# CERTIFIED STOPPING FOR FIXED-CODEBOOK QUANTIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

When pairwise refinement stops in fixed-codebook quantization, how much joint improvement remains possible? We develop a certified stopping framework that separates conditional search, certificate optimization, and relaxation strength, connecting each source of uncertainty to the computation that can reduce it. A budgeted pair oracle maintains a feasible incumbent and a valid conditional lower bound after any number of finite branch scores. A globally valid curvature split extends these bounds to the coupled quadratic. Optimizing signed block curvature and the anchor recovers a block-moment relaxation when $H \succ 0$ and every block codebook has full affine span. Its certified optimization interval decomposes exactly into restricted-master and finite-separation terms. On eight exactly enumerable six-coordinate restrictions of captured GPT-2 reconstruction objectives, one pair-fixed point remains above the finite optimum, while four objectives retain strictly positive block-moment relaxation gaps after certificate optimization. Across all eight stored rational objectives, the optimized certificate is tighter than continuous-box and fixed-curvature controls. The hierarchy distinguishes when to score more candidates, improve the conic state, or change the relaxation.

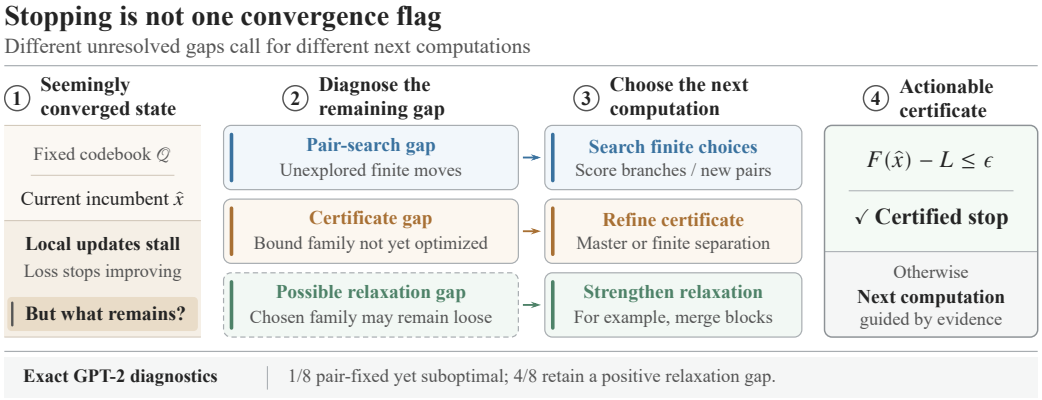



Figure 1: Certified stopping separates finite pair search, optimization within a certificate family, and the relaxation's remaining gap. Each layer identifies which computation can improve the guarantee.

## 1 INTRODUCTION

Pairwise refinement in fixed-codebook quantization can reach a state where every selected pair is conditionally optimal while joint changes still improve the objective. This raises a stopping question: how much improvement remains possible, and which computation can tighten the current guarantee? We study this question through certificates that remain valid under finite search budgets.

Three notions of completion govern this decision (Figure 1). A selected pair can be optimal with all other coordinates fixed; optimization can reach the best value in a chosen certificate family; and a feasible codeword can attain the finite optimum. The associated computational choices are to search additional finite configurations, optimize the certificate, or strengthen its relaxation.



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

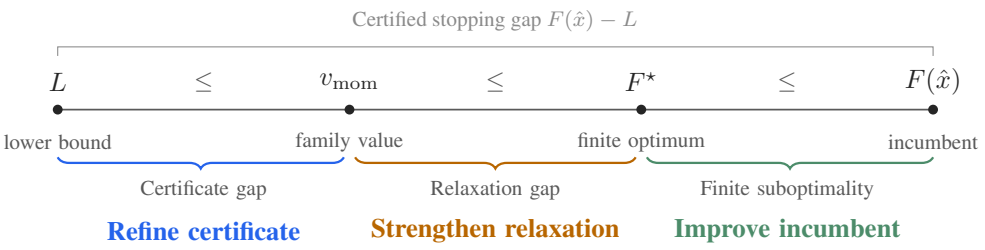



Figure 2: Different gaps call for different computations. The ordering $L \leq v_{\mathrm{mom}} \leq F^\star \leq F(\hat{x})$ separates certificate optimization within the current family, block-moment relaxation error, and finite suboptimality. Their associated targets are to optimize the certificate, strengthen the relaxation (e.g., by merging blocks), or search additional finite choices. These are not execution steps; spacing is schematic, and the component gaps are not generally separately observable. A moment-feasible upper value $U$ gives $v_{\mathrm{mom}} - L \leq U - L$, but need not upper-bound $F^\star$. Closing the family interval alone does not certify finite optimality.

Our framework uses an anytime pair oracle that maintains an incumbent and a valid lower bound after any strict score budget, with two monotone frontiers for positive semidefinite pairs. Modified block curvature lifts these bounds to the coupled quadratic. Optimizing curvature and anchor yields a block-moment characterization whose certified optimization interval separates restricted-master error from finite-separation error.

We use four terms throughout (Figure 2). The *pair-search gap* is conditional on fixed external coordinates. The *certificate-optimization gap* is $v_{\mathrm{mom}} - L$, bounded by a computed interval width $U - L$. The *block-moment relaxation gap* is $F^\star - v_{\mathrm{mom}}$. The *certified suboptimality* $F(\hat{x}) - L$ is the actionable global guarantee on a returned codeword. In global statements, $L$, $U$, and $V$ denote the certificate lower value, moment-primal upper value, and restricted-master dual value.

**Contributions.**

- **Anytime finite-pair certificates.** A budgeted pair oracle returns valid conditional intervals after arbitrary finite score budgets, with lazy branch bounds for positive semidefinite and signed pair curvature.
- **A global stopping hierarchy.** We lift block bounds to the coupled objective, identify the optimized signed-curvature/anchor family with a block-moment relaxation under the stated assumptions, and decompose its certified optimization interval into restricted-master and finite-separation terms.
- **Exact model-derived diagnostics.** Eight enumerable GPT-2 restrictions distinguish pair convergence, certificate-family convergence, and finite optimality on the same objectives, with continuous-box and fixed-curvature controls.

The empirical study follows the same hierarchy. We extract eight fixed six-coordinate restrictions from captured GPT-2 reconstruction objectives, retain the original three-bit grids and complete external residual, and enumerate every finite configuration exactly. These problems make $F^\star$ observable, allowing the local search state, block-moment family, and final certified suboptimality to be compared on one objective. Kernel and downstream model studies address complementary questions about local-oracle cost and the relation between reconstruction and perplexity.

## 2 A COMPLETE RECONSTRUCTION OBJECTIVE

We fix the grids and reconstruction target throughout a refinement run. A certificate then bounds improvement over one finite feasible set and one objective.

Let $W \in \mathbb{R}^{m \times d}$ be a weight matrix in output-by-input layout and $X \in \mathbb{R}^{N \times d}$ be calibration activations. Quantized weights $Z$ obey fixed affine grids

$$Z_{ri} \in \mathcal{Q}_{ri} = \{s_{ri}(u - \zeta_{ri}) : u = 0, \ldots, L-1\}, \qquad s_{ri} > 0. \tag{1}$$

The offset $\zeta_{ri}$ need not be an integer in the analysis. Our GPTQ runs use integer zero points. For $H = X^\top X/N + \rho I \succeq 0$, the usual damped reconstruction objective is $\mathrm{tr}((Z-W)H(Z-W)^\top)$. A more general target adds a linear term:

$$F(Z) = \mathrm{tr}(EHE^\top) + 2\langle E, C\rangle + K, \qquad E = Z - W. \tag{2}$$

Rows are independent for fixed $H, C$ and grids. For one row, write vectors as columns and define the half-gradient $g = H(z-w) + c$. All coordinates contribute to $g$, even when only a pair will change.

For $B = \{i, j\}$ and a feasible displacement $d_B = q_B - z_B$, the exact objective increment is

$$f_B(d_B) = 2g_B^\top d_B + d_B^\top H_{BB} d_B. \tag{3}$$

This expression includes the external residual through $g_B$. Replacing it by a quadratic formed only from $w_B, z_B, H_{BB}$ drops cross terms and changes the optimization problem. Appendix B gives an explicit positive-definite example where a block-only improvement increases the full objective. The distinction matters even when the codebook and update are otherwise identical.

**Why pairs can help.** The first source of residual uncertainty arises from interactions between discrete choices. At a scalar coordinate minimum, every feasible one-coordinate displacement has nonnegative increment. Two such increments may still have a negative sum after adding $2H_{ij}d_i d_j$. Pair search can exploit this coupling. When $H$ is diagonal, the cross term vanishes and scalar exact minimization already solves the row problem. This elementary mechanism motivates pair controls, while the magnitude of their benefit remains an empirical question.

**Budget convention.** The current feasible pair is retained with known increment zero. We count a finite score whenever a first-coordinate branch is evaluated at its exact finite second-coordinate minimizer. A budget $K_b$ limits these scores beyond the retained incumbent. Lower-bound evaluations, center calculations, residual updates, pair selection and data movement are separate work. A run can stop before its first finite score if the lower bound already proves the incumbent optimal. This convention explains noninteger mean counts near one.

## 3 Anytime certificates for budgeted pair search

Suppress $B$ and write $f(x,y) = ax^2 + 2bxy + cy^2 + 2px + 2qy$, where $a = H_{ii}$, $b = H_{ij}$, $c = H_{jj}$ and $(p, q) = g_B$. Coordinates are displacements from the incumbent, so zero is feasible. Assume $a, c > 0$ and $ac - b^2 \geq 0$; the pair may be singular. Let $[l_x, u_x]$ and $[l_y, u_y]$ be the bounding intervals of the two displacement grids.

For a fixed finite $x$, continuous minimization of $y$ gives

$$y_c(x) = \mathrm{clip}\left(-\frac{q + bx}{c}, l_y, u_y\right), \qquad \phi(x) = f(x, y_c(x)). \tag{4}$$

Let $y_f(x)$ be a nearest point on the finite displacement grid to the unconstrained center, with clipping at its endpoints. Then $v(x) = f(x, y_f(x))$ is the exact minimum of that finite branch and $\phi(x) \leq v(x)$. In absolute coordinates, affine rounding uses $s_j\,\mathrm{clip}(\mathrm{round}(q_j^{\mathrm{center}}/s_j + \zeta_j), 0, L-1) - s_j\zeta_j$. Moving the offset outside the rounding operation is valid only under an appropriate integer-offset convention.

**Theorem 1** (Budgeted conditional gap)**.** *Let $S$ be any evaluated set of first-coordinate choices, and $U_{\mathrm{pair}} = \min(0, \min_{x\in S} v(x))$. Suppose $r$ is a lower bound on every unevaluated branch, with $r = +\infty$ when none remain. For the exact finite pair optimum $f^\star$,*

$$\underline{f} = \min(U_{\mathrm{pair}}, r) \leq f^\star \leq U_{\mathrm{pair}}, \qquad U_{\mathrm{pair}} - f^\star \leq \Gamma = U_{\mathrm{pair}} - \underline{f}. \tag{5}$$

*In particular, $r \geq U_{\mathrm{pair}}$ certifies pair optimality; $r \geq U_{\mathrm{pair}} - \tau$ certifies a conditional error of at most $\tau \geq 0$.*

*Proof.* Every evaluated branch has been minimized exactly. The incumbent is feasible, and every unvisited branch is bounded below by $r$. Partitioning the finite domain into those sets yields the lower bound, while feasibility gives the upper bound. Subtraction proves the gap statement. □

One implementation evaluates all $L$ lower bounds and sorts the branches by $\phi$. The next unvisited bound is then $r$. This is a transparent reference implementation and reaches the finite pair optimum with at most $L$ scores, in contrast to enumerating $L^2$ combinations. Its full bound table and sort, however, are a cost even when only one finite branch is needed. The next section removes that preprocessing for this convex relaxation.

**Interpretation.** The theorem turns finite search progress into a quantitative description of remaining conditional uncertainty. The quadratic and all external coordinates are fixed, and the resulting interval records how much additional improvement remains possible within the selected pair.

## 4 Lazy conditional bounds

The relaxation $\phi$ is convex because it partially minimizes a jointly convex quadratic over a convex interval. A minimizer can be found without evaluating the finite grid. First allow $x$ to range over $\mathbb{R}$. Eliminating $x$ leaves

$$\psi(y) = \left(c - \frac{b^2}{a}\right) y^2 + 2\left(q - \frac{bp}{a}\right) y - \frac{p^2}{a}. \tag{6}$$

For $s = c - b^2/a > 0$, set $y_0 = \mathrm{clip}(-(q - bp/a)/s, l_y, u_y)$. For $s = 0$, choose an endpoint minimizing the remaining linear term, with either endpoint allowed when that term is zero. Then $x_0 = -(p + by_0)/a$ minimizes $\phi$ on $\mathbb{R}$, and $\bar{x} = \mathrm{clip}(x_0, l_x, u_x)$ minimizes it on the bounded interval.

Split the finite $x$ grid at the greatest grid point not exceeding $\bar{x}$. One frontier moves left from this point; the other moves right from its successor. Convexity makes each frontier's sequence of lower bounds nondecreasing. Comparing their two heads merges them into a globally ordered sequence. After consuming one head, only that frontier's successor needs a new lower bound. The lesser remaining head is $r$ in Theorem 1.

**Theorem 2** (Lazy bound count)**.** *Under the assumptions of Section 3, the two-frontier procedure enumerates first-coordinate lower bounds in nondecreasing order, up to ties. After $k$ finite scores it has evaluated at most $\min(L, k + 2)$ lower bounds and performed one continuous-center solve. It returns the interval in Equation (5) at any strict score budget and requires only two stored frontier heads, excluding the input grids and objective state.*

*Proof.* The center construction and convexity imply monotonicity away from $\bar{x}$. The minimum unvisited bound is therefore always a current head. Initially at most two heads are evaluated; consuming a head introduces at most one new bound. No grid point is visited twice. The count follows, and the head invariant supplies the lower bound required by Theorem 1. □

The complete procedure is given in Table 24.

## 5 From pair bounds to a global certificate

The pair interval fixes all external coordinates. To certify joint changes, we instead under-estimate the complete quadratic by a sum of finite pair functions. This changes the curvature used by the oracle while preserving the full gradient.

### 5.1 Finite block bounds certify the coupled objective

Write the row objective as $F(x) = x^\top Hx - 2b^\top x + c_0$, with $b = Hw - c$ and $c_0$ collecting the constants in Equation (2). Let $\mathcal{Q} = \prod_i \mathcal{Q}_i$ and $F^\star = \min_{x \in \mathcal{Q}} F(x)$. Partition coordinates into pairs (and one singleton if needed), with $\mathcal{Q}_p = \prod_{i \in p} \mathcal{Q}_i$. Choose symmetric block-diagonal $D = \mathrm{blkdiag}(D_p)$ such that $G = H - D \succeq 0$. At an anchor $z$, set $g = Hz - b$ and $f_p^D(d) = 2g_p^\top d + d^\top D_p d$. The anchor may be continuous; the quantized incumbent remains feasible separately.

**Theorem 3** (From finite block bounds to a global certificate)**.** *If* $l_p \le \min_{q_p \in \mathcal{Q}_p} f_p^D(q_p - z_p)$ *for every block, then*

$$L_D(z) = F(z) + \sum_p l_p \le F^\star,$$
$$F(x) - L_D(z) = \sum_p [f_p^D(x_p - z_p) - l_p] + (x - z)^\top G(x - z). \tag{7}$$

*Consequently every feasible* $\hat{x}$ *satisfies* $0 \le F(\hat{x}) - F^\star \le F(\hat{x}) - L_D(z)$.

*Proof.* Expand the quadratic around $z$ and substitute $H = D + G$. Each block term is nonnegative by its finite lower bound, and the remaining quadratic is nonnegative because $G \succeq 0$. Taking the minimum over feasible $x$ proves the result. □

This identity explains why the original conditional gaps cannot be summed. Their curvature is $H_{pp}$, but $H - \mathrm{blkdiag}(H_{pp})$ need not be positive semidefinite. A valid split is essential. For an exact damped Gram, $D = \rho I$ is immediately valid. It is inexpensive to specify but can lose much of the useful coupling information. Appendix C gives a strict pair-local obstruction even for positive definite matrices arbitrarily close to the identity.

**A captured example: what pair convergence misses.** The first predeclared QKV case in Section 6 makes the distinction concrete. Normalize each of its six affine grids to $\{-1, -5/7, \ldots, 5/7, 1\}$ and retain the three recorded pairs $(1,2), (3,4), (5,6)$. Exact pair sweeps stop at

$$\hat{x} = (1, 3, 1, 1, -1, 1)/7, \qquad F(\hat{x}) = 0.$$

Every alternative within each selected pair has nonnegative increment, as checked over all 64 choices per pair. Nevertheless, the joint codeword $x^\star = (1, 1, 3, 1, -1, 1)/7$ is better. Moving only coordinate 2 costs $0.03004680$; moving only coordinate 3 costs $0.03084345$. Moving both introduces a cross term of $-0.06091832$, leaving the net improvement $F(x^\star) = -2.80714 \times 10^{-5}$. These displayed values are rounded; exact rational enumeration certifies the optimum over all $8^6$ codewords. The two changing coordinates belong to different selected pairs, so this does not contradict exact conditional minimization. It identifies precisely the coupling that a global stopping statement must retain.

### 5.2 The optimized certificate is a block-moment relaxation

For fixed $D$ and exact finite block minima, optimizing the continuous anchor gives

$$J_D^\star = \min_\pi \left\{ F(m) + \sum_p \mathrm{tr}(D_p V_p) \right\},$$

where $\pi_p$ is a distribution over the finite codewords of block $p$, with mean $m_p$ and covariance $V_p$. The distributions describe a relaxation: their moments need not come from one feasible joint codeword. Optimizing over all symmetric admissible $D$ gives the block-moment value $v_{\mathrm{mom}}$ (Appendix C.2). Under $H \succ 0$ and full affine span of every finite block codebook, strong duality identifies $v_{\mathrm{mom}}$ exactly with the optimum of the block-diagonal signed-curvature/anchor family, and $v_{\mathrm{mom}} \le F^\star$. The global certificate above itself only requires $H - D \succeq 0$; the stronger conditions enter this moment-equivalence statement.

The difference is observable even on a small exact example. A four-coordinate binary instance has best PSD-local lower value $-1/4$, unrestricted signed-curvature value $-1/8$, and finite optimum zero. Matching rational primal and dual witnesses establish all three values (Appendix D.5). Allowing signed local curvature removes one restriction, but leaves a block-moment relaxation gap of $1/8$. Once the family is optimized, additional inner scores act only on finite separation and therefore leave this relaxation gap unchanged.

Signed curvature therefore has a concrete role. For an arbitrary symmetric pair, continuous elimination of its second coordinate gives at most three quadratic pieces. Splitting each piece at its discrete

extremum gives at most six monotone streams. Merging their heads supplies the same valid conditional interval, with at most $\min(L, k+6)$ lower-bound evaluations after $k$ finite scores. The positive semidefinite case retains its sharper $k+2$ bound. Appendix C.3 proves the signed construction, including zero and negative diagonal cases. This is an arithmetic count; the signed reference is not a compiled latency result.

