation Problems". In: *Journal of the ACM* 23.3 (1976), pp. 555–565. DOI: 10.1145/321958.321975.

[12] Silvano Martello and Paolo Toth. *Knapsack Problems: Algorithms and Computer Implementations*. Wiley, 1990.

[13] Richard M. Karp. "Reducibility among Combinatorial Problems". In: *Complexity of Computer Computations*. Plenum, 1972, pp. 85–103.

[14] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York: W. H. Freeman and Company, 1979.

[15] Chinonso Onah et al. "QUEST: QUantum-Enhanced Shared Transportation". In: *2025 IEEE International Conference on Quantum Computing and Engineering (QCE)*. Vol. 01. 2025, pp. 2149–2160. DOI: 10.1109/QCE65121.2025.00235.

[16] Juan F. R. Hernandez et al. *Quantum optimization beyond QUBO for industrial logistics and scheduling*. 2026. arXiv: 2605.30252 [quant-ph]. URL: https://arxiv.org/abs/2605.30252.

[17] Chinonso Onah et al. *Quantum and classical approaches to the optimization of highway platooning: the two-vehicle matching problem*. 2026. arXiv: 2603.18919 [quant-ph]. URL: https://arxiv.org/abs/2603.18919.

[18] Michael R. Garey, David S. Johnson, and Ravi Sethi. “The Complexity of Flowshop and Jobshop Scheduling”. In: *Mathematics of Operations Research* 1.2 (1976), pp. 117–129. DOI: 10.1287/moor.1.2.117.

[19] Jan Karel Lenstra, Alexander H. G. Rinnooy Kan, and Peter Brucker. “Complexity of Machine Scheduling Problems”. In: *Annals of Discrete Mathematics* 1 (1977), pp. 343–362. DOI: 10.1016/S0167-5060(08)70743-X.

[20] Harold W. Kuhn. “The Hungarian Method for the Assignment Problem”. In: *Naval Research Logistics Quarterly* 2.1-2 (1955), pp. 83–97. DOI: 10.1002/nav.3800020109.

[21] James Munkres. “Algorithms for the Assignment and Transportation Problems”. In: *Journal of the Society for Industrial and Applied Mathematics* 5.1 (1957), pp. 32–38. DOI: 10.1137/0105003.

[22] Jack Edmonds. “Paths, Trees, and Flowers”. In: *Canadian Journal of Mathematics* 17 (1965), pp. 449–467. DOI: 10.4153/CJM-1965-045-4.

[23] Chinonso Onah, Roman Firt, and Kristel Michielsen. *Empirical Quantum Advantage in Constrained Optimization from Encoded Unitary Designs*. 2026. arXiv: 2511.14296 [cs.ET]. URL: https://arxiv.org/abs/2511.14296.

[24] Chinonso Onah, Stuart Hadfield, and Kristel Michielsen. *Separating Geometry From Interference in Constrained Quantum Optimization*. 2026. arXiv: 2607.13630 [quant-ph]. URL: https://arxiv.org/abs/2607.13630.

[25] Chinonso Onah and Kristel Michielsen. *Fundamental Limitations of QAOA on Constrained Problems and a Route to Exponential Enhancement*. 2025. arXiv: 2511.17259 [quant-ph]. URL: https://arxiv.org/abs/2511.17259.

[26] Chinonso Onah and Kristel Michielsen. *Optimal, Qubit-Efficient Quantum Vehicle Routing via Colored-Permutations*. 2026. arXiv: 2604.04570 [quant-ph]. URL: https://arxiv.org/abs/2604.04570.

[27] Xavier Bonet-Monroig et al. “Low-cost error mitigation by symmetry verification”. In: *Physical Review A* 98.6 (Dec. 2018), p. 062339. DOI: 10.1103/PhysRevA.98.062339. URL: https://link.aps.org/doi/10.1103/PhysRevA.98.062339.

[28] Ramiro Sagastizabal et al. “Experimental error mitigation via symmetry verification in a variational quantum eigensolver”. In: *Physical Review A* 100.1 (July 2019), p. 010302. DOI: 10.1103/PhysRevA.100.010302. URL: https://link.aps.org/doi/10.1103/PhysRevA.100.010302.

[29] Kristan Temme, Sergey Bravyi, and Jay M. Gambetta. “Error mitigation for short-depth quantum circuits”. In: *Physical Review Letters* 119.18 (Nov. 2017), p. 180509. DOI: 10.1103/PhysRevLett.119.180509. URL: https://link.aps.org/doi/10.1103/PhysRevLett.119.180509.

[30] Suguru Endo, Simon C. Benjamin, and Ying Li. “Practical Quantum Error Mitigation for Near-Future Applications”. In: *Physical Review X* 8.3 (Aug. 2018), p. 031027. DOI: 10.1103/PhysRevX.8.031027. URL: https://link.aps.org/doi/10.1103/PhysRevX.8.031027.

[31] Sergey Bravyi et al. “Mitigating measurement errors in multiqubit experiments”. In: *Physical Review A* 103.4 (Apr. 2021), p. 042605. DOI: 10.1103/PhysRevA.103.042605. URL: https://link.aps.org/doi/10.1103/PhysRevA.103.042605.