### 5.3 The certificate interval decomposes exactly

Curvature and anchor can be optimized jointly. With $\varphi_p(q) = q^\top D_p q + 2(u_p - b_p)^\top q$, the family value is the optimum of

$$\max_{D,u,s,t} \; c_0 - s + \sum_p t_p, \qquad \begin{bmatrix} H - D & u \\ u^\top & s \end{bmatrix} \succeq 0, \qquad t_p \le \varphi_p(q) \quad (q \in \mathcal{Q}_p). \tag{8}$$

Every constraint is convex in the displayed variables. The singular Schur-complement formulation explicitly enforces $u \in \mathrm{range}(H - D)$. Appendix D proves equivalence to the anchor family and gives its moment primal. A feasible primal objective $U$ is an upper bound on $v_{\mathrm{mom}}$. To see what this witness represents, write $M_p = \mathbb{E}_{\pi_p}[q_p q_p^\top]$. Its variables are a mean $m$ and a symmetric matrix $X$ satisfying

$$X - mm^\top \succeq 0, \qquad X_{pp} = M_p, \qquad U = c_0 - 2b^\top m + \mathrm{tr}(HX). \tag{9}$$

Every finite codeword supplies such a witness by taking point masses and $X = xx^\top$. The converse fails: the cross-block entries of $X$ need only satisfy the positive semidefinite constraint, rather than arise from a joint distribution on the finite grid. This enlargement is where the block-moment relaxation gap enters. The mean $m$ may itself fall between codewords, and evaluating $F(m)$ is not the primal objective: the additional term $\mathrm{tr}(H(X - mm^\top))$ accounts for its retained second moments.

The finite constraints can be numerous. A restricted master keeps only a support $T_p \subseteq \mathcal{Q}_p$ per pair. For the following decomposition, let $U$ and $V = c_0 - s + \sum_p t_p$ be feasible primal and dual values for this same restricted master, with each primal distribution supported on $T_p$. If the pair oracle returns $l_p \le \min_{\mathcal{Q}_p} \varphi_p \le v_p$, replace $t_p$ by $\min(t_p, l_p)$. The corrected full-grid lower value $L$ satisfies

$$L \le v_{\mathrm{mom}} \le U, \qquad U - L = \underbrace{U - V}_{\text{restricted-master error}} + \underbrace{\sum_p (t_p - l_p)_+}_{\text{finite separation error}}. \tag{10}$$

Weak duality on these shared supports gives $U - V \ge 0$. The second term is the correction needed to make the finite constraints valid over the full grid. The identity decomposes the computable width $U - L$; the unknown optimization gap $v_{\mathrm{mom}} - L$ is bounded by that width. A codeword with $v_p < t_p$ proves a violated constraint. A straddling interval calls for further separation or an unresolved stop.

There are thus two valid intervals with different targets: $[L, U]$ contains the best certificate value, while $[L, F(\hat{x})]$ contains the finite optimum. If the former is narrow but the latter is wide, more work on the same certificate family has limited value. Numerical solver output must also satisfy the cone and moment equations. The reference repairs both using rational residual bounds; repairing positivity alone is insufficient. The complete arguments, finite exchange and preserved failure examples are in Appendix D.

**Return to the captured example.** For the same QKV objective, the repaired moment witnesses give $L \approx -8.29400312 \times 10^{-5}$ and $U - L < 7.84 \times 10^{-10}$. Thus its reported gap $F(\hat{x}) - L \approx 8.29400 \times 10^{-5}$ contains two different substantial parts: the achievable finite improvement $2.80714 \times 10^{-5}$ and a block-moment relaxation gap for the chosen pair partition between $5.4867879 \times 10^{-5}$ and $5.4868664 \times 10^{-5}$ (outward-rounded endpoints). Improving the codeword can remove the first part. More accurate optimization of this moment family can remove at most $U - L$, and cannot remove the second part.

**Further structural refinements.** Covariance balance can sometimes show that a supplied split already reaches the family target without a joint master (Appendix E). Grid spacing and effective coupling rank give a separate upper bound on block-moment relaxation error, with an attaining family (Appendix C.1).

Table 1: Exact stopping quantities on all eight captured restrictions. Pair error is $F(\hat{x}) - F^\star$ after convergence on the selected pairs. Certificate-optimization width is $U - L$. The block-moment relaxation-gap interval is $[\max(0, F^\star - U), F^\star - L]$. Values are rounded for display; exact rational endpoints are retained. Zero pair errors are exact.

| Captured projection / row | Pair error | Certificate opt. width | Relaxation gap interval |
|---|---|---|---|
| QKV / first | $2.81\times10^{-5}$ | $7.84\times10^{-10}$ | $[5.49\times10^{-5},\ 5.49\times10^{-5}]$ |
| QKV / last | 0 | $1.44\times10^{-9}$ | $[3.25\times10^{-5},\ 3.25\times10^{-5}]$ |
| Attention out / first | 0 | $1.06\times10^{-9}$ | $[0,\ 2.88\times10^{-10}]$ |
| Attention out / last | 0 | $1.86\times10^{-9}$ | $[4.63\times10^{-7},\ 4.65\times10^{-7}]$ |
| MLP up / first | 0 | $1.20\times10^{-9}$ | $[0,\ 5.85\times10^{-10}]$ |
| MLP up / last | 0 | $1.38\times10^{-9}$ | $[0,\ 2.78\times10^{-10}]$ |
| MLP down / first | 0 | $8.82\times10^{-8}$ | $[0,\ 2.91\times10^{-10}]$ |
| MLP down / last | 0 | $8.81\times10^{-10}$ | $[1.11\times10^{-5},\ 1.11\times10^{-5}]$ |

# 6 Exact stopping diagnostics on captured GPT-2 quadratics

The theory separates conditional search, certificate optimization, and relaxation strength. We evaluate these components on exactly enumerable restrictions of stored reconstruction objectives, with additional measurements covering oracle cost and downstream behavior.

## 6.1 Captured quadratics with an exact finite reference

Before execution, we fixed all four first-block GPT-2 projection records from the retained capture, the first and last output row of each, and the first three pairs of their stored matching. This gives eight six-coordinate cases on the original three-bit affine grids. All outside coordinates stay at the stored CD2 values, and their contribution remains in the complete gradient. Each case is therefore a coordinate restriction of the captured quadratic rather than an isolated Hessian-block fit. Affine grid changes and normalization are exact rational transformations; Appendix F specifies them.

For each case, exhaustive integer arithmetic evaluates all $8^6 = 262{,}144$ codewords. Exact selected-pair sweeps give a pair-fixed point. The signed block-moment reference uses all 64 local codewords per pair, with rationally repaired primal and dual certificates. A separate implementation checks the finite optimum, moment identities, and PSD constraints without importing the solver or proposed pair oracle. The six-dimensional restrictions make every stopping component observable against the exact finite optimum.

**Conditional convergence and finite optimality.** Seven pair-fixed points attain the exact finite optimum. The first QKV row instead retains error $2.8071 \times 10^{-5}$ although none of its three selected pairs can improve. The captured objective realizes the cross-pair mechanism of Section 5: conditional pair convergence and finite optimality are distinct stopping statements.

**Optimized certificates reveal structural residuals.** The certificate-optimization widths are below $8.82 \times 10^{-8}$ in all eight cases. Four objectives nevertheless have strictly positive certified block-moment relaxation gaps, ranging from approximately $4.63 \times 10^{-7}$ to $5.49 \times 10^{-5}$. For the other four, the certified upper bound on the relaxation gap is below $5.86 \times 10^{-10}$. Thus a tight family interval can either nearly meet the finite optimum or reveal a remaining structural gap, depending on the objective.

**Finite block moments tighten the captured certificates.** Figure 3 compares the optimized block-moment certificate with $D = \mu I$ for a certified positive $\mu < \lambda_{\min}(H)$ at the same repaired moment mean and with the ordinary continuous-box relaxation. The optimized lower bound is tighter than both controls on all eight stored objectives. The original pair-fixed-anchor bounds are also retained, separating anchor choice from curvature choice. Together these controls isolate the information supplied by finite block moments.

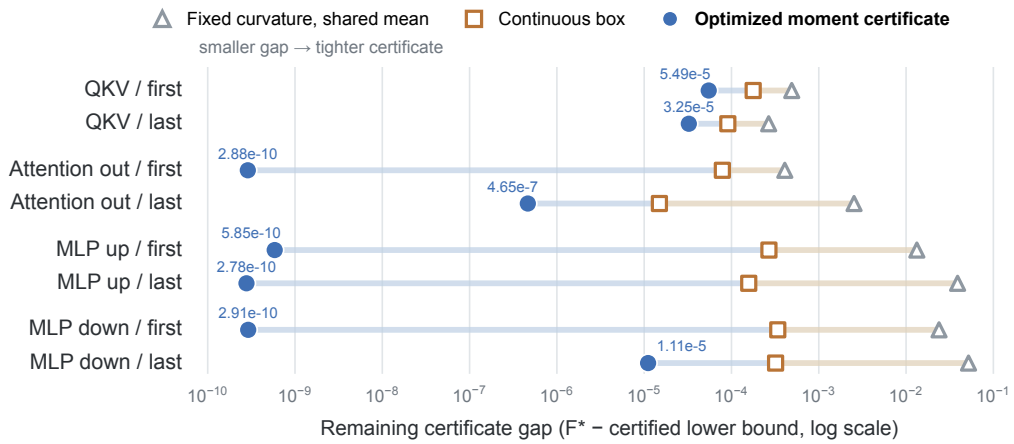



Figure 3: Certificate tightness on the same eight finite objectives, plotted as $F^\star$ minus each certified lower bound. The optimized block-moment certificate is closest to the finite optimum in every case. The fixed-curvature control uses the repaired moment mean as a shared anchor, while the continuous-box control is the classical convex relaxation. All displayed bounds are backed by exact rational checks.

Table 2: Frozen-state separation isolates the finite-separation term of Equation (10). $K_b$ is a per-pair score budget; the last column totals actual scores across the three pairs. Separation error reaches exactly zero in all eight cases at $K_b = 8$.

| Captured projection / row | Remaining separation error $K_b = 0$ | $K_b = 1$ | $K_b = 2$ | Total scores at $K_b = 8$ |
|---|---|---|---|---|
| QKV / first | $1.56 \times 10^{-4}$ | $1.36 \times 10^{-5}$ | 0 | 6 |
| QKV / last | $5.88 \times 10^{-5}$ | $1.78 \times 10^{-5}$ | $1.03 \times 10^{-5}$ | 8 |
| Attention out / first | $7.78 \times 10^{-5}$ | 0 | 0 | 3 |
| Attention out / last | $4.11 \times 10^{-4}$ | $3.68 \times 10^{-5}$ | 0 | 4 |
| MLP up / first | $3.94 \times 10^{-4}$ | $2.63 \times 10^{-5}$ | 0 | 5 |
| MLP up / last | $1.11 \times 10^{-3}$ | $2.92 \times 10^{-4}$ | 0 | 5 |
| MLP down / first | $6.19 \times 10^{-4}$ | $1.95 \times 10^{-4}$ | $1.10 \times 10^{-4}$ | 5 |
| MLP down / last | $1.93 \times 10^{-3}$ | $1.52 \times 10^{-3}$ | $1.17 \times 10^{-4}$ | 7 |

**Budgeted separation resolves a measurable part of certificate error.** Table 2 holds the corrected signed dual state fixed and varies pair-score budgets zero, one, two, and eight. One score per pair removes all separation error in one case; two do so in five; budget eight removes it in all eight, using only three to eight actual finite scores across the three pairs. The previously identified block-moment relaxation gaps remain because they belong to a different layer of the stopping hierarchy. All 96 local budget intervals are independently verified.

### 6.2 Oracle cost and reconstruction relevance

The convex-oracle study checks 24,576 captured pair-row instances and 196,608 numerical method/budget cases, with a separate rational replay of 1,536 intervals. Compiled CPU timings give a 5.1–6.1-fold speedup over constructing and sorting a full branch-bound table. Against direct conditional scanning, the lazy implementation is slightly slower at 3 bits (0.910 vs. 0.858 ms) and faster at 4 bits (0.924 vs. 1.319 ms); Appendix J gives the complete protocol and table. These measurements characterize the local convex oracle, while the signed global reference has its own computational contract.

The model studies test why the certified objective must be specified. In single-layer tests, pair repair reduces mean calibration loss by about 9.9% relative to GPTQ, but adds only about 0.15% beyond the remaining CD3 loss (Figure 6). For whole-model asymmetric three-bit quantization, changing the reconstruction target from runtime to full-reference reduces pair perplexity from 102.27 to 74.83

(Figure 5). These studies establish objective mismatch and the role of stronger scalar controls; the exact stopping claims are tested on the eight fixed quadratics above.

## 7 Related work

**Discrete refinement and optimization.** Fixed-grid quantization provides a setting where discrete refinement decisions are explicit. GPTQ, QuantEase, and CDQuant study layer-wise and coordinate-wise optimization on fixed grids (Frantar et al., 2023; Behdin et al., 2023; Nair & Suggala, 2024); SchurQuant and OJBKQ explore stronger discrete formulations (Lee et al., 2026; Wang et al., 2026); and ReQuant revisits assignments as post-processing (Ma et al., 2026). We complement these refinement procedures by characterizing the computational meaning of a refinement state and the uncertainty remaining after finite search. QuIP# changes the representation through lattice codebooks (Tseng et al., 2024), whereas our analysis fixes the codebook and studies stopping behavior. Conditional minimization and bounded Schnorr–Euchner enumeration provide related search foundations (Chang & Han, 2008; Ahn et al., 2017).

**Certification and optimization relaxations.** Quadratic reformulation and semidefinite bounds are established tools for discrete optimization (Billionnet et al., 2008; Park & Boyd, 2018). Building on nonconvex separable underestimators (Buchheim & Traversi, 2018), we connect the optimized finite-block family to its exact moment target and make unfinished finite separation an explicit part of the stopping interval. Schur complements and verified SDP supply the conic and numerical foundations (Boyd & Vandenberghe, 2004; Jansson et al., 2008). Appendix K gives the detailed comparison.

## 8 Discussion

**Stopping as resource allocation.** The certified gap $F(\hat{x}) - L$ upper-bounds the suboptimality of the returned codeword. It supplies an actionable stopping rule without requiring the finite optimum:

$$F(\hat{x}) - L \leq \epsilon \quad \Longrightarrow \quad F(\hat{x}) - F^\star \leq \epsilon.$$

The terms in $U - L$ direct further computation. If $v_p < t_p$, add the violated codeword and reoptimize the restricted master. If $l_p < t_p \leq v_p$, score further pairs to resolve the interval. A large restricted-master term calls for improving the conic state.

**When the certificate family is nearly optimized.** A small $U - L$ bounds the improvement available from further optimization within the current family. If $F(\hat{x}) - L$ remains large, computation can target a better codeword or a stronger relaxation. Merging blocks cannot weaken the optimized certificate, and one full-codeword block is exact; the cost is a larger finite oracle (Corollary 5). The eight captured restrictions distinguish these targets by enumeration: one pair-fixed point retains finite suboptimality, and four optimized families retain positive relaxation gaps. On larger problems with unknown $F^\star$, these two contributions generally remain unresolved. The framework identifies when the current family is nearly exhausted; choosing between further finite search and stronger relaxation requires additional evidence.

**Reconstruction and downstream quality.** The stopping guarantee applies to the specified quadratic and fixed codebooks. The model studies show that lower reconstruction error can accompany worse perplexity. Objective-level certification and downstream validation thus answer complementary questions: how accurately the chosen objective has been optimized, and how its solution affects model quality. Target and calibration choices determine the link between them.

**Scaling the global certificate.** Global certification for full rows must balance conic optimization, finite separation, and moment repair. The pair timings measure one component; structured solvers and adaptive block merging are routes to controlling total cost. A stopped search and an exhausted certificate are different computational states. Reliable stopping distinguishes them.

## AI-USE STATEMENT

AI assistance through Codex was used extensively in hypothesis refinement, mathematical derivation and proof writing, literature discovery and checking, experiment design, data-preprocessing and method implementation, synthetic mathematical test cases, experimental analysis, figure generation and manuscript drafting. The accompanying finite-certificate artifact permits independent examination of the eight saved rational objectives and their witnesses. This statement describes the research and writing assistance used; it does not substitute for author verification.

## REPRODUCIBILITY STATEMENT

The source archive contains the manuscript, tables and figure PDFs. The separate artifact supplies all eight captured rational objectives, their saved witnesses, input hashes and an independent exact verifier. Appendix A proves the center and ordering results; Appendix L specifies data and model evaluation. Appendix F documents the finite-objective construction, enumeration, moment verification and original capture checks. Appendix M gives the executable replay and its scope.

## REFERENCES

Junil Ahn, Jaewon Chang, and Chiho Lee. A computationally efficient Schnorr-Euchner enumeration for solving integer least-squares problem in wireless communications. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E100-A(1):327–331, 2017. doi: 10.1587/transfun.E100.A.327. URL https://globals.ieice.org/en_transactions/fundamentals/10.1587/transfun.E100.A.327/_p.

Yamato Arai and Yuma Ichikawa. Quantization error propagation: Revisiting layer-wise post-training quantization. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/df2034a516cbd617a96492cc476276c9-Abstract-Conference.html.

Kayhan Behdin, Ayan Acharya, Aman Gupta, Qingquan Song, Siyu Zhu, Sathiya Keerthi, and Rahul Mazumder. QuantEase: Optimization-based quantization for language models. *arXiv preprint arXiv:2309.01885*, 2023. URL https://arxiv.org/abs/2309.01885.

Alain Billionnet, Sourour Elloumi, and Marie-Christine Plateau. Quadratic 0–1 programming: Tightening linear or quadratic convex reformulation by use of relaxations. *RAIRO Operations Research*, 42(2):103–121, 2008. doi: 10.1051/ro:2008011. URL https://www.numdam.org/article/RO_2008__42_2_103_0.pdf.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf.

Christoph Buchheim and Emiliano Traversi. Quadratic combinatorial optimization using separable underestimators. *INFORMS Journal on Computing*, 30(3):424–437, 2018. doi: 10.1287/ijoc.2017.0789. URL https://optimization-online.org/2012/10/3651/.

Xiao-Wen Chang and Qing Han. Solving box-constrained integer least squares problems. *IEEE Transactions on Wireless Communications*, 7(1):277–287, 2008. URL https://www.cs.mcgill.ca/~chang/download.php?filename=ChaH08.pdf.

Santanu S. Dey and Jingye Xu. Asymptotically tight lagrangian dual of smooth nonconvex problems via improved error bound of shapley–folkman lemma. *arXiv preprint arXiv:2601.19003*, 2026. URL https://arxiv.org/abs/2601.19003.

Benjamin Dubois-Taine and Alexandre d'Aspremont. Frank–wolfe meets shapley–folkman: A systematic approach for solving nonconvex separable problems with linear constraints. *Mathematical Programming*, 219:501–551, 2026. doi: 10.1007/s10107-025-02271-8. URL https://arxiv.org/abs/2406.18282.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.17323.

Christian Jansson, Denis Chaykin, and Christian Keil. Rigorous error bounds for the optimal value in semidefinite programming. *SIAM Journal on Numerical Analysis*, 46(1):180–200, 2008. doi: 10.1137/050622870. URL https://tore.tuhh.de/entities/publication/5d120153-6625-4e11-bb2b-a75207250823.

Gunjun Lee, Sehwan Son, Younjoo Lee, Byungjun Kim, and Jung Ho Ahn. SchurQuant: Groupwise discrete optimization for layer-wise LLM quantization. *arXiv preprint arXiv:2608.15567*, 2026. URL https://arxiv.org/abs/2608.15567.

Yongge Ma, Guoan Wang, Feiyu Wang, Yaoming Li, Qian Zhang, Zihan Yan, Yinjun Han, and Tong Yang. ReQuant: Fixed-grid discrete refinement for post-training quantization. *arXiv preprint arXiv:2608.07019*, 2026. URL https://arxiv.org/abs/2608.07019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843.

Pranav Ajit Nair and Arun Sai Suggala. CDQuant: Greedy coordinate descent for accurate LLM quantization. *arXiv preprint arXiv:2406.17542*, 2024. URL https://arxiv.org/abs/2406.17542.

Jaehyun Park and Stephen Boyd. A semidefinite programming method for integer convex quadratic minimization. *Optimization Letters*, 12(3):499–518, 2018. doi: 10.1007/s11590-017-1132-y. URL https://web.stanford.edu/~boyd/papers/pdf/int_least_squares.pdf.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language-models.pdf.

James Saunderson, Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1395–1416, 2012. doi: 10.1137/120872516. URL https://arxiv.org/abs/1204.1220.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. URL https://msp.org/pjm/1958/8-1/pjm-v8-n1-p14-s.pdf.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48630–48656. PMLR, 2024. URL https://proceedings.mlr.press/v235/tseng24a.html.

Madeleine Udell and Stephen Boyd. Bounding duality gap for separable problems with linear constraints. *Computational Optimization and Applications*, 64(2):355–378, 2016. doi: 10.1007/s10589-015-9819-4. URL https://web.stanford.edu/~boyd/papers/pdf/duality_bound.pdf.

Xinyu Wang, Ziyu Zhao, Peng Lu, Yu Gu, and Xiao-Wen Chang. OJBKQ: Objective-joint Babai-Klein quantization. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2373–2381, 2026. doi: 10.24963/ijcai.2026/264. URL https://www.ijcai.org/proceedings/2026/264.

# A Complete arguments and implementation invariants

## A.1 The continuous center, including singular pairs

For fixed $y$, strict positivity of $a$ gives the unique minimizing $x(y) = -(p + by)/a$. Substitution yields Equation (6). Positive semidefiniteness implies $s = c - b^2/a \geq 0$. If $s > 0$, minimization over $[l_y, u_y]$ clips the unconstrained scalar minimizer. If $s = 0$, the objective is linear or constant and the endpoint rule is exact. Consequently $(x_0, y_0)$ minimizes $f$ over $\mathbb{R} \times [l_y, u_y]$, and $x_0$ is a global minimizer of $\phi$ on $\mathbb{R}$.

To justify the final clipping step, consider any convex real-valued function $h$ with a minimizer $x_0$. If $x_0 < l_x$, convexity makes $h$ nondecreasing on $[l_x, \infty)$, so its interval minimum is attained at $l_x$. The case $x_0 > u_x$ is symmetric. Otherwise $x_0$ itself is feasible. This proves the bounded center formula even if the minimizer set is an interval. There is no inverse of the full layer Hessian in this construction.

For ordered affine grid points $x_0^{\mathrm{grid}} < \cdots < x_{L-1}^{\mathrm{grid}}$, split at $j = \max\{t : x_t^{\mathrm{grid}} \leq \bar{x}\}$. The left sequence $j, j-1, \ldots, 0$ and the right sequence $j+1, j+2, \ldots, L-1$ each have nondecreasing $\phi$. The lesser available head is therefore the least remaining bound. Empty sequences have head $+\infty$. Inductively, advancing a consumed head preserves this invariant. This proves the ordering and the stopping lower bound without materializing a sorted table.

The maximum count $k+2$ includes a successor bound computed after the last finite score: that head is needed to report the final gap. Some runs require fewer bounds because a frontier is exhausted or the incumbent already meets the stopping criterion. Each row also requires one center solve. The count does not charge divisions, branch decisions or an active-row selection as if they were free; these operations enter measured implementation time.

## A.2 Conditional finite minimization

For a fixed displacement $x$, write

$$f(x, y) = c\left(y + \frac{q + bx}{c}\right)^2 + ax^2 + 2px - \frac{(q + bx)^2}{c}.$$

With $c > 0$, nearest-grid rounding of the scalar center minimizes this expression over the finite $y$ grid. The grid endpoints enforce clipping. At a midpoint tie either nearest point has the same exact objective. Our search retains the previous incumbent on a non-improving tie, so equal costs need not force a change of codeword. The exact reference orders bounds as rationals and uses the same stated feasibility convention. Numerical parity checks compare objective values when multiple optimal codewords are possible.

The general budget theorem does not require an evaluated prefix; any lower bound on the unvisited branches suffices. Ordered evaluation is useful because the next head supplies such a bound immediately. A full scan of the $L$ branches finds a global optimum within the selected pair. For a prescribed tolerance $\tau$, one can replace the strict-improvement stopping test by $r \geq U - \tau$ while still enforcing the score budget. The reported interval remains $[\min(U, r), U]$.

## A.3 Why certificate drift depends on the full residual

Let $z_R$ be the old external state and compare the full row cost at two feasible pair codewords $\widehat{q}_B$ and $q_B$. Changing $z_R$ to $z_R + \Delta_R$ changes this cost difference by

$$2(\widehat{q}_B - q_B)^\top H_{BR}\Delta_R.$$

Terms depending only on $R$ cancel, as do the fixed linear terms. Since the old cost difference is at most $\Gamma$ for every competitor, its new maximum is at most $\Gamma$ plus the maximum of this linear expression. The Cartesian codebook allows each coordinate to attain the appropriate endpoint independently, giving Equation (43). If grids or $C$ also change, this expression is incomplete and must not be reused unchanged.

The bound applies to the last certified incumbent while external updates occur. A sequence of external changes may be accumulated into their total $\Delta_R$ before forming $\eta$, retaining cancellations.

Summing absolute diameter penalties for individual changes is valid but may be looser. Neither construction turns overlapping pair gaps into a bound on the optimum over all coordinates; the competing set remains the selected pair.

### A.4 THE ALIGNED QUADRATIC AND ITS VERIFICATION

Write $E = Z - W$. The residual in Equation (44) is $X_r E^\top + R_\mu$. Expanding its squared Frobenius norm and adding the damping term gives

$$F_\mu(Z) = \mathrm{tr}(EHE^\top) + 2\langle E, R_\mu^\top X_r/N\rangle + \|R_\mu\|_F^2/N.$$

With $SH = C$ and symmetric $H$, expanding $(E+S)H(E+S)^\top$ produces the cross term $2\langle E, C\rangle$ and constant $\mathrm{tr}(SHS^\top) = \langle S, C\rangle$, proving

$$F_\mu(Z) = \mathrm{tr}((Z - W_\star)H(Z - W_\star)^\top) + K - \langle S, C\rangle. \tag{11}$$

The center is changed solely for the optimizer; feasible weights still come from Equation (1) fitted to original $W$.

Independent tests compare direct activation-space loss with the quadratic, compare its gradient with automatic differentiation, check the completion constant, exhaustively minimize an $8 \times 8$ pair grid, and require exact runtime-endpoint weights. The all-projection run computes empirical moments in FP32 and solves for the center in FP64. It checks relative linear-solve residuals below $10^{-9}$ and completion residuals below $10^{-8}$. These checks validate the stored-coefficient calculation; they do not claim exact real arithmetic for the original activation stream.

## B TWO EXPLICIT RECONSTRUCTION MECHANISMS

### B.1 A SCALAR LOCAL MINIMUM IMPROVED BY A PAIR

Take

$$H = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}, \quad w = (0.6, -0.6)^\top, \quad z = (0,0)^\top, \quad \mathcal{Q}_1 = \mathcal{Q}_2 = \{-1, 0, 1\}.$$

The half-gradient is $g = (-0.06, 0.06)^\top$. Every nonzero feasible scalar move has increment either $0.88$ or $1.12$, so $z$ is a strict scalar coordinate minimum. The pair move $(1, -1)$ has increment $2(-0.06 - 0.06) + 2 - 1.8 = -0.04$. The objective falls from $0.072$ to $0.032$. The eigenvalues of $H$ are $0.1$ and $1.9$, so singularity is unnecessary for this phenomenon. For diagonal $H$, the objective separates and every coordinate-wise exact minimum is a global finite-grid minimum.

### B.2 A BLOCK-ONLY DECREASE THAT RAISES THE COMPLETE OBJECTIVE

Use the same $H$, but let $w = (0.49, 0.49)^\top$, $z = (1, 0)^\top$ and $\mathcal{Q}_1 = \mathcal{Q}_2 = \{0, 1\}$. Changing the first coordinate from 1 to 0 decreases the isolated first-coordinate square from $0.2601$ to $0.2401$. Its change is $-0.02$. The omitted external term is $2(0.9)(-1)(-0.49) = 0.882$, giving a full-objective increase of $0.862$. Indeed, the complete objective increases from $0.05038$ to $0.91238$. This example explains why local block-only traces cannot validate a complete-layer reconstruction claim. Our natural-text block-eight control measures that same distinction on actual activations.

## C GLOBAL CERTIFICATES AND THE FINITE DISTRIBUTION RELAXATION

Keep a fixed finite product codebook and $F(x) = x^\top Hx - 2b^\top x + c_0$. For symmetric block-diagonal $D$ with $G = H - D \succeq 0$, the identity in Equation (7) follows by expanding $F(x)$ around the anchor. It remains true for a continuous anchor $a$. With a feasible seed $s_p$, the existing displacement oracle uses local Hessian $D_p$ and half-gradient $(Ha - b)_p + D_p(s_p - a_p)$; its lower increment is shifted by $f_p^D(s_p - a_p; a)$. This explains why the original conditional Hessian $H_{pp}$ cannot simply be reused in a global certificate.

**A strict local obstruction with weak coupling.** Take two active binary coordinates and independent identity-cost dummies to make two pairs. On the active coordinates let $H = \left[\begin{smallmatrix} 1 & -t \\ -t & 1 \end{smallmatrix}\right]$, $0 < t < 1$, and $w_1 = w_2 = (2 - t)/(4(1 - t))$. At the zero codeword, changing either active coordinate alone costs $t/2 > 0$, while changing both costs $-t$. Each original pair is strictly locally optimal, but the row is not. The active condition number $(1 + t)/(1 - t)$ tends to one as $t \to 0$. Setting active local curvature to $1 - t$ leaves the PSD coupling $t(1, -1)^\top(1, -1)$; the modified global oracle identifies the joint improvement. Hence weak coupling does not justify summing original local gaps.

**Dual and distribution values.** For analysis factor $G = R^\top R$ and define

$$\phi_p(q; y) = q^\top D_p q + 2(R_p^\top y - b_p)^\top q, \tag{12}$$

$$d_D(y) = c_0 - \|y\|^2 + \sum_p \min_{q \in \mathcal{Q}_p} \phi_p(q; y), \tag{13}$$

$$J_D(\pi) = c_0 + \sum_p \mathbb{E}[q^\top D_p q - 2b_p^\top q] + \|\mu\|^2, \quad \mu = \sum_p R_p \mathbb{E} q. \tag{14}$$

Conjugacy of the squared norm and minimax on the finite probability simplices give $\min_\pi J_D(\pi) = \max_y d_D(y)$. One may restrict $y$ to a compact ball containing the bounded means before applying the minimax theorem of Sion (1958). The dual is continuous, coercive and 2-strongly concave, so its maximum exists. An orthogonal component outside $\mathrm{range}(R)$ only subtracts its squared norm; consequently $y = Ra$ at an optimum. This proves the anchor formulation, also when $G$ is singular. For any $l_p \le \min \phi_p(\cdot; y)$, the exact relaxation gap is

$$J_D(\pi) - \left( c_0 - \|y\|^2 + \sum_p l_p \right) = \sum_p (\mathbb{E}\phi_p - l_p) + \|\mu - y\|^2. \tag{15}$$

No local PSD assumption is needed for these finite-domain statements.

**Constructive recovery.** Let $r$ be the dimension of the span of $R_p(q - q_p^0)$ over all blocks and codewords. For any finite represented distribution, form columns $(e_p, R_p q)$ and its positive weight vector $\omega$. If the columns are dependent, take a nonzero null vector $v$ and choose its sign so that $\sum_j v_j(q_j^\top D_{p(j)} q_j - 2b_{p(j)}^\top q_j) \le 0$. Normalization makes $v$ have both signs. Updating $\omega \leftarrow \omega + \alpha v$ with $\alpha = \min_{v_j < 0} \omega_j/(-v_j)$ preserves the coupled mean and normalizations, never increases $J_D$, and deletes a positive weight. On termination there are at most $P + r$ positive weights and hence at most $r$ mixed blocks. This is finite termination, not a practical timing bound. Independent block rounding has exact expectation

$$\mathbb{E}F(X) = J_D(\pi) + \sum_p \mathrm{Var}(R_p X_p). \tag{16}$$

Conditional expectations produce a deterministic codeword no worse than this expectation; conditional errors add. For a Cartesian scalar-grid block, its image lies in a centrally symmetric box image whose opposite vertices are feasible. Its enclosing radius is half its diameter $\Delta_p$, so the variance is at most $\Delta_p^2/4$. This supplies a width-based bound using only the largest $r$ block contributions. These arguments instantiate classical convexification and recovery geometry (Udell & Boyd, 2016; Dubois-Taine & d'Aspremont, 2026); face-aware variance bounds also appear in Dey & Xu (2026).

### C.1 Grid spacing and approximate recovery

If $D$ is diagonal and nonnegative, replace each scalar distribution by the mixture on the adjacent grid points bracketing its mean. The mean is preserved and $\mathbb{E}q_i^2$ cannot increase: consecutive segments of the interpolant of $q_i^2$ have increasing slopes $q_j + q_{j+1}$. Thus $J_D$ does not increase. Compress next. Each remaining mixed coordinate is still supported on an interval of width at most $h_i$, and has image variance at most $\|R_i\|^2 h_i^2/4$. Equation (42) follows from Equation (16). Starting with a certified relaxation error $\delta$ adds at most $\delta$ to the returned-codeword bound. Zero diagonal entries are allowed in this construction.

For general positive definite $D_p$, let $z_p(y)$ minimize $\phi_p$ over the continuous box and choose

$$u_p \geq \max_{|e_i| \leq h_i/2} e^\top D_p e, \qquad R_p^\top R_p \preceq \theta_p D_p. \tag{17}$$

For a pair, the exact box maximum is $u_p = (D_{11}h_1^2 + D_{22}h_2^2 + 2|D_{12}|h_1 h_2)/4$. Nearest grid rounding of $z_p$ preserves every active endpoint; at interior coordinates its gradient vanishes. Its first-order error is therefore zero, and its objective error is at most $u_p$. Box optimality then gives, for every finite $q$,

$$\|q - z_p(y)\|_{D_p}^2 \leq \phi_p(q; y) - \min_{\mathcal{Q}_p} \phi_p(\cdot; y) + u_p. \tag{18}$$

At optimal primal and dual distributions, Equation (15) forces all supported points to be local minimizers. Their image variance is at most $\theta_p u_p$. Compression preserves these supports; thus

$$0 \leq F^\star - J_D^\star \leq \sum_{j=1}^{\min(P,r)} \big(\min\{\Delta_p^2/4, \theta_p u_p\}\big)_{(j)}. \tag{19}$$

For a distribution with certified gap $\delta$, its compressed local excesses $e_p = \mathbb{E}\phi_p - \min \phi_p$ sum to at most $\delta$. Averaging Equation (18) bounds each variance by $\theta_p(u_p + e_p)$. Hence approximate recovery costs at most $(1 + \max_p \theta_p)\delta$ plus the right side of Equation (19) and any rounding errors. Actual variances and the direct returned-codeword gap can be tighter.

**Sharp family.** Use three identical even $L$-level grids with spacing $h$, centered at zero, $D = \beta I$ with $\beta > 0$, and $R = \sqrt{\rho}(1, 1, 1)$ with $\rho > 0$. Set $b = 0, c_0 = 0$. Opposite nearest points in two coordinates and an equal nearest mixture in the third attain $J_D^\star = 3\beta h^2/4$ with zero coupled mean. Every codeword coordinate is an odd multiple of $h/2$, so its sum has magnitude at least $h/2$ and its squared norm at least $3h^2/4$. Nearest balanced signs attain both inequalities, giving $F^\star - J_D^\star = \rho h^2/4$. Direct sums of $r$ copies prove sharp rank dependence. At fixed span $W$, the exact gap is $\rho W^2/[4(L-1)^2]$ per copy. The spectral nearest-lattice rounding bound is classical (Park & Boyd, 2018); this example concerns the tighter finite-distribution relaxation, not uniform dominance of SDP algorithms.

### C.2 What selecting the curvature can recover

For this subsection assume $H \succ 0$ and that each finite block affinely spans its coordinates. Fixed coordinates may first be removed. Let $m_p = \mathbb{E}_{\pi_p} q$ and $M_p = \mathbb{E}_{\pi_p} q q^\top$. Define

$$\begin{aligned} v_{\mathrm{mom}} = \min_{\pi, m, X} \quad & c_0 - 2b^\top m + \operatorname{tr}(HX) \\ \text{subject to} \quad & \pi_p \in \operatorname{simplex}(\mathcal{Q}_p), \quad m_p = \mathbb{E}_{\pi_p} q, \quad X_{pp} = M_p, \\ & \begin{bmatrix} 1 & m^\top \\ m & X \end{bmatrix} \succeq 0. \end{aligned} \tag{20}$$

**Proposition 4** (Curvature and block moments)**.** *The supremum of $J_D^\star$ over symmetric block-diagonal $D$ satisfying $H - D \succeq 0$ equals $v_{\mathrm{mom}}$. Requiring $D_p \succeq 0$ instead gives the value $v_+$ of Equation (20) with $X_{pp} \succeq M_p$ in place of equality. Thus $v_+ \leq v_{\mathrm{mom}} \leq F^\star$.*

*Proof.* Dualize $X_{pp} = M_p$ with symmetric multipliers $D_p$ and Lagrangian term $\operatorname{tr}(D_p(M_p - X_{pp}))$. The remaining $X$ term is $\operatorname{tr}((H-D)X)$. Since the Schur complement gives $X \succeq mm^\top$, its infimum is finite exactly when $H - D \succeq 0$, and then it is $m^\top(H - D)m$. The resulting minimization over

finite block distributions is $J_D^\star$. A strictly positive distribution on each full-span block has positive definite covariance; independent blocks give a positive definite augmented moment matrix. This is a Slater point, and the finite local moments and PSD condition bound all entries of $X$. Conic strong duality proves equality. For $X_{pp} \succeq M_p$ the multipliers must be PSD. Adding a positive multiple of the identity gives strict feasibility; $H \succ 0$ bounds minimizing level sets. The same argument proves the second statement. Rank-one codeword moments prove the final inequality. □

This is a finite-block specialization of the relation between quadratic reformulation and SDP bounds (Billionnet et al., 2008).