[32] Piotr Czarnik et al. “Error mitigation with Clifford quantum-circuit data”. In: *Quantum* 5 (Nov. 2021), p. 592. DOI: 10.22331/q-2021-11-26-592. URL: https://doi.org/10.22331/q-2021-11-26-592.

[33] Stuart Hadfield et al. “From the Quantum Approximate Optimization Algorithm to a Quantum Alternating Operator Ansatz”. In: *Algorithms* 12.2 (2019), p. 34. DOI: 10.3390/a12020034.

[34] Franz G. Fuchs and Ruben Pariente Bassa. “Constraint Preserving Mixers for the Quantum Approximate Optimization Algorithm”. In: *Algorithms* 15.6 (2022), p. 202. DOI: 10.3390/a15060202.

[35] Nicolas PD Sawaya, Albert T Schmitz, and Stuart Hadfield. “Encoding trade-offs and design toolkits in quantum algorithms for discrete optimization: coloring, routing, scheduling, and other problems”. In: *Quantum* 7 (2023), p. 1111.

[36] B. Tsvelikhovskiy, I. Safro, and Y. Alexeev. “Symmetries and Dimension Reduction in Quantum Approximate Optimization Algorithm”. Version 2. In: *arXiv preprint arXiv:2309.13787* (2023). arXiv: 2309.13787 [quant-ph]. URL: https://arxiv.org/abs/2309.13787.

[37] Abhishek Awasthi et al. *Constraint Preserving XY-Mixers under Trotterized Adiabatic Evolution*. 2026. arXiv: 2605.02465 [quant-ph]. URL: https://arxiv.org/abs/2605.02465.

[38] Pradeep Niroula et al. “Constrained quantum optimization for extractive summarization on a trapped-ion quantum computer”. In: *Scientific Reports* 12 (2022), p. 17171. DOI: 10.1038/s41598-022-20853-w. URL: https://doi.org/10.1038/s41598-022-20853-w.

[39] Filip B. Maciejewski et al. “Improving Quantum Approximate Optimization by Noise-Directed Adaptive Remapping”. In: *Quantum* 9 (Nov. 2025), p. 1906. ISSN: 2521-327X. DOI: 10.22331/q-2025-11-06-1906. URL: http://dx.doi.org/10.22331/q-2025-11-06-1906.

[40] Stuart Hadfield, Filip B. Maciejewski, and Davide Venturelli. *Noise-Directed Adaptive Remapping for Integer Optimization: from qubits to (encoded) qudits*. 2026. arXiv: 2606.28234 [quant-ph]. URL: https://arxiv.org/abs/2606.28234.

[41] Filip Maciejewski et al. “Quantum Approximate Optimization via Noise-Directed Adaptive Warm-Starting”. In: *arXiv preprint arXiv:2607.09368* (2026).

[42] Libor Caha, Alexander Kliesch, and Robert Koenig. “Twisted Hybrid Algorithms for Combinatorial Optimization”. In: *Quantum Science and Technology* 7.4 (2022), p. 045013. DOI: 10.1088/2058-9565/ac7f4f. URL: https://doi.org/10.1088/2058-9565/ac7f4f.

[43] Tatsuhiko Shirai and Nozomu Togawa. “Postprocessing Variationally Scheduled Quantum Algorithm for Constrained Combinatorial Optimization Problems”. In: *IEEE Transactions on Quantum Engineering* 5 (2024). Art. no. 3101114, pp. 1–14. DOI: 10.1109/TQE.2024.3376721. URL: https://doi.org/10.1109/TQE.2024.3376721.

[44] Jayadev Misra and David Gries. “Finding Repeated Elements”. In: *Science of Computer Programming* 2.2 (1982), pp. 143–152. DOI: 10.1016/0167-6423(82)90012-0. URL: https://doi.org/10.1016/0167-6423(82)90012-0.

[45] Graham Cormode and S. Muthukrishnan. “An Improved Data Stream Summary: The Count-Min Sketch and Its Applications”. In: *Journal of Algorithms* 55.1 (2005), pp. 58–75. DOI: 10.1016/j.jalgor.2003.12.001. URL: https://doi.org/10.1016/j.jalgor.2003.12.001.

[46] Sanjeev Arora. “Polynomial-time approximation schemes for Euclidean Traveling Salesman and other geometric problems”. In: *Journal of the ACM* 45.5 (1998), pp. 753–782.

[47] Mark M. Wilde. *Quantum Information Theory*. 2nd ed. Cambridge University Press, 2017.

[48] Richard Durstenfeld. “Algorithm 235: Random Permutation”. In: *Communications of the ACM* 7.7 (1964). Linear-time in-place implementation of the Fisher–Yates shuffle, p. 420. DOI: 10.1145/364520.364540.

[49] M. Ayrton et al. *Qiskit: An Open-source Framework for Quantum Computing*. Version 0.47. 2023. URL: https://qiskit.org.

[50] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.