**Corollary 5** (Block merging and the exact endpoint)**.** *Let every block of a partition $\mathcal{P}$ be contained in a block of $\mathcal{P}'$, with each block codebook the Cartesian product of its scalar grids. Then*

$$v_{\mathrm{mom}}(\mathcal{P}) \le v_{\mathrm{mom}}(\mathcal{P}') \le F^\star.$$

*For the single-block partition $\mathcal{P}' = \{\{1, \dots, d\}\}$, $v_{\mathrm{mom}}(\mathcal{P}') = F^\star$.*

*Proof.* Marginalizing each coarse-block distribution onto its constituent fine blocks preserves $m$, $X$, and the objective in Equation (20). Every coarse feasible point is therefore fine feasible, proving the first inequality. Point masses at finite codewords prove the second. With one block, $X = \mathbb{E}[qq^\top]$ and $m = \mathbb{E}q$, so the objective equals $\mathbb{E}F(q) \ge F^\star$; a point mass at a minimizer attains equality. □

Equivalently, every fine-block curvature is admissible for the merged family, and its merged finite minimum remains the sum of the original minima. The single-block choice $D = H$ makes the residual curvature zero. A block of $k$ coordinates with $L$ choices each has $L^k$ codewords; the lazy oracle developed here applies to pairs. Merging can strengthen the certificate without guaranteeing strict improvement or lower cost.

**Strict rational separation.** On $\{-1, 1\}^2$, take $H = \left[\begin{smallmatrix} 1 & -1/2 \\ -1/2 & 1 \end{smallmatrix}\right]$, $b = (-1/2, 2)$ and $c_0 = 0$. The four values are $4, -2, 8, -2$. For $D_+ = \mathrm{diag}(3/4, 0)$, anchor $a_+ = (0, 1)$ gives $L_{D_+}(a_+) = -9/4$. A matching feasible relaxed-moment point has $m = a_+$ and $X = \left[\begin{smallmatrix} 1 & 1/2 \\ 1/2 & 5/4 \end{smallmatrix}\right]$; its covariance is $(1, 1/2)^\top(1, 1/2)$ and $X_{ii} \ge M_i = 1$. Thus $v_+ = -9/4$ exactly. For $D_s = \mathrm{diag}(7/8, -1)$ and $a_s = (-1/8, 31/32)$, $G_s = H - D_s \succeq 0$ and $G_s a_s = b$. This gives $L_{D_s}(a_s) = -17/8$. The matching exact-block moment has $m = a_s$ and $X = \left[\begin{smallmatrix} 1 & 1/8 \\ 1/8 & 1 \end{smallmatrix}\right]$; its covariance is $(63/1024)(4, 1)^\top(4, 1)$. Therefore $v_{\mathrm{mom}} = -17/8 < F^\star = -2$. Both gaps are $1/8$. Adding the constant $13/3$ expresses the same example as a squared distance with target $(2/3, 7/3)$, without changing any gap.

### C.3 The signed pair oracle

Consider $f(x, y) = \alpha x^2 + 2\beta xy + \gamma y^2 + 2px + 2qy$ on finite affine grids, with no sign restriction on the symmetric Hessian. If $\gamma > 0$, elimination over the second endpoint interval $[l, u]$ has at most three pieces: its two clipped endpoints and its interior. The endpoint polynomials are $\alpha x^2 + 2(p + \beta t)x + \gamma t^2 + 2qt$ for $t = l, u$; the interior polynomial is $(\alpha - \beta^2/\gamma)x^2 + 2(p - \beta q/\gamma)x - q^2/\gamma$. If $\gamma \le 0$, a second-coordinate endpoint is optimal, and their cost difference is affine in $x$. There are then at most two pieces. Grid indices in each region are contiguous and obtained from the thresholds by clipping and integer rounding.

**Proposition 6** (Signed lazy enumeration)**.** *For any symmetric pair quadratic on finite affine grids, the continuous branch lower bounds admit a merge of at most six monotone streams. After $k$ finite branch scores at most $\min(L_x, k + 6)$ distinct bounds have been evaluated. The same budgeted interval as Theorem 1 remains valid. Complete search uses at most $L_x$ finite branch scores. For $\gamma \le 0$, at most four streams suffice.*

*Proof.* On a contiguous affine-grid segment a quadratic has at most two monotone streams. A convex piece is traversed outward from a discrete minimizer; a concave piece is split at a discrete maximizer and traversed inward from its endpoints. A linear piece needs one direction. Clipped nearest rounding of the vertex finds these discrete extrema without a bound table. The disjoint region and stream partitions cover every first-grid index. Merge the heads using a heap of size at

most six. Its minimum is the smallest unvisited lower bound; consuming a head introduces at most one successor. This proves the count. A finite second-coordinate minimum uses nearest rounding if $\gamma > 0$ and an endpoint otherwise, completing the budget and exhaustion arguments. □

The count is conservative and does not assert optimality of the constant six. Region, vertex, endpoint-selection and merge arithmetic is additional work. Our exact reference also permits a zero-score budget and singleton grids. Nine targeted pair fixtures check all stream partitions and bound values against independent endpoint/stationary-point relaxation and full finite enumeration. Matching rational moment witnesses check Proposition 4's strict example without a numerical SDP solver. No compiled signed-kernel or model-performance claim follows from these checks.

The exact checks cover signed-curvature search, grid-spacing and PSD-tail bounds, and the implicit global certificate. The portable replay described in Appendix M separately verifies the captured finite objectives.

# D JOINT CURVATURE OPTIMIZATION WITH CERTIFIED FINITE SEPARATION

This section makes the family in Appendix C.2 directly optimizable and gives a stopping test that includes incomplete finite search. The matrix-fractional epigraph and singular Schur complement are classical (Boyd & Vandenberghe, 2004, Example 3.4 and Appendix A.5.5); separable quadratic underestimation is established (Buchheim & Traversi, 2018). Our construction specifies their finite-oracle interface and endpoint corrections for the supplied quadratic. Classical verified SDP checks linear equalities together with positivity (Jansson et al., 2008, Theorem 4.1); our finite-moment repair instantiates that requirement.

## D.1 A CONVEX REPRESENTATION AND ITS PRIMAL WITNESS

Let $D$ be symmetric and block diagonal, without a local sign restriction. For variables $u, s, t$, define

$$\mathcal{Z} = \begin{bmatrix} H - D & u \\ u^\top & s \end{bmatrix} \succeq 0, \quad \varphi_p(q) = q^\top D_p q + 2(u_p - b_p)^\top q, \tag{21}$$

$$\mathcal{D} = \max_{D,u,s,t} \left\{ c_0 - s + \sum_p t_p : t_p \le \varphi_p(q) \quad (q \in \mathcal{Q}_p) \right\}. \tag{22}$$

The cone condition in Equation (21) is part of the maximization. Every constraint and the objective are affine in its variables. For valid finite lower values $l_p$, set $L = c_0 - s + \sum_p l_p$. Direct expansion gives, for every finite codeword,

$$F(x) - L = \begin{bmatrix} x \\ -1 \end{bmatrix}^\top \mathcal{Z} \begin{bmatrix} x \\ -1 \end{bmatrix} + \sum_p [\varphi_p(x_p) - l_p] \ge 0. \tag{23}$$

Write $G = H - D$. The singular Schur complement is $G \succeq 0$, $u \in \mathrm{range}(G)$ and $s \ge u^\top G^\dagger u$. The range condition cannot be omitted. Choose $a$ with $Ga = u$. At the best $s = a^\top G a$, the bound equals the continuous-anchor bound in Appendix C. Conversely, any admissible curvature and anchor give this feasible cone point. Thus Equation (22) optimizes curvature and anchor together, including singular $G$, without alternating bilinear variables.

For finite block distributions $\pi_p$, let $m_p = \mathbb{E}q$, $M_p = \mathbb{E}qq^\top$. The moment primal minimizes

$$U = c_0 - 2b^\top m + \mathrm{tr}(HX), \qquad \mathcal{S} = \begin{bmatrix} X & -m \\ -m^\top & 1 \end{bmatrix} \succeq 0, \qquad X_{pp} = M_p. \tag{24}$$

Its optimum is $v_{\mathrm{mom}}$. The duality argument in Appendix C.2 gives $\mathcal{D} = v_{\mathrm{mom}}$ for $H \succ 0$ and full affine span of each block; fixed scalar coordinates can first be removed. In particular, endpoint-corner distributions with full support give a strictly feasible moment matrix for nondegenerate Cartesian grids. For any feasible primal and full-grid dual lower values,

$$U - L = \mathrm{tr}(\mathcal{Z}\mathcal{S}) + \sum_p (\mathbb{E}_{\pi_p}\varphi_p - l_p) \ge 0. \tag{25}$$

Indeed, the trace is $\mathrm{tr}((H - D)X) - 2u^\top m + s$; using $X_{pp} = M_p$ and collecting terms proves the identity. A PSD trace product is nonnegative, as are the local excesses. Hence $[L, U]$ encloses the optimum of the entire curvature/anchor family for this partition. For a pure feasible $x$, $L \le F^\star \le F(x)$ is a different interval. In particular, a small $U - L$ does not imply a small block-moment relaxation gap.

## D.2 RESTRICTED SUPPORTS, INCOMPLETE SEPARATION AND STOPPING

Initialize $T_p \subseteq \mathcal{Q}_p$ with all distinct endpoint corners. The restricted master keeps only constraints at $q \in T_p$ in Equation (22), or equivalently restricts the primal distributions to $T_p$. Its primal is still feasible for the full moment problem, so any feasible objective $U$ bounds $v_{\mathrm{mom}}$ from above. Its restricted dual alone need not bound the full-grid optimum from below.

**Proposition 7** (Certified exchange interval)**.** *Suppose the restricted master supplies feasible primal and dual points, with objectives $U$ and $V = c_0 - s + \sum_p t_p$. Let $l_p \le \min_{\mathcal{Q}_p} \varphi_p \le v_p$ be finite-oracle intervals, where $v_p$ is achieved by a returned codeword. Define*

$$\epsilon_{\text{master}} = U - V, \qquad L = V - \sum_p (t_p - l_p)_+. \tag{26}$$

*Then $L \le v_{\text{mom}} \le U$ and $U - L = \epsilon_{\text{master}} + \sum_p (t_p - l_p)_+$. If $v_p < t_p$, the returned point proves a new violated constraint. If $l_p < t_p \le v_p$, the available budget is inconclusive.*

*Proof.* Replace each restricted scalar by $t'_p = \min(t_p, l_p)$. It now bounds every point of $\mathcal{Q}_p$ while leaving the cone unchanged. Its objective is $L$, so weak duality gives the interval. Subtracting from $U$ proves the gap decomposition. A returned point of value below $t_p$ cannot belong to $T_p$, because the restricted dual is feasible there. An interval straddling $t_p$ establishes neither a violation nor complete separation. □

At a feasible seed $z_p$, the signed displacement oracle takes local Hessian $D_p$ and half-gradient $D_p z_p + u_p - b_p$. Adding $\varphi_p(z_p)$ to its increment endpoints yields $[l_p, v_p]$. The caller can add a proved violating point, raise only an inconclusive block's budget, or return the current certified family interval. No anchor inversion is required. The $K + 6$ bound applies per pair-oracle call; repeated budgets, master work, support scoring and matrix correction are additional costs.

With exact master solves and exact finite separation, every nonterminal round adds at least one new point. There are consequently at most $\sum_p (|\mathcal{Q}_p| - |T_p^{(0)}|)$ nonterminal rounds. Once no violation exists, the restricted optimum is full-grid dual feasible and the family interval closes. This conditional finite-domain argument is not a polynomial-time guarantee. Approximate solves instead use Proposition 7; a numerical status code is not its feasibility premise. Intersecting all valid intervals gives a running maximum lower bound and minimum upper bound, even when individual candidate bounds are nonmonotone.

### D.3 Exact corrections of numerical proposals

Remove fixed coordinates and map each remaining endpoint interval to $[-1, 1]$ by an exact affine substitution, transforming $H, b, c_0$ as well. The following corrections apply to that normalized problem.

**Dual feasibility.** Suppose a reconstructed symmetric candidate $\mathcal{Z}_0$ has a certified lower spectral bound $-\epsilon_D$, with $\epsilon_D \ge 0$. Set $D' = D_0 - \epsilon_D I$, $s' = s_0 + \epsilon_D$ and $u' = u_0$. Then $\mathcal{Z}' = \mathcal{Z}_0 + \epsilon_D I \succeq 0$. Since $\|q_p\|^2 \le d_p$, an old finite lower value $l_p$ remains valid after subtracting $\epsilon_D d_p$. The total lower-bound allowance is

$$\epsilon_D(d + 1). \tag{27}$$

Re-querying the corrected curvature can yield a tighter bound. Signed curvature permits this correction; an additional constraint $D_p \succeq 0$ would need separate treatment.

**Primal feasibility.** First clip negative proposed probabilities to zero and normalize exactly, using the endpoint reference if a block has zero remaining mass. Recompute $m_p, M_p$ and set $X_{pp} = M_p$ exactly; retain symmetric proposed cross moments. Only then bound $\mathcal{S}_0 \succeq -\epsilon_P I$. The independent uniform endpoint reference has $m_{\text{ref}} = 0$, $X_{\text{ref}} = I$ and $\mathcal{S}_{\text{ref}} = I$. Mix both distributions and moments:

$$\pi'_p = \frac{\pi_p + \epsilon_P \pi_{p,\text{ref}}}{1 + \epsilon_P}, \quad \mathcal{S}' = \frac{\mathcal{S}_0 + \epsilon_P I}{1 + \epsilon_P}, \quad U' = \frac{U_0 + \epsilon_P U_{\text{ref}}}{1 + \epsilon_P}. \tag{28}$$

This preserves $X'_{pp} = M'_p$, the bottom-right entry one and positivity. The endpoint support is already in each restricted master, so the repaired primal remains feasible there. Correcting only matrix positivity would not preserve the finite moment relaxation.

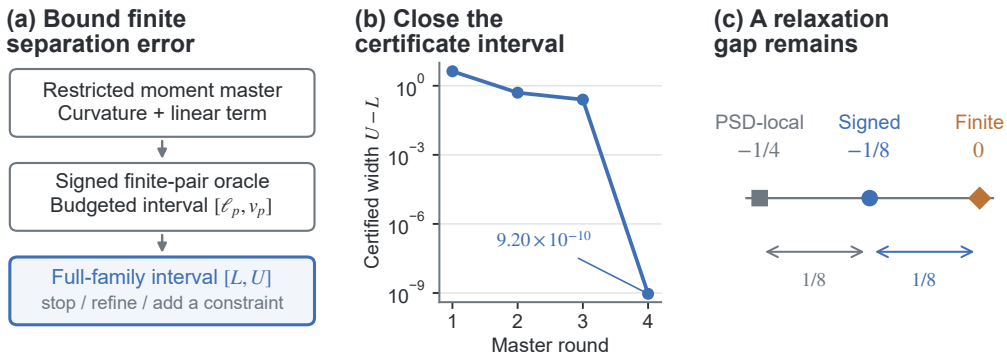



Figure 4: Distinct stopping gaps require different computations. (a) Budgeted finite separation controls search error during curvature optimization. (b) Intersections of verified intervals for the six-coordinate odd-coupling instance; raw candidate lower bounds need not improve monotonically. (c) A separate four-coordinate binary instance distinguishes the local-PSD restriction from the remaining block-moment relaxation gap. Each layer requires its own check. Both instances have exact primal/dual witnesses (Appendix D.5). Master solves, support scoring and matrix correction are additional to finite-oracle work.

**A bound independent of numerical eigenvalue accuracy.** For any exact rational symmetric $A$ and any rational factor $T$, compute $E = A - TT^\top$ exactly and set $\epsilon = \max_i \sum_j |E_{ij}|$. Then $\|E\|_2 \le \epsilon$ and $A \succeq -\epsilon I$. A numerical eigendecomposition can propose $T$ by clipping negative eigenvalues and rationalizing the resulting factor. Validity depends on the exact residual, not the eigensolver's accuracy. Our dense reference uses denominator $10^{10}$ for numerical proposals and adds a positive $10^{-12}$ margin to the residual allowance. Independent exact leading principal determinants verify the corrected positive definite matrices. This dense rational work is not counted as finite branch evaluations.

### D.4 Two failures prevented by the validity conditions

**Dropping the singular range condition.** Take $H = I_2$, $b = (0, 1)$, $c_0 = 0$, $\mathcal{Q} = \{-1, 1\}^2$, $D = \mathrm{diag}(0, 1)$, $G = \mathrm{diag}(1, 0)$, $u = (0, 1)$ and $s = 0$. Although $u^\top G^\dagger u = 0$, the vector $u$ is outside $\mathrm{range}(G)$ and $\mathcal{Z}$ is indefinite. The putative lower value is 1, whereas the true optimum is 0. Adding $I$ to $\mathcal{Z}$ gives valid value $-2$ after local correction, with loss $3 = (d + 1)\epsilon_D$.

**Repairing positivity but violating finite moments.** Use the binary two-coordinate witness of Appendix C.2, with $H = \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix}$, $b = (-1/2, 2)$ and $c_0 = 0$. At $m = (0, 1)$, the matrix $X_0 = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$ has an indefinite covariance. Increasing only $X_{22}$ to $5/4$ makes $X - mm^\top = (1, 1/2)(1, 1/2)^\top \succeq 0$, but the second finite moment must still equal 1. This invalid point has objective $-9/4$, below the valid moment optimum $-17/8$. Matrix positivity alone therefore cannot certify a primal upper bound for the required relaxation.

### D.5 Executed small instances and analytical witnesses

The reference uses CVXPY 1.6.5 and Clarabel 0.10.0, with absolute gap, relative gap and feasibility tolerances $10^{-11}$ and at most 250 master iterations. Every numerical candidate is reconstructed and corrected before use. Endpoint corners initialize each pair support; finite score budgets $0, 1, 2, L_x$ are increased only as needed. The stopping tolerance on the certified family interval is $10^{-6}$.

Table 3: Joint certificate reference. Exact optima have matching rational primal/dual witnesses; widths are from independently verified corrected numerical output. Rounds count restricted master solves, not equal work.

| Instance | $v_{\mathrm{mom}}$ | $F^\star$ | Rounds | Final $U - L$ |
|---|---|---|---|---|
| Odd coupling, 6 coordinates | $1/2$ | $19/36$ | 4 | $9.20 \times 10^{-10}$ |
| Signed witness, 4 coordinates | $-1/8$ | $0$ | 1 | $1.61 \times 10^{-9}$ |

**Odd-coupling instance.** Let $v = (1,0,1,0,1,0)^\top$, $H = \frac{3}{4}I + \frac{1}{4}vv^\top$, $b = 0$, $c_0 = 0$ and use $\{-1, -1/3, 1/3, 1\}$ on each coordinate, partitioned as $(1,2),(3,4),(5,6)$. For the dual witness, take $D = \frac{3}{4}I$, $u = 0$, $s = 0$: the coupling is PSD and the sum of finite local minima is $1/2$. For the matching moment primal, take mean zero, diagonal entries $1/9$, and the active-coordinate submatrix $X_{\{1,3,5\}} = \frac{1}{18}L_{K_3}$, where $L_{K_3}$ is the complete three-vertex graph Laplacian. All other off-diagonals vanish. This matrix is PSD, agrees with uniform central-corner pair moments, and has $v^\top X v = 0$, proving objective $1/2$. Every finite coordinate has squared magnitude at least $1/9$; the sum of three active odd multiples of $1/3$ has squared magnitude at least $1/9$. Central points attaining both bounds give $F^\star = \frac{1}{2} + \frac{1}{36}$.

The four master rounds use 4, 5, 6 and 7 support points per pair, out of 16 possible. They add nine points in total. Across all budget attempts, the signed oracle records 18 finite branch scores and 72 branch-bound evaluations; current-support scoring, the SDP solves, rational corrections and verification are additional work. One master reports `optimal_inaccurate`. Raw lower candidates are nonmonotone; Figure 4(b) uses intersections of the saved valid intervals.

**Signed-curvature instance.** Let $H = I_4$ except $H_{13} = H_{31} = -1/2$, $b = (-1/2, 0, 2, 0)$ and $c_0 = 0$, with binary $\{-1, 1\}$ coordinates and pairs $(1,2),(3,4)$. Take $D = \mathrm{diag}(7/8, 1, -1, 1)$, $u = b$, $s = 2$. Writing $w = (1/4, 0, -1, 0)$ gives $\mathcal{Z} = 2(w,-1)(w,-1)^\top \succeq 0$ and lower value $-1/8$. A matching primal has $m = (-1/8, 0, 31/32, 0)$, unit diagonal second moments, $X_{13} = X_{31} = 1/8$ and other cross moments zero. Its active covariance is $\frac{63}{1024}(4,1)(4,1)^\top$; independent dummy coordinates have variance one. Feasible binary pair distributions realize these local moments. Direct substitution gives $U = -1/8$, while finite enumeration gives $F^\star = 0$. The local-PSD value is $-1/4$, obtained by extending the two-coordinate witnesses with two unit-cost dummies. The returned signed reference thus separates the curvature-family restriction from the remaining block-moment relaxation gap. Endpoint corners are already the complete local codebooks here: one master and no extra oracle separation calls suffice. This is not a zero-cost solve.

**Independent replay and scope.** An independent script imports neither the solver nor the research kernels. It checks all five corrected master states, 4,112 full codewords, 200 full local values, every recorded finite interval, exact moment equalities, new cuts and Equation (25). Five uncorrected candidate matrices have exact negative leading principal minors; all corrected matrices pass the independent positivity checks. Reported branch counts are checked against their budgets, not independently profiled as instructions. The two adverse examples are also verified exactly. These examples test validity and distinct mathematical gaps; they establish neither large-scale SDP efficiency nor model-quality gains.

The completed raw records and independent reports bind the checked source versions by SHA-256; the figure builder checks the saved record hashes. These checks validate the reference construction above. The portable replay in Appendix M verifies the captured-objective certificates without another solver run.

## E When a fixed split reaches the joint certificate

The joint master is a reference for certificate strength. This section asks when a supplied inexpensive split already reaches it. The answer depends on completing local covariances, not just minimizing their sum. The scalar balance geometry and the block-coherence sufficient theorem are classical (Saunderson et al., 2012). We connect them to the finite-block family interval, including a quantita-

tive completion cost, explicit signed binary certificates, and a distinction between certification and rounding.

### E.1 Exact reduction to projected covariance completion

Fix $H = D + R^\top R$ with symmetric block-diagonal $D$ and $k$-row $R$. For finite distributions $\pi_p$, let $m_p = \mathbb{E}q_p$, $M_p = \mathbb{E}q_p q_p^\top$ and $V_p = M_p - m_p m_p^\top$. Write

$$J_D(\pi) = c_0 - 2b^\top m + \sum_p \operatorname{tr}(D_p M_p) + \|Rm\|^2. \tag{29}$$

This equals $F(m) + \sum_p \operatorname{tr}(D_p V_p)$ in the main text. Define

$$\Psi_R(V) = \min_{C \succeq 0,\ C_{pp} = V_p} \operatorname{tr}(RCR^\top). \tag{30}$$

The feasible set contains $\operatorname{blkdiag}(V_p)$ and is compact: PSD bounds each cross entry by the square root of its two diagonal entries. Thus the minimum is attained. Substituting $X = mm^\top + C$ in the moment primal of Appendix D proves

$$v_{\mathrm{mom}} = \min_\pi \{J_D(\pi) + \Psi_R(V(\pi))\}, \qquad L \le J_D^\star \le v_{\mathrm{mom}} \le J_D(\pi) + \Psi_R(V(\pi)). \tag{31}$$

Here $L$ is any valid fixed-split lower bound. Identification with the joint dual uses the positive-definite and full-affine-span assumptions stated in Appendix D. A completion preserves local moments, but need not arise from a joint distribution with the prescribed finite marginals. Its objective is an upper value for $v_{\mathrm{mom}}$, not necessarily for $F^\star$.

**Proposition 8** (Projection and lifting)**.** *Set $A_p = R_p V_p R_p^\top$. Then*

$$\Psi_R(V) = \Phi(A) := \min_{Y \succeq 0,\ Y_{pp} = A_p} \sum_{p,q} \operatorname{tr}(Y_{pq}). \tag{32}$$

*The identity includes singular and zero $A_p$.*

*Proof.* Project a feasible $C$ by $Y_{pq} = R_p C_{pq} R_q^\top$. Conversely define

$$K_p = V_p R_p^\top A_p^\dagger, \quad E_p = V_p - V_p R_p^\top A_p^\dagger R_p V_p, \quad C = KYK^\top + \operatorname{blkdiag}(E_p), \tag{33}$$

where $K$ is block diagonal. A square-root factor of $V_p$ expresses $E_p$ as its congruence with an orthogonal complementary projector; hence $E_p \succeq 0$ and $R_p E_p R_p^\top = 0$. PSD of $Y$ forces the rows and columns of $Y_{pq}$ into the ranges of $A_p$ and $A_q$: a vector in a diagonal block's kernel has zero quadratic form and therefore lies in the full PSD matrix's kernel. Consequently $R_p K_p = A_p A_p^\dagger$ acts as identity on the relevant range. Equation (33) has diagonal blocks exactly $V_p$, maps back to $Y$, and preserves the objective. □

The projected matrix has dimension $kP$; it is not automatically smaller than the original matrix. The next results avoid constructing that matrix in specific cases.

### E.2 Rank-one and common-direction formulas

For nonnegative lengths $a_1, \dots, a_P$, define $\Delta(a) = (2\max_p a_p - \sum_p a_p)_+$. Vectors of these lengths have minimum resultant length $\Delta(a)$. The triangle inequality proves the lower bound. For attainability, if the current resultant ranges over $[l, u]$, adding a length $a$ yields the connected interval with maximum $u + a$ and minimum $\max(0, l - a, a - u)$. Induction gives $[\Delta, \sum_p a_p]$; planar vectors attain every value. Zero lengths are allowed. Applying this Gram-vector argument to (32) gives, for $R = r^\top$,

$$\Psi_r(V) = \Delta\big((\sqrt{r_p^\top V_p r_p})_p\big)^2. \tag{34}$$

It follows from (31) that the fixed split reaches $v_{\mathrm{mom}}$ if and only if *some* minimizer of $J_D$ has balanced lengths. Necessity uses attainment and the two nonnegative excesses in $J_D(\pi) + \Psi - J_D^\star$. One unbalanced minimizer is insufficient to prove the split suboptimal; another may be balanced.

**Proposition 9** (Directional lower bound and commuting exactness)**.** *For any orthonormal basis* $(q_j)_{j=1}^k$*, put* $a_{pj} = \sqrt{q_j^\top A_p q_j}$ *and* $\Delta_j = \Delta((a_{pj})_p)$*. Then* $\sum_j \Delta_j^2 \le \Phi(A)$*. Equality holds if this basis simultaneously diagonalizes all* $A_p$*, in particular for commuting PSD matrices.*

*Proof.* For feasible $Y$, each matrix $G_j = (q_j^\top Y_{pq} q_j)_{p,q}$ is PSD with diagonal $a_{pj}^2$. Its entry sum is at least $\Delta_j^2$ by the scalar result. Trace in the orthonormal basis gives the stated lower bound. If every $A_p$ is diagonal in $Q = (q_1, \dots, q_k)$, choose independent optimal scalar Gram matrices $G_j$ and set $Y_{pq} = Q\,\mathrm{diag}((G_j)_{pq})_j Q^\top$. An orthogonal transformation and permutation make $Y$ a direct sum of the $G_j$, proving PSD. It has the required diagonal blocks and attains the bound. Symmetric commuting matrices have a common orthonormal eigenbasis, obtained by successive diagonalization on invariant eigenspaces. □

Thus $J_D(\pi) - L + \sum_j \Delta_j^2$ bounds the possible improvement of the joint family over $L$ under this condition. For noncommuting matrices, directional costs remain lower bounds on the supplied completion problem, not upper bounds and not uniform lower bounds over all distributions. Dropping off-diagonal entries would change the required moments.

### E.3 General low-rank sufficiency and partial cancellation

**Proposition 10** (A small-matrix sufficient zero test)**.** *Let* $T = \sum_p A_p$*. If* $T = 0$*, then* $\Phi(A) = 0$*. More generally, if for some* $0 < \epsilon \le 1$*,*

$$(1-\epsilon)T - 2A_p \succeq 0 \quad \textit{for every } p, \tag{35}$$

*then* $\Phi(A) = 0$*. Neither commutativity nor invertibility is required.*

*Proof.* Factor $V_p = S_p S_p^\top$ using its nonzero support and set $S = \mathrm{blkdiag}(S_p)$, $Z = RS$. The same PSD range argument as above makes a completion equivalent to $C = SWS^\top$, where $W \succeq 0$ and $W_{pp} = I$. Zero cost is equivalent to $WZ^\top = 0$. Thus $\mathcal{U} = \mathrm{range}(Z^\top)$ must be realizable as a subspace of the nullspace of a block correlation matrix. Its projector is $Z^\top T^\dagger Z$, whose $p$-th diagonal block has the same nonzero eigenvalues as $T^{\dagger/2} A_p T^{\dagger/2}$. On the range of $T$, (35) bounds these eigenvalues by $(1-\epsilon)/2$. Hence the block coherence is strictly below $1/2$. Corollary 5.11 of Saunderson et al. (2012) supplies the required $W$. For $T = 0$, every $A_p = 0$ and the block diagonal completion already has zero cost. □

This is an application of the established block-coherence theorem, not a new coherence threshold. The test is only sufficient: scalar variances $(1, 1, 4)$ fail it for every positive $\epsilon$, although lengths $(1, 1, 2)$ are balanced and the completion cost is zero.

When complete cancellation fails, take supplied matrices $0 \preceq K_p \preceq A_p$ and $E_p = A_p - K_p$. Adding $\mathrm{blkdiag}(E_p)$ to a completion for $K$ proves

$$\Phi(A) \le U_K + \sum_p \mathrm{tr}(E_p), \qquad v_{\mathrm{mom}} - L \le J_D(\pi) - L + U_K + \sum_p \mathrm{tr}(E_p), \tag{36}$$

for any certified $U_K \ge \Phi(K)$. Use zero from Proposition 10 or the common-basis formula from Proposition 9, and take the minimum with the independent-completion upper $\mathrm{tr}(T)$. The auxiliary $K_p$ need not have finite distribution realizations; the completed and lifted covariance restores the original local moments. Finding a useful PSD-ordered core is additional work, and simply removing off-diagonal entries is invalid.

For supplied rational matrices, exact PSD elimination checks the margins in (35) using $O(Pk^3)$ rational arithmetic operations. A positive pivot permits a Schur complement; a zero pivot requires the remaining row to vanish, and a negative pivot rejects PSD. This follows by congruence and includes singular matrices. The checker does not form a pseudoinverse, projector or full moment matrix. Moment construction, split certification, finite-oracle work and bit growth are excluded from the operation count. For the common-basis routine, rational orthogonality and diagonalization are checked exactly; general irrational eigenbases require a different verified representation.

Square roots need not be exact. If $\alpha_p \le a_p \le \beta_p$, then

$$\max(0, \max_p[\alpha_p - \sum_{q\neq p} \beta_q]) \le \Delta(a) \le \max(0, \max_p[\beta_p - \sum_{q\neq p} \alpha_q]). \tag{37}$$

For a rational squared length $n/d$ and denominator $N = 2^t$, let $h = \mathrm{isqrt}(\lfloor nN^2/d \rfloor)$. The endpoints $h/N$ and $(h+1)/N$ enclose the root; both are $h/N$ when $h^2 d = nN^2$. Integer comparisons prove validity. Squaring and summing the nonnegative imbalance endpoints gives outward completion bounds. An irrational balance equality may retain a small positive allowance.

### E.4 A SOLVED BINARY FAMILY AND A PAIR-PARTITION BOUNDARY

Consider centered binary grids $\mathcal{Q}_i = \{-h_i, h_i\}$, $h_i > 0$, $H = \mathrm{diag}(\beta_i) + rr^\top$, $\beta_i > 0$, and $b = 0$. Write $a_i = |r_i|h_i$, $C_0 = c_0 + \sum_i \beta_i h_i^2$ and $\Delta = (2\max_i a_i - \sum_i a_i)_+$. Then

$$v_{\mathrm{mom}} = C_0 + \Delta^2, \qquad F^\star = C_0 + \min_{\sigma\in\{-1,1\}^d} (\sum_i \sigma_i a_i)^2. \tag{38}$$

The objective depends only on $X$; every scalar moment point can use mean zero and $X_{ii} = h_i^2$. Equation (34) proves the first identity, and direct substitution proves the second.

If $\Delta > 0$, let $k$ be the unique dominant index. Choose a finite $x$ with $r_k x_k = a_k$ and $r_i x_i = -a_i$ for $i \neq k$, with either sign when $r_i = 0$. The explicit diagonal curvature

$$\widehat{D}_{ii} = \beta_i + r_i\Delta/x_i \tag{39}$$

gives the matching lower and finite value $C_0 + \Delta^2$. To prove admissibility, conjugate $H - \widehat{D}$ by $\mathrm{diag}(x)$, write $s_i = r_i x_i$ and $y_i = y_k + d_i$ for $i \neq k$. Expansion gives

$$y^\top[ss^\top - \Delta\,\mathrm{diag}(s)]y = \big(\sum_{i\neq k} a_i d_i\big)^2 + \Delta \sum_{i\neq k} a_i d_i^2 \ge 0. \tag{40}$$

Every local binary value $\widehat{D}_{ii} q_i^2$ is constant, so $u = s = 0$ in the joint cone proves the lower bound. Also $r^\top x = \Delta$. The curvature can be negative. The balanced case is joint-family exact for the original split but still contains the signed-partition problem; it need not be finite-exact. Translation must preserve $b = 0$ for this zero-field result to apply.

More generally, for *any* symmetric $H$, centered binary grids and $b = 0$, partitions into singletons and pairs give the same moment value. Indeed any scalar feasible $X \succeq 0$, $X_{ii} = h_i^2$ can take $m = 0$. For a pair set $\rho = X_{ij}/(h_i h_j) \in [-1, 1]$ and assign $\pi(\sigma_i, \sigma_j) = (1 + \rho\sigma_i\sigma_j)/4$. These nonnegative probabilities realize its exact pair moments; the reverse inclusion is immediate. Pairing alone therefore cannot remove this relaxation gap. A triple block can enforce a strictly stronger constraint, as the three-sign odd-coupling example demonstrates.

### E.5 COMPRESSION, DIRECTIONAL INCOMPATIBILITY AND EXACT CHECKS

For three binary coordinates with $D = I$, $r = (1, 1, 1)$ and $b = c_0 = 0$, uniform distributions give $J_D = 3$ and balanced variances, certifying $v_{\mathrm{mom}} = 3$ although $F^\star = 4$. Fixing the first two coordinates to $+1, -1$ and leaving the third uniform preserves the *coupled* mean and $J_D$, but not each local mean. The independent-rounding variance drops from 3 to 1 while the completion cost rises from 0 to 1. Retaining the original distribution preserves the sharper family upper. The support-compression theorem does not promise this covariance balance and is not contradicted.

For three scalar blocks with $V_i = 1$ and $R = [(1, 0)^\top\ (0, 1)^\top\ (1/2, 1/2)^\top]$, total lengths $1, 1, 1/\sqrt{2}$ are balanced, yet $\Psi = 1/2$. The coordinate directional lower is $1/4 + 1/4$; covariance $zz^\top$, $z = (1, 1, -1)$, attains it. Treating total variances as scalar lengths would falsely assert zero completion. The exact common-basis routine rejects these noncommuting matrices.

Table 4: Exact fixed-split, joint-family and finite optima. Closing the curvature gap need not close the discrete gap. These are proof witnesses, not performance measurements.

| Finite quadratic | $J_D^\star$ | $v_{\mathrm{mom}}$ | $F^\star$ |
|---|---|---|---|
| Rank one, balanced odd | 3 | 3 | 4 |
| Rank one, dominant signed | $9/4$ | $17/2$ | $17/2$ |
| Rank two, both directions balanced | 6 | 6 | 8 |
| Rank two, one dominant direction | 6 | 7 | 8 |

The two rank-two finite witnesses use three binary pairs, $D = I_6$, $b = c_0 = 0$. Their coupling blocks are $[I_2, I_2, I_2]$ and $[\mathrm{diag}(3, 1), I_2, I_2]$. Uniform pair distributions attain the fixed value 6. Directional polygon completions give moment values 6 and 7; admissible diagonal curvatures $(1, 1, 1, 1, 1, 1)$ and $(4, 1, 0, 1, 0, 1)$ match them. In both cases the best finite value is 8. Thus the second split can improve by precisely 1, while the first already reaches the joint target. Separate explicit witnesses check noncommuting covariance cancellation, a singular total covariance and a PSD core with upper $19/2$ instead of independent variance $31/2$. The latter is an upper bound, not an exact completion claim.

The exact verifiers check centered moments and primal/dual witnesses, enumerating 133 rank-one and 128 low-rank codewords. Eight additional covariance witnesses include singular cases. Independent principal-minor checks validate rational elimination; reports retain source hashes, root enclosures and invalid inputs. These mathematical checks use no SDP solver and imply no model-quality or runtime advantage.

## F CAPTURED-QUADRATIC ALIGNMENT PROTOCOL AND VERIFICATION

The unchanged source capture contains the first-block QKV, attention-output, MLP-expansion and MLP-contraction projections. We use each first and last output row and the first three stored disjoint matching pairs. Their six coordinates retain the original eight-point affine grids. Every outside coordinate stays at its original stored CD2 value. Selection precedes the new experiment, with all eight cases retained. The continuous-box control is an explicitly later addition to the same cases.

Let $z$ be the stored row, $w$ its captured center, $I$ the six active coordinates and $g_I = H_{I,:}(z - w)$. The complete gradient is recomputed as a sum of exact products of stored binary rationals, including every coordinate outside $I$. For grid midpoint $a_i$ and halfwidth $s_i > 0$, write $q_I = a + Sx$ with $x_i \in \{-1, -5/7, -3/7, -1/7, 1/7, 3/7, 5/7, 1\}$ and $S = \mathrm{diag}(s)$. With $d = a - z_I$, the objective increment is

$$x^\top S H_{II} S x + 2[S(g_I + H_{II} d)]^\top x + 2 g_I^\top d + d^\top H_{II} d.$$

Dividing all coefficients by the largest transformed Hessian diagonal gives the normalized objective used in the figures and table. The scale is stored. This preserves the exact finite problem and its optimum; it does not turn a six-coordinate restriction into an unrestricted row problem.

All $8^6$ points are enumerated with integer coefficients obtained by clearing denominators. An independent meet-in-the-middle implementation compares 512 left and 512 right triples, including all cross terms, and reproduces every optimum exactly. Strict-improvement pair sweeps start from the nearest exact canonical codewords to the stored CD2 values. Ties retain the incumbent. The verifier checks all 64 choices for each of the three pairs at the final point.

The complete-support moment reference supplies numerical proposals only. The retained rational correction makes its primal and dual matrices positive definite while preserving the finite block distributions and their moment equalities. Independent exact elimination verifies positivity, and separate sums reconstruct every mean, second moment and objective. The largest family interval width is below $8.82 \times 10^{-8}$. A numerical solver warning is retained; the proof of the reported interval uses the corrected matrices, not the status string.

For the classical continuous-box control, a numerical convex solve supplies a rationally rounded and clipped feasible point $a$. With $g = Ha - b$, convexity gives the exact tangent lower value $c_0 - a^\top H a - 2\sum_i |g_i|$ on $[-1, 1]^6$, while $F(a)$ gives an upper value for that continuous problem. The verifier reconstructs both. The fixed-split control uses a positive rational $\mu$ with exact $H - \mu I \succ 0$ verification. Its anchor is either the pair-fixed point or the repaired moment mean; the latter uses information from the same dense reference.

At each fixed corrected signed dual, local budgets zero, one, two and eight use the same endpoint seed. Exhaustive pair evaluation independently checks their 96 intervals. Master construction, input acquisition, support scoring and rational correction are additional work; none of the reported branch counts measures deployed speed. No model inference is performed in this captured-quadratic study.

## G FURTHER STOPPING REFINEMENTS

**When a simple split already suffices.** For a supplied split $H = D + R^\top R$ and finite block distributions $\pi$, set $A_p = R_p V_p R_p^\top$. Their minimum covariance-completion cost $\Psi$ gives

$$0 \le v_{\mathrm{mom}} - L \le J_D(\pi) - L + \Psi. \tag{41}$$

If all $A_p$ share an orthonormal eigenbasis, let $a_{pj}$ be the square root of their eigenvalues in direction $j$. Then $\Psi = \sum_j (2 \max_p a_{pj} - \sum_p a_{pj})_+^2$. Rank one is the scalar polygon formula. For general $k$-row coupling, the $k \times k$ tests $(1 - \epsilon) \sum_q A_q - 2A_p \succeq 0$ for some $\epsilon > 0$ suffice for $\Psi = 0$, by classical block coherence (Saunderson et al., 2012). Thus balanced covariance and a small fixed-split gap can rule out substantial further curvature improvement without a dense master. Appendix E proves the reduction, a partial cancellation bound, and explicit balanced/dominant binary families. Compression can improve rounding while destroying covariance balance; retain the distribution witnessing the tighter family upper bound.

**Discreteness has a separate scale.** For diagonal $D \succeq 0$, write $G = R^\top R$, let $h_i$ be the largest adjacent grid gap, and let $r$ be the effective coupling rank. Adjacent mixtures, support compression

and conditional rounding give

$$0 \le F^\star - J_D^\star \le \frac{1}{4} \sum_{j=1}^{\min(d,r)} \left(\|R_i\|^2 h_i^2\right)_{(j)}, \tag{42}$$

Here $R_i$ is column $i$ of $R$, and $(j)$ orders the terms decreasingly. This specializes established convexification and constructive recovery methods (Udell & Boyd, 2016; Dubois-Taine & d'Aspremont, 2026); the quarter coefficient and rank dependence are attained by an explicit even-cardinality grid family. General positive definite blocks and approximate optimization are treated in Appendix C.1. These are fixed-quadratic guarantees. The model experiments below evaluate the original convex pair refinement; the new global and signed results are supported by proofs and exact small-instance checks.

## H LIFETIME OF CONDITIONAL CERTIFICATES

**External updates.** During a sweep, a later pair update changes the objective seen by an earlier pair. The following bound quantifies this effect without claiming that frozen gaps can be summed into a global certificate.

**Proposition 11** (Certificate drift)**.** *Suppose $\widehat{q}_B$ has old conditional gap at most $\Gamma$. Change only external coordinates $R$ by $\Delta_R$, keeping $H, C$ and the pair grids fixed. Set $\eta = H_{BR}\Delta_R$ and choose $q_i^{\mathrm{ext}} = \min \mathcal{Q}_i$ for $\eta_i \ge 0$, and $q_i^{\mathrm{ext}} = \max \mathcal{Q}_i$ otherwise. Then*

$$\Gamma_{\mathrm{new}} \le \Gamma + 2\sum_{i\in B} \eta_i(\widehat{q}_i - q_i^{\mathrm{ext}}) \le \Gamma + 2\sum_{i\in B} |\eta_i| \operatorname{diam}(\mathcal{Q}_i). \tag{43}$$

*Proof.* For any competing $q_B$, the change in its cost difference from $\widehat{q}_B$ is exactly $2(\widehat{q}_B - q_B)^\top \eta$. The old difference is at most $\Gamma$. Maximizing the additional linear term over the finite Cartesian grid gives the first bound; coordinate diameters give the second. □

The first bound uses the incumbent location and is at least as sharp as the diameter bound. A controller could use it to trigger a new pair solve when the tolerated drift is exceeded. Our experiments verify the inequality on exhaustive small cases; they do not evaluate such an adaptive controller.

## I ARITHMETIC AND RECONSTRUCTION TARGETS

**Stored-quadratic certificates.** The preceding results hold in exact arithmetic. Float64 evaluation can round a lower bound upward. Our rational reference interprets input floats as exact binary rationals and performs clipping, ordering, minimization and stopping comparisons exactly, exporting lower bounds downward and upper/gap bounds upward. This certifies the supplied quadratic, not preceding activation or covariance computations. The fast implementation remains a numerical diagnostic. A representation change also matters: FP32 storage can destroy a zero canonical gap. Appendix R bounds this change using pair codeword errors and the full-gradient change from all coordinate casts.

**Target alignment on unchanged grids.** An exact pair optimum may still optimize an unsuitable reconstruction target. We use OJBKQ target interpolation (Wang et al., 2026) and the corrected-center perspective of QEP (Arai & Ichikawa, 2025). Let $X_r$ and $X_f$ be runtime and full-precision activations from the same training tokens, and define

$$T_\mu = ((1-\mu)X_f + \mu X_r)W^\top, \quad F_\mu(Z) = \frac{\|X_r Z^\top - T_\mu\|_F^2}{N} + \rho\|Z - W\|_F^2. \tag{44}$$

With $R_\mu = (1-\mu)(X_r - X_f)W^\top$, Equation (2) holds for $H = X_r^\top X_r/N + \rho I$, $C = R_\mu^\top X_r/N$ and $K = \|R_\mu\|_F^2/N$. For $H \succ 0$, solving $SH = C$ and replacing $W$ by $W_\star = W - S$ completes the square, with additive constant $K - \langle S, C\rangle$. The same routines then apply on grids fitted from $W$. At $\mu = 1$, $C = K = 0$ recovers the runtime objective. Gaps and drift require fixed $H, C$; a new sample or target requires recomputing them.

# J PRESERVED KERNEL AND RECONSTRUCTION STUDIES

We use GPT-2 (Radford et al., 2019) and pinned WikiText-2 training and validation data (Merity et al., 2017). The study is an exploratory mechanism and reconstruction audit on one model. All protocol extensions were fixed before their own runs, after examining the preceding results. The test split remains unused. Source hashes, selected-token hashes, per-context losses and complete case coverage are retained. Appendix L specifies the model revision, sampling, code provenance, cost conventions and numerical checks.

## J.1 DOES THE STOPPING RESULT SURVIVE EXACT CHECKS?

We freeze 32 selected pairs from each of three complete feed-forward expansion layers, at 3 and 4 bits, after two full-residual scalar sweeps. The 128 evenly spaced output rows give 24,576 pair-row instances. Four budgets per precision are evaluated against exhaustive finite-grid optima. Both full-bound and lazy implementations match the full-budget optimum within $10^{-10}$ in all 196,608 numerical method/budget cases. At budgets one and two, approximately 97.6–97.9% and 99.8–99.9% of instances respectively have a reported gap below the fixed numerical threshold. These are numerical diagnostics, not exact arithmetic success rates.

The independent rational replay selects the first and last output row of every frozen pair: 384 instances and 1,536 budget intervals. All intervals are valid for the stored rational problems. Four float64 lower bounds exceed the exact optimum, with maximum excess $2.45 \times 10^{-20}$. The magnitude is immaterial for reconstruction quality, but the sign rules out a blanket machine-certificate claim for ordinary float64 evaluation.

Mean finite scores remain about 1.016, while mean lower-bound evaluations fall from 8 and 16 to about 3. The first Python/PyTorch implementation was slower despite this reduction (Appendix L.4). We therefore compile both bound-search variants and a full conditional-scan baseline with the same single-thread CPU framework, without fast-math transformations. All 221,184 compiled method/budget cases satisfy the stated numerical-oracle, feasibility and count checks.

Table 5: Compiled single-thread CPU search on 12,288 frozen rows per precision. Five warmups per cell and 40 interleaved repetitions; milliseconds include validation and result allocation. Packing and first dispatch are separate.

| Bits | Search | Median (ms) | 10th–90th percentiles (ms) |
|---|---|---|---|
| 3 | All bounds | 4.627 | [4.327, 5.065] |
| 3 | Lazy bounds | 0.910 | [0.841, 1.177] |
| 3 | Conditional scan | 0.858 | [0.752, 1.015] |
| 4 | All bounds | 5.646 | [5.295, 6.246] |
| 4 | Lazy bounds | 0.924 | [0.835, 1.122] |
| 4 | Conditional scan | 1.319 | [1.221, 1.613] |

Table 5 uses five warmups and 40 interleaved repetitions on 12,288-row batches. Lazy search is faster than full bound generation and sorting, but slightly slower than a conditional scan at 3 bits and faster at 4 bits. Packing, first dispatch and full-layer residual updates are outside these warm subproblem timings. The worst case remains $L$ finite scores.

## J.2 HOW MUCH DOES A PAIR ADD BEYOND SCALAR REFINEMENT?

Each single-layer run quantizes one complete `c_fc` matrix from layer 0, 5 or 11, using one of three training draws. Grids are shared across methods within a run: 3 or 4 bits, static groups of 128, damping 0.01. We compare released GPTQ, two or three complete-residual scalar sweeps (CD2/CD3), and CD2 followed by one pair sweep. Correlation-based, adjacent and random pair choices separate the effect of coupling from another pass. Exhaustive $L^2$ and conditional $L$ pair scans give identical selected weights and objectives in all 18 layer/draw/precision comparisons.

Figure 6 shows mean calibration reductions of 9.92% and 9.90% over GPTQ and validation RMSE reductions of 1.67% and 1.60% at 3 and 4 bits. All nine RMSE comparisons improve at each precision. Most of the gain is already obtained by CD3: the pair adds about 0.15% of its remaining objective. Conditional scanning saves factors 8 and 16 in finite scores relative to $L^2$ enumeration, but only a few percent of total refinement wall time. The full residual and the additional scalar sweep are therefore essential controls for assessing the pair contribution.

### J.3 Does lower runtime reconstruction improve the whole model?

We next quantize all 48 transformer projections sequentially and evaluate 250,802 predicted validation tokens in disjoint 1,024-token contexts, including the final partial context. Full-precision perplexity is 31.0403. With symmetric grids and 4,096 training tokens, 3-bit perplexities are 155.30 for GPTQ, 231.53 for CD3 and 183.78 for CD2 plus pairs. At 4 bits they are 40.495, 40.368 and 40.663. Every refinement lowers its own frozen layer objective, but that descent does not impose an ordering on model NLL.

Table 6: All-projection three-bit validation perplexity (lower is better). Methods share the indicated grid rule and nested training sample; each run uses its own sequential activations.

| Training tokens | Grid | GPTQ | CD3 | CD2 + pair |
|---|---|---|---|---|
| 4,096 | Symmetric | 155.30 | 231.53 | 183.78 |
| 4,096 | Asymmetric | 103.15 | 101.15 | 116.62 |
| 16,384 | Symmetric | 145.97 | 130.09 | 168.98 |
| 16,384 | Asymmetric | 93.59 | 100.28 | 102.27 |

To investigate the poor 3-bit outcome, we cross symmetric/asymmetric grids with 4,096/16,384 training tokens. The larger sample contains all original windows. Table 6 retains all four cells. Pair refinement has higher perplexity than matching GPTQ throughout; scalar refinement changes sign. Grid symmetry and calibration size materially change the absolute quality, but do not make pair improvement reliable in these controls. Since preceding projections differ across sequential runs, layer ratios refer to each run's own seed rather than one common objective across methods.

### J.4 Does an aligned target change the outcome?

With asymmetric 3-bit grids and 16,384 training tokens, we test $\mu \in \{1, 0.5, 0\}$ for CD3 and CD2 plus pairs. Only training activations enter the center solve; runtime endpoints reproduce all 48 prior projection hashes and validation NLL. Supplementary Table 9 retains all six outcomes. Pair PPL improves from 102.27 at $\mu = 1$ to 74.83 at $\mu = 0$, versus GPTQ's 93.59. These exploratory comparisons change both the objective and the sequential activation path.

At full reference, the pair's mean runtime-objective ratio to its seed rises to 1.73 while its aligned ratio falls to 0.826. These are complete conditional scans; Appendix S tests budgeted sweeps on all three draws.

Two predetermined further draws and ten scalar sweeps (CD10) give twelve outcomes (Figure 7). Pair-minus-CD10 NLL is $-0.0945$, $-0.0117$ and $-0.0222$; the draw-1 contextual interval includes zero. Pairs improve on matched GPTQ throughout, but CD3 is best on draw 2. CD10 still changes coordinates on its last sweep. Appendix P retains all results and their uncertainty scope.

All nine aligned fixed-grid initializer variants lower local loss yet worsen PPL relative to the old CD10 (Appendix Q). A released-QEP off/on control improves PPL on all three draws while also changing dynamic grids and runtime paths (Appendix T). Local loss alone does not order final model quality.

### J.5 A stronger codebook control

Released QuIP# Hadamard, LDLQ and E8P/RVQ routines provide a representation control, with independently checked packing and FP16 attributes. At three bits and ten reconstruction passes,

mean single-layer calibration error is 0.453 times GPTQ and 0.502 times scalar pair repair. Whole-model tuning is not monotone: ten passes worsen all three zero-pass perplexities. Appendix Z retains every outcome and the complete index/attribute/unquantized-tensor accounting. Forward evaluation uses reconstructed FP32 weights; matched index precision does not establish matched deployment memory or packed inference speed.

# K Detailed comparison of the retained methods

GPTQ uses approximate second-order information for one-shot scalar quantization (Frantar et al., 2023). QuantEase develops cyclic coordinate optimization, including initialization from GPTQ (Behdin et al., 2023); CDQuant uses greedy coordinate descent (Nair & Suggala, 2024). Our full-residual coordinate controls instantiate this established optimization principle. They isolate the incremental effect of pair search rather than treating another scalar sweep as a new quantizer. SchurQuant combines groupwise discrete optimization with grid refitting and explicitly studies reconstruction versus end-model behaviour (Lee et al., 2026). Our fixed grids keep improvements and search gaps on one finite feasible set.

Finite-box least squares and efficient bounded Schnorr–Euchner enumeration have a substantial history (Chang & Han, 2008; Ahn et al., 2017). OJBKQ directly connects finite-box quantization to retained Babai candidates, randomized Klein search and an interpolated reconstruction target (Wang et al., 2026). Seed retention, box bounds and candidate ordering are established tools. Our specific ordering uses the clipped continuous pair bound, charging its evaluations separately from finite scores. The signed extension and global stopping interval use the same finite-oracle budget convention.

QuIP# combines Hadamard incoherence with lattice vector codebooks (Tseng et al., 2024). Its released routines provide a representation control at matched index precision. Pair certificates measure search error on a fixed Cartesian codebook; vector codebooks change that representation. Separate reconstruction and storage accounting distinguishes both effects from packed inference performance.

Quadratic convex reformulation and SDP lower bounds have a substantial history (Billionnet et al., 2008; Park & Boyd, 2018). Nonconvex separable underestimators are also established (Buchheim & Traversi, 2018). We identify the exact finite-block moment target and how local sign restrictions affect it. Constructive convexification using Frank–Wolfe and support compression is existing work (Udell & Boyd, 2016; Dubois-Taine & d'Aspremont, 2026); our grid-specific analysis supplies explicit spacing constants and connects them to budgeted finite oracles. The joint conic representation uses classical Schur complements (Boyd & Vandenberghe, 2004). Verified SDP already enforces equalities and positivity together (Jansson et al., 2008); our endpoint-mixture repair preserves the finite moments, while unfinished separation contributes an explicit gap term.

# L Experiment protocol and provenance

## L.1 Model, data and evaluation units

| | |
|---|---|
| Model | `openai-community/gpt2` |
| Model revision | `607a30d783dfa663caf39e06633721c8d4cfcd7e` |
| Dataset revision | `b08601e04326c79dfdd32d625aee71d232d685c3` |
| GPTQ revision | `2d65066eeb06a5c9ff5184d8cebdf33662c67faf` |
| QuIP# revision | `1d8f873e9a2a8b86b12bb1064c312c5689b77d98` |

The original capture records include file and selected-token SHA-256 digests. No test file is loaded by these protocols. The GPT-2 tokenizer joins corpus text using the frozen scripts. Training sampling selects non-overlapping 256-token windows by index. The initial single-layer study uses 16 windows per draw and draws indexed 0, 1 and 2. Validation for that study uses 16 fixed 256-token windows with seed 123. The three studied expansion matrices have output-by-input shape $3072 \times 768$. Only one matrix is replaced in each single-layer run.

Whole-model evaluation instead covers the complete validation token stream in successive disjoint 1,024-token contexts. Each context predicts all tokens after its first; the final short context is retained.

If context $t$ contains $n_t$ predicted tokens and has mean NLL $\ell_t$, corpus NLL is $\sum_t n_t \ell_t / \sum_t n_t$, and perplexity is its exponential. We do not average context perplexities or discard the short tail. The tokenization has 251,048 tokens and evaluation makes 250,802 predictions. Its FP32 NLL is 3.4352876217. These details determine the reported perplexity convention.

The all-projection quantization order within each of 12 blocks is `attn.c_attn`, `attn.c_proj`, `mlp.c_fc`, `mlp.c_proj`. Activations for later projections come from the already quantized prefix of that run. The full block output is propagated before the next block. Grid and calibration follow-ups use a 64-window sample that includes all original 16 windows; 48 extra windows use seed 100. This is one nested draw, not 64 independent training replications.

For target alignment, each block has an unchanged reference copy evaluated on the full-precision activation path. Corresponding module-input hooks collect $X_f$ and $X_r$ for the same training batches. The reference path advances through original blocks; the runtime path advances through that case's quantized blocks. Original projection weights are restored between cases. The two $\mu = 1$ endpoints are checked against all stored projection index, weight, scale and zero-point hashes from the preceding sensitivity experiment. Evaluation NLL must also agree within $10^{-7}$.

### L.2 Baseline and refinement definitions

GPTQ calls pinned, unmodified upstream source with static groups of 128, block size 128, damping factor 0.01, no activation ordering and no MSE grid fitting. The initial experiment uses symmetric grids; the follow-up adapter adds the upstream asymmetric option and is parity-tested at the original symmetric setting. This is a reconstructed-weight evaluation, not a packed CUDA-kernel run.

Complete scalar refinement maintains $G = (Z - W)H$ (or the corresponding shifted-center gradient). For each coordinate, it minimizes the scalar quadratic on that coordinate's existing grid and updates all gradient columns using the accepted displacement. Pair refinement uses the same complete residual and greedily selected disjoint pairs ordered by normalized Hessian correlation. The search tests every first-coordinate codeword and conditionally minimizes the second; an exhaustive $L^2$ version is retained as a parity oracle. Adjacent and random pair controls keep the objective and codebooks fixed. These controls are not complete reproductions of the QuantEase, CDQuant or SchurQuant paper pipelines.

The QuIP# adapter selects unchanged pure-PyTorch function and class bodies from its pinned source, excluding CUDA imports and inference registration. It uses released randomized Hadamard transforms, LDLQ and E8P/RVQ codebooks. Packed indices and FP16 scale/sign attributes are decoded independently for reconstruction. The two-bit base codebook is checked against exhaustive nearest-point search over all 65,536 entries. Three- and four-bit residual stages are checked by stage-wise decode, without claiming global RVQ nearest neighbours. Zero and the released CLI default ten quantization tuning iterations are reported. Gradient fine-tuning, low-rank correction and CUDA inference are outside this matched single-layer protocol.

### L.3 Certificate selection and arithmetic

The frozen replay selects the first 16 greedy correlation pairs and 16 evenly spread remaining pairs per layer and precision after CD2. It uses 128 evenly spaced output rows of each pair, with all pair coordinates held fixed during its budget sweep. Budgets are $\{1, 2, 4, 8\}$ at three bits and $\{1, 2, 4, 16\}$ at four bits. The small-gap threshold is $10^{-12}$ and the exhaustive numerical-oracle comparison tolerance is $10^{-10}$. The exact reference chooses stored-row positions 0 and 127 of every pair, corresponding to the first and last selected output rows, without outcome-dependent selection.

Each exact input coefficient and grid value is reconstructed as the rational represented by its stored binary float. The reference minimizes finite branches with rational comparisons and returns fraction strings. For a float export, it compares the rounded value with the exact fraction and uses the next representable float in the required direction when needed. Thus exported lower bounds cannot exceed their exact lower values and exported upper/gap values cannot fall below their exact upper values. The rational implementation is deliberately an audit reference, with no claim of useful encoding throughput.

Table 7: Full-budget frozen-pair replay. Counts are means per pair row; time is milliseconds per 128-row CPU batch in a single descriptive pass. Lazy search also performs one center solve per row.

| Bits | Search | Finite scores | Lower bounds | Time (ms) |
|---|---|---|---|---|
| 3 | All bounds | 1.016 | 8.000 | 0.764 |
| 3 | Lazy bounds | 1.016 | 2.995 | 1.279 |
| 4 | All bounds | 1.017 | 16.000 | 0.745 |
| 4 | Lazy bounds | 1.017 | 3.006 | 1.223 |

### L.4 Uncertainty and timing

The nine layer/draw summaries are arithmetic means and min–max ranges of ratios to each run's own GPTQ baseline. They are descriptive: the same model weights and validation windows are reused. Whole-model contextual intervals use a paired circular moving-block bootstrap of eight consecutive validation contexts, 5,000 resamples and seed 62819, with token-weighted aggregation in each resample. They condition on the fixed trained model and calibration draw and do not adjust for exploratory protocol selection.

Experiments use a GeForce RTX 3060 Laptop GPU with 6 GiB memory, PyTorch 2.5.1+cu121 and Transformers 4.57.3. TF32 is disabled. GPU experiment timers synchronize at stage boundaries. Whole-model encoding times include calibration collection and sequential propagation but exclude model loading and final validation. Aligned-target times also include reference moment collection and center solves. They are consequently not a direct speed comparison with the earlier runtime-only endpoint implementation. Single-layer added-search times exclude the shared GPTQ seed. Frozen lazy CPU replay times use one pass without warmup or repetitions, so they are descriptive implementation observations rather than a controlled latency benchmark. No inference throughput is measured.

The compiled follow-up uses Numba with one CPU thread, no parallel JIT and no fast-math option. Both bound-search methods and the complete conditional scan share this framework. Inputs are packed once from the same frozen records. Validation and output allocation are included in every timed call; NPZ decompression, packing and the first JIT/cache dispatch have separate recorded times. Five warmups precede 40 repetitions per precision/method cell, interleaved in a deterministic randomized order (seed 48932). The benchmark runs after this task's model experiments have finished. Reported 10th–90th percentiles describe runtime variation on this machine, not uncertainty over hardware. The kernel returns float64 diagnostics and does not implement outward-rounded certificates or a full-layer encoder.

## M Portable finite-certificate replay

The accompanying artifact contains the eight six-coordinate rational objectives, primal and dual witnesses, pair-fixed points, fixed-split bounds and finite-budget records used in the captured certificate study. Its README gives the execution instructions and tested environment. An independent verifier checks input hashes, enumerates all $8^6$ configurations per objective, and verifies matrix feasibility, moment consistency, pair optimality and every saved budget interval. It imports neither the conic solver nor the proposed pair oracle. The expected verification covers eight cases and $2{,}097{,}152$ finite configurations.

This replay starts from the supplied rational objectives and witnesses. Model activation capture, conic optimization, downstream perplexity and timing are separate procedures documented in the relevant appendices. The artifact supports independent checking of the finite certificates; it does not rerun those procedures.

## N ADDITIONAL ALIGNED-TARGET MEASUREMENTS

Table 8 records the separate aligned and runtime objectives. The mean runtime ratio can exceed one while the aligned loss decreases. Figure 5 places these outcomes alongside the full-precision reference and paired contextual uncertainty.

Table 8: Mean ratios to each projection's own GPTQ seed, with total encoding and calibration time. Objectives and activations differ across runs. Timing includes teacher moment collection and center solves.

| $\mu$ | Refinement | Aligned ratio | Runtime ratio | Time (s) |
|---|---|---|---|---|
| 1 | CD3 | 0.9041 | 0.9041 | 166.1 |
| 1 | CD2 + pair | 0.9023 | 0.9023 | 357.3 |
| 0.5 | CD3 | 0.8622 | 1.1073 | 156.9 |
| 0.5 | CD2 + pair | 0.8588 | 1.0967 | 354.9 |
| 0 | CD3 | 0.8313 | 1.7137 | 158.0 |
| 0 | CD2 + pair | 0.8257 | 1.7275 | 353.9 |

Table 9: Aligned-target all-projection validation. GPTQ perplexity is 93.59. Intervals describe paired contextual variation conditional on this model and calibration draw; they do not adjust for selection across targets.

| $\mu$ | Refinement | PPL | $\Delta$NLL / GPTQ | Contextual 95% interval |
|---|---|---|---|---|
| 1 | CD3 | 100.28 | $+0.0691$ | $[+0.0569, +0.0814]$ |
| 1 | CD2 + pair | 102.27 | $+0.0887$ | $[+0.0719, +0.1061]$ |
| 0.5 | CD3 | 87.00 | $-0.0729$ | $[-0.0846, -0.0608]$ |
| 0.5 | CD2 + pair | 87.22 | $-0.0705$ | $[-0.0829, -0.0578]$ |
| 0 | CD3 | 86.12 | $-0.0832$ | $[-0.0941, -0.0712]$ |
| 0 | CD2 + pair | 74.83 | $-0.2236$ | $[-0.2381, -0.2096]$ |

For CD3, perplexity at $\mu = 1, 0.5, 0$ is 100.28, 87.00, 86.12, respectively. For CD2 plus pairs, perplexity at $\mu = 1, 0.5, 0$ is 102.27, 87.22, 74.83, respectively. Table 9 reports all six cases and their paired contextual intervals. Each layer decreases its own aligned objective. Because target choice also changes the sequential activation path, these model-level differences cannot be attributed solely to a common frozen pair gap. No target is selected as a final configuration from this exploratory comparison.

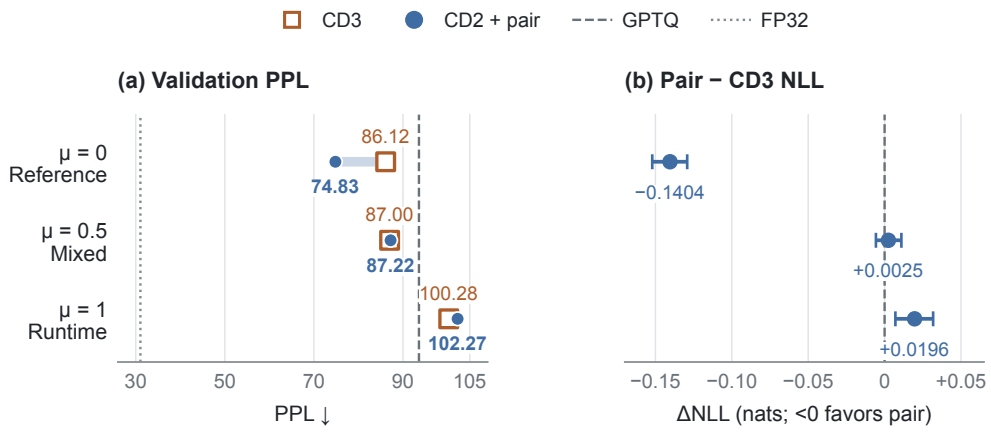



Figure 5: Downstream diagnostic of reconstruction-target choice. Together with Table 8, these results illustrate objective mismatch. (a) Segments connect CD3 and CD2 plus pair at each $\mu$; FP32 and GPTQ are PPL references. (b) Pair-minus-CD3 NLL with paired contextual 95% intervals conditional on this draw and model, without target-selection adjustment.

## O Complete single-layer reconstruction figure

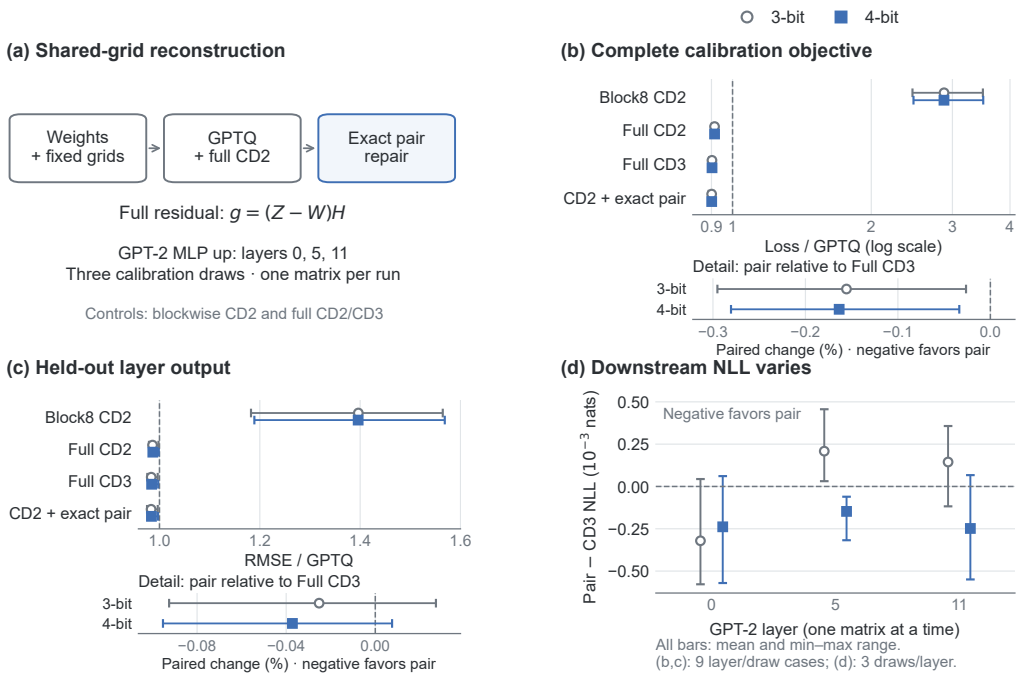



Figure 6: Single-layer reconstruction with shared finite grids. (a) Each run quantizes one matrix. (b,c) Ratios to GPTQ retain all four optimizer controls; detail panels show $100(\text{pair}/\text{CD3} - 1)$ for the same metric. Means and min–max bars cover nine layer/draw cases per precision. Calibration improves over CD3 in every case, while its RMSE advantage is not uniform. (d) Pair-minus-CD3 NLL by layer, with ranges over three draws. These are descriptive ranges, not confidence intervals.

## P CALIBRATION REPLICATION AND STRONGER SCALAR CONTROLS

The full-reference target was selected on the discovery draw. Two further draws use predetermined base/extension seeds 1/101 and 2/102, following the same nested 16-plus-48-window rule as discovery (0/100). Each has 16,384 training tokens. Independent tokenization and selection reconstruction match all three selected-token hashes. The corpus and validation contexts remain shared.

Table 10: All twelve full-reference outcomes. Intervals use paired circular moving blocks of eight validation contexts, 5,000 resamples, seed 62819. They condition on each fitted model/calibration run and do not estimate variability across calibration draws.

| Draw | Method | PPL | ΔNLL / CD10 | Contextual 95% interval |
|---|---|---|---|---|
| Discovery | GPTQ | 93.588 | +0.1291 | [+0.1124, +0.1444] |
| Discovery | CD3 | 86.118 | +0.0459 | [+0.0370, +0.0544] |
| Discovery | CD10 | 82.252 | +0.0000 | [+0.0000, +0.0000] |
| Discovery | CD2 + pair | 74.834 | −0.0945 | [−0.1082, −0.0812] |
| Draw 1 | GPTQ | 92.718 | +0.1621 | [+0.1472, +0.1774] |
| Draw 1 | CD3 | 81.161 | +0.0290 | [+0.0184, +0.0401] |
| Draw 1 | CD10 | 78.841 | +0.0000 | [+0.0000, +0.0000] |
| Draw 1 | CD2 + pair | 77.921 | −0.0117 | [−0.0266, +0.0032] |
| Draw 2 | GPTQ | 96.467 | +0.1133 | [+0.1018, +0.1249] |
| Draw 2 | CD3 | 79.444 | −0.0809 | [−0.0902, −0.0718] |
| Draw 2 | CD10 | 86.137 | +0.0000 | [+0.0000, +0.0000] |
| Draw 2 | CD2 + pair | 84.248 | −0.0222 | [−0.0339, −0.0107] |

CD10 is not assumed converged. The number of coordinates changed on its tenth sweep is 335,940, 336,235, 335,356 for discovery, draw 1 and draw 2. All refinements decrease their own frozen-layer objective; their sequential paths and objectives differ. New GPTQ runs also collect teacher moments and solve an unused aligned center for diagnostics. Their total times are not competitive encoder timings. CPU verification ran concurrently during some cases. One interrupted pair case was restarted from its original weights; only its complete rerun enters this table. Raw records preserve the interruption and all completed results.

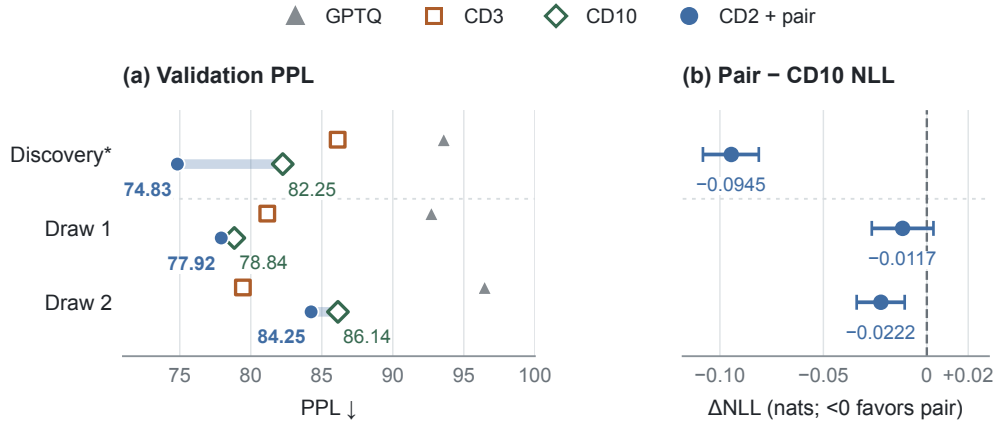



Figure 7: Full-reference replication across calibration draws from one corpus; validation contexts are shared. (a) All four methods and outcomes; segments connect CD10 and pair repair within each draw. The starred discovery draw selected the target. (b) Pair-minus-CD10 NLL and paired contextual 95% intervals, conditional on each fitted model and draw. These are not calibration-sampling intervals. CD10 uses ten fixed scalar sweeps; pair repair uses a full conditional scan.

## Q ALIGNED FIXED-GRID INITIALIZATION CONTROL

This control applies the pinned GPTQ error-propagation routine to the aligned center on original-weight grids. The center is supplied in FP32, the same damped metric is supplied without additional damping, and static asymmetric three-bit group-128 grids remain fixed. The Gram matrix is explicitly symmetrized; its recorded relative asymmetry is zero in all 432 projection records. All three existing training draws and all nine repair variants were prescribed before the experiment.

Table 11: All nine aligned-initializer outcomes. Repair follows fixed-grid GPTQ at the aligned center. NLL differences use the original-initializer CD10 on the same draw; intervals describe paired contextual uncertainty conditional on each fitted model, using the same bootstrap as Table 10.

| Draw | Repair | PPL | ΔNLL / CD10 | Contextual 95% interval |
|---|---|---|---|---|
| Discovery | None | 105.363 | $+0.2476$ | $[+0.2357, +0.2592]$ |
| Discovery | CD3 | 123.665 | $+0.4078$ | $[+0.3927, +0.4226]$ |
| Discovery | CD2 + pair | 117.752 | $+0.3588$ | $[+0.3476, +0.3699]$ |
| Draw 1 | None | 128.177 | $+0.4860$ | $[+0.4665, +0.5063]$ |
| Draw 1 | CD3 | 111.064 | $+0.3427$ | $[+0.3263, +0.3599]$ |
| Draw 1 | CD2 + pair | 104.063 | $+0.2776$ | $[+0.2623, +0.2933]$ |
| Draw 2 | None | 120.153 | $+0.3328$ | $[+0.3188, +0.3471]$ |
| Draw 2 | CD3 | 94.161 | $+0.0891$ | $[+0.0758, +0.1024]$ |
| Draw 2 | CD2 + pair | 120.277 | $+0.3339$ | $[+0.3212, +0.3474]$ |

Within each run's own layer states, the mean ratio of initializer loss to the runtime-target diagnostic seed is 0.762–0.844. Thus the new initializer lowers mean aligned reconstruction loss, but every final model is worse than original-initializer CD10. Repair also lowers its own frozen objective, with variable model-level effects. The loss ratios are within-run comparisons, not a common frozen objective across methods. This is a controlled initializer adaptation, not a replication of QEP. Released QEP uses separately damped, scaled correction and dynamic grids on corrected weights, with a down-projection exclusion. The present control uses the full-reference center, original static grids and no exclusion. Its timing also includes an unused runtime-target GPTQ seed, teacher moments and center solves; CPU tests overlapped some cases. No competitive encoder-latency claim is made.

## R TRANSFERRING A GAP TO THE STORED CODEBOOK

The grid used by a real-arithmetic search and the values stored by a model are different objects. This section bounds their difference for one frozen row objective. It does not cover a new calibration matrix, a changed target, or downstream NLL. Consider

$$F(x) = (x-w)^\top H(x-w) + 2b^\top(x-w), \qquad H = H^\top \succeq 0.$$

For each coordinate let $C_i(k)$ and $S_i(k)$ be finite canonical and stored codebooks indexed by the same set. For FP32 storage, $S_i(k) = \mathrm{float32}(C_i(k))$, interpreted afterwards as an exact real value. Arbitrary same-index perturbations are also allowed. Neither uniform spacing, ordering, nor injectivity of the stored map is required.

Fix selected indices $\widehat{k}_i$ for all coordinates, with $q_i = C_i(\widehat{k}_i)$, $r_i = S_i(\widehat{k}_i)$ and $\rho = r - q$. For a pair $B$, define $A = H_{BB}$ and

$$g_c = [H(q-w)+b]_B, \qquad g_s = [H(r-w)+b]_B, \qquad \eta = g_s - g_c = [H\rho]_B.$$

Thus $\eta$ includes storage changes outside $B$. Suppose the canonical final-state conditional gap already satisfies

$$G_c = F(q) - \min_{k_B} F(q_{-B}, C_B(k_B)) \le \Gamma.$$

Any preceding discrete-update drift must already be included in $\Gamma$. The desired quantity is $G_s = F(r) - \min_{k_B} F(r_{-B}, S_B(k_B))$, with stored external coordinates.

**Proposition 12** (Storage transfer)**.** *For a candidate index pair set $d = C_B(k_B) - q_B$ and $e = S_B(k_B) - r_B - d$. If $T$ bounds*

$$P(d, e) = -2\eta^\top d - 2(g_s + Ad)^\top e \tag{45}$$

*for every candidate, then $G_s \le \Gamma + T$.*

*Proof.* Let $D_c(k)$ and $D_s(k)$ denote the respective objective decreases from $q$ and $r$. Expanding the two complete quadratics gives

$$D_s(k) - D_c(k) = -2\eta^\top d - 2g_s^\top e - 2d^\top Ae - e^\top Ae = P(d, e) - e^\top Ae \le P(d, e),$$

where only $A \succeq 0$ is needed for the last step. Maximizing over the common index set proves the result. Duplicate stored values are harmless: minimization over indices equals minimization over the resulting stored set. The incumbent has $d = e = 0$ in both problems. □

**An explicit local envelope.** For each coordinate in $B$, define

$$\ell_i = \min_k C_i(k) - q_i, \quad u_i = \max_k C_i(k) - q_i,$$
$$R_i = \max(|\ell_i|, |u_i|), \quad \kappa_i = \max_k |S_i(k) - C_i(k) - \rho_i|.$$

The triangle inequality yields the valid transfer bound

$$T_{\text{abs}} = 2|\eta|^\top R + 2|g_s|^\top \kappa + 2R^\top |A|\kappa. \tag{46}$$

If only maximum absolute codeword errors $\epsilon_i = \max_k |S_i(k) - C_i(k)|$ are available, use $\kappa_i \le \epsilon_i + |\rho_i|$. Measuring actual stored values avoids assuming a relative-error model at underflow or exactly representable values.

A tighter bound retains the direction of $\eta$ and bounds the conditional gradient over the displacement rectangle. Set

$$D_\eta = \sum_i \max(-2\eta_i \ell_i, -2\eta_i u_i), \qquad m_i^- = g_{s,i} + \sum_j \min(A_{ij}\ell_j, A_{ij}u_j),$$

$$m_i^+ = g_{s,i} + \sum_j \max(A_{ij}\ell_j, A_{ij}u_j), \qquad M_i = \max(|m_i^-|, |m_i^+|).$$

Then $T_{\text{env}} = D_\eta + 2\sum_i \kappa_i M_i \le T_{\text{abs}}$. Both bounds need only local coefficients and two full-gradient columns after an $O(|K_i| + |K_j|)$ scan of the codebooks.

**A constant-size box relaxation.** Retain signed error endpoints $e_i^- = \min_k(S_i(k) - C_i(k)) - \rho_i$ and $e_i^+ = \max_k(S_i(k) - C_i(k)) - \rho_i$. The function $P$ is affine in each of its four scalar arguments separately. Maximizing one argument at a time therefore shows that

$$T_{16} = \max_{d_i \in \{\ell_i, u_i\},\ e_i \in \{e_i^-, e_i^+\}} P(d, e)$$

bounds every candidate using sixteen evaluations. The box contains zero, so $0 \le T_{16} \le T_{\text{env}}$. Degenerate intervals do not affect the argument. Restoring the negative quadratic term and still checking only corners would be invalid: a concave quadratic can maximize in the interior.

The displacement/error relation of one coordinate can be retained at linear cost. Enumerate its actual $(d_i(k), e_i(k))$ values and maximize $P$ over the four corners of the other coordinate's rectangle. Let the maximum be $T_i$; define $T_j$ symmetrically. Both remain relaxations of all actual candidates, so $T_{\text{lin}} = \min(T_i, T_j)$ is valid and

$$0 \le T_{\text{lin}} \le T_{16} \le T_{\text{env}} \le T_{\text{abs}}, \qquad G_s \le \Gamma + T_{\text{lin}}.$$

This costs $4(|K_i| + |K_j|)$ contrast evaluations. For small grids it can cost as much as or more than Cartesian enumeration; the asymptotic count alone does not establish a faster encoder. Transfer work is additional to the original finite-search budget.

**Why a zero gap needs correction.** Consider a two-dimensional example with $H = I$ and $b = 0$. Let $v = 2^{-23}$, $C_0 = 1 + v/8$, $C_1 = 1 + 5v/8$ and $w = ((C_0 + C_1)/2, 0)$. Select $C_1$ in the first coordinate; let the second coordinate's codebook be the singleton zero. The canonical choices tie. Their FP32 values are 1 and $1 + v$, respectively. Direct substitution gives $G_c = 0$ but $G_s = 2^{-48}$, with $T_{16} = 5\,2^{-49}$ and $T_{\mathrm{lin}} = 2^{-47}$. A small transfer bound can therefore be useful even though it cannot preserve an exact-zero claim.

**Exact and numerical checks.** Independent full-dimensional enumeration with exact fractions checks the transfer inequalities and bound hierarchy, including singular Gram matrices, duplicate stored values, and the tie above. The numerical local implementation and versioned complete-residual sweep have separate finite-oracle tests. The sweep forms the full storage-gradient change once and accounts for visited and unvisited coordinates, while recording transfer work separately. These checks do not constitute model-level timing or quality evidence. Ordinary FP64 evaluation of $\Gamma$, gradients and transfer bounds remains a numerical diagnostic. A machine certificate additionally requires exact or outward-rounded evaluation of those quantities. The transfer preserves conditional scope; summing pair bounds does not certify the global optimum.

## S BUDGETED FULL-MODEL INTEGRATION

This follow-up applies budgets $K_b = 1, 2, 8$ to all 48 projections on each of the three existing training draws. The asymmetric three-bit group-128 grids, full-reference target, runtime-GPTQ seed and two scalar sweeps are fixed by the preceding protocol. All nine cases were prescribed before any budgeted-model outcome. Validation uses the same 250,802 predicted tokens; the test split remains reserved. Each case follows its own sequential activation path.

Table 12: All nine budgeted-model outcomes. Differences use the original-initializer CD10 on the same draw. Contextual intervals follow Table 10; they condition on each fitted model and do not estimate calibration uncertainty.

| Draw | $K_b$ | PPL | $\Delta$NLL / CD10 | Contextual 95% interval |
|---|---|---|---|---|
| Discovery | 1 | 90.734 | $+0.0981$ | $[+0.0858, +0.1102]$ |
| Discovery | 2 | 85.128 | $+0.0344$ | $[+0.0202, +0.0479]$ |
| Discovery | 8 | 78.247 | $-0.0499$ | $[-0.0651, -0.0359]$ |
| Draw 1 | 1 | 81.787 | $+0.0367$ | $[+0.0212, +0.0526]$ |
| Draw 1 | 2 | 74.773 | $-0.0530$ | $[-0.0645, -0.0420]$ |
| Draw 1 | 8 | 75.961 | $-0.0372$ | $[-0.0521, -0.0227]$ |
| Draw 2 | 1 | 86.176 | $+0.0004$ | $[-0.0112, +0.0126]$ |
| Draw 2 | 2 | 73.888 | $-0.1534$ | $[-0.1650, -0.1422]$ |
| Draw 2 | 8 | 82.626 | $-0.0416$ | $[-0.0517, -0.0318]$ |

Perplexity is nonincreasing with budget on 1 of the three draws. A larger frozen-pair budget cannot worsen its retained incumbent in exact arithmetic, but this does not order the final models: changed codewords alter later targets and optimization paths. No budget is selected as a final configuration from these validation outcomes. An independent complete conditional sweep starts from the same canonical CD2 state and matching at all 144 full-budget projections. Its final objective agrees within the prescribed tolerance in 144/144 cases; the maximum recorded relative difference is 0, and the minimum code-agreement fraction is 1.000000. This comparison controls the search endpoint on the same seed. It does not require equality with the historical full-pair runs, which entered the sweep using stored FP32 seeds. The first historical discovery divergence for budget eight occurs at block 0, MLP down-projection; preceding projection weights match. Canonicalization changes the seed representation. Appendix U resolves the first retained-code difference in a captured replay using exact full-row arithmetic.

Table 13: Actual stage accounting, in seconds. Finite and bound counts are means per pair row; one continuous-center solve per row is additional. The independent endpoint replay runs only at budget eight.

| Draw | $K_b$ | Finite | Bounds | Pair (s) | Audit (s) | Replay (s) |
|---|---|---|---|---|---|---|
| Discovery | 1 | 0.987 | 2.948 | 346.9 | 526.8 | 0.0 |
| Discovery | 2 | 1.018 | 2.978 | 344.3 | 527.7 | 0.0 |
| Discovery | 8 | 1.019 | 2.979 | 344.3 | 583.4 | 52.8 |
| Draw 1 | 1 | 0.987 | 2.948 | 345.9 | 532.4 | 0.0 |
| Draw 1 | 2 | 1.018 | 2.978 | 350.3 | 536.2 | 0.0 |
| Draw 1 | 8 | 1.019 | 2.979 | 353.5 | 594.8 | 53.0 |
| Draw 2 | 1 | 0.987 | 2.948 | 344.0 | 532.4 | 0.0 |
| Draw 2 | 2 | 1.018 | 2.978 | 341.0 | 531.9 | 0.0 |
| Draw 2 | 8 | 1.019 | 2.979 | 346.5 | 590.2 | 52.3 |

The serial compiled-CPU/GPU pair stage takes 341.0–353.5 seconds. Its timer includes packing, search transfers, residual updates and final gap diagnostics. The full audit also includes moments, center and seed construction, orchestration, hashing, checkpoint writes and the endpoint replay; restoration, initial embeddings and final validation are outside that timer. CPU verification overlapped parts of this campaign. These instrumented costs do not establish competitive encoder latency. Reducing candidate counts alone has not demonstrated a complete-stage speedup.

Table 14: Conditional numerical diagnostics and measured encoding memory. Means weight every pair row equally. Final zero is the fraction of reported bounds at or below zero tolerance, not an exact-arithmetic certificate.

| Draw | $K_b$ | Mean local gap | Mean final bound | Final zero (%) | Peak MiB |
|---|---|---|---|---|---|
| Discovery | 1 | 6.24e-07 | 6.36e-04 | 0.197 | 1142.3 |
| Discovery | 2 | 3.20e-08 | 6.52e-04 | 0.142 | 1142.3 |
| Discovery | 8 | 0.00e+00 | 6.51e-04 | 0.148 | 1142.3 |
| Draw 1 | 1 | 6.15e-07 | 6.34e-04 | 0.204 | 1142.3 |
| Draw 1 | 2 | 3.27e-08 | 6.47e-04 | 0.141 | 1142.3 |
| Draw 1 | 8 | 0.00e+00 | 6.46e-04 | 0.132 | 1142.3 |
| Draw 2 | 1 | 6.23e-07 | 6.36e-04 | 0.196 | 1142.3 |
| Draw 2 | 2 | 3.15e-08 | 6.46e-04 | 0.140 | 1142.3 |
| Draw 2 | 8 | 0.00e+00 | 6.44e-04 | 0.144 | 1142.3 |

Local gaps refer to the incumbent at its pair update; final bounds add the full external-coordinate drift at the end of the sweep. They are conditional on the other coordinates, cannot be summed into a global gap, and apply to canonical FP64 grid values before the FP32 return cast. Appendix R treats that separate storage change. The saved records contain per-layer means, maxima and counts, so individual-pair quantiles and fresh final-state oracle gaps cannot be reconstructed from hashes alone. Ordinary floating-point diagnostics remain distinct from outward-valid certificates. CUDA allocated peaks include the resident model, calibration buffers and encoding work; counters are reset after initial embeddings and read before validation. Reserved peaks are retained in the raw record. This is neither CPU RSS nor packed inference memory. The explicit search-payload accounting is not a measurement of all device traffic.

## T RELEASED QEP FUNCTION-LEVEL OFF/ON CONTROL

We adapt the pinned author correction and dynamic-grid GPTQ routines of QEP (Arai & Ichikawa, 2025) to the same GPT-2 model, training draws and validation contexts. This distinguishes the fixed-grid center experiment in Appendix Q from correction followed by dynamic grid fitting. The six cases were prescribed before execution. The off condition skips the correction in this same author routine; it is not the earlier static-grid GPTQ baseline. Both conditions quantize all 48 projections

at three bits with asymmetric group-128 grids. The correction uses scale $\alpha = 0.5$ and damping 1.0, separately from GPTQ damping 0.01. In each on run it applies to 36 projections; the 12 MLP down-projections remain quantized but uncorrected. An original block copy supplies teacher activations, and quantized runtime activations propagate separately. Native Conv1D orientation and the moment normalization conversion are checked explicitly.

Table 15: All six completed QEP off/on outcomes. Negative differences favor correction. Intervals use the same paired contextual resampling as Table 10; they condition on the fitted models and do not estimate calibration uncertainty.

| Draw | Off PPL | On PPL | $\Delta$NLL (on $-$ off) | Contextual 95% interval |
|---|---|---|---|---|
| Discovery | 102.583 | 76.740 | $-0.2902$ | $[-0.3027, -0.2779]$ |
| Draw 1 | 108.084 | 81.258 | $-0.2853$ | $[-0.2981, -0.2718]$ |
| Draw 2 | 100.757 | 92.955 | $-0.0806$ | $[-0.0932, -0.0681]$ |

Correction improves validation NLL on all three draws in this function-level adaptation. It also changes fitted grids and downstream runtime states. The comparison therefore supports this combined off/on intervention; it does not isolate a corrected center on a common frozen grid or establish parity with the released end-to-end QEP pipeline. Together with Appendix Q, it shows why the unfavorable fixed-grid initializer result cannot be generalized to QEP.

Table 16: Descriptive local reconstruction and observed audit cost. Local error is the median across 48 projections of teacher MSE per output. Times are seconds; calibration and adapter sums are subintervals of the complete case.

| Draw | Correction | Local MSE | Case (s) | Calibration (s) | Adapter (s) |
|---|---|---|---|---|---|
| Discovery | Off | 0.07395 | 70.78 | 12.14 | 40.84 |
| Discovery | On | 0.05242 | 78.31 | 13.97 | 46.17 |
| Draw 1 | Off | 0.07258 | 75.66 | 14.49 | 42.70 |
| Draw 1 | On | 0.05215 | 91.36 | 15.26 | 55.87 |
| Draw 2 | Off | 0.07272 | 76.16 | 15.05 | 42.31 |
| Draw 2 | On | 0.05111 | 79.97 | 15.12 | 46.18 |

The local statistic evaluates $\|X_r Z^\top - X_f W^\top\|_F^2/(N d_{\text{out}})$ from saved scalar moment diagnostics; runtime states differ between cases. It is not a common-objective causal contrast or a direct feature replay of the full model. Separate tiny Conv1D checks established direct/moment agreement before the model run. All six cases and 288 projection records pass source, calibration, metric, grid-event, correction and exclusion checks. Original tensors are represented by hashes in these records, so their numerical contents cannot be independently replayed from this summary alone. Case times include teacher/runtime collection, diagnostics, provenance, propagation and checkpoint writes; validation is separate. Desktop GPU load and overlapping CPU diagnostics prevent a controlled encoder-latency interpretation.

## U Exact first-code divergence in a captured full sweep

The historical full-pair sweep enters from a promoted stored FP32 CD2 state; the budgeted integration uses its canonical FP64 grid representation. Appendix S verifies equality of complete search endpoints when the seed is held fixed. Here we resolve the first retained-code difference between representations in a saved first-block MLP down-projection ($768 \times 3072$), with the same full quadratic, target and matching.

A read-only observer records all 1,536 pair decisions per output row in the unchanged conditional routine. Removing the observer reproduces its original abstract syntax tree. Each observed CPU and CUDA sweep agrees exactly with its corresponding uninstrumented control and captured GPU endpoint. All nine saved arrays—matching, post-update values, codes, acceptance flags and branches—are bitwise identical between the CPU and CUDA traces. This is observed parity on this captured

sweep, not a general device guarantee. The recorded incremental pair gradients differ in their last bits across devices; equality of the saved decisions does not imply equality of every intermediate numerical gradient. The first differing retained code occurs at zero-based pair ordinal 430, coordinates $(818, 2867)$, output row 722. Both local incumbents have codes $(3, 3)$ and identical values. The alternative $(2, 4)$ has displacement $d = (-0.0546056702733,\ 0.0649074986577)$. At that instant 126 external coordinates differ, including 18 already visited ones. The pair coordinates themselves do not differ.

**Exact local explanation.** Treating the saved binary coefficients as rationals, independently recompute the two full 3,072-term pair gradients and enumerate all $8^2$ feasible points. Write $\Delta F(d; g) = 2g^\top d + d^\top H_{BB} d$ for the objective increment; negative values improve the incumbent. The common alternative has increment $+8.744120876 \times 10^{-12}$ at the canonical state and $-2.576600673 \times 10^{-12}$ at the stored-seed state (rounded from exact fractions). The unique optima are $(3, 3)$ and $(2, 4)$, respectively. Since the local incumbent and candidate coincide across states, only the external residual changes this comparison:

$$\Delta F(d; g_s) - \Delta F(d; g_c) = 2d^\top (g_s - g_c). \tag{47}$$

The exact shift is $-1.132072155 \times 10^{-11}$, which reverses the sign. The stored-path decrease also exceeds the routine's $10^{-15}$ acceptance threshold in magnitude. Every conditional second-coordinate choice is checked against the exact finite enumeration; the smallest recorded distance to a rounding half-integer is 0.09568. Thus this first decision is explained by external-state drift across a conditional decision boundary, and persists without rounded incremental gradients or half-integer tie handling.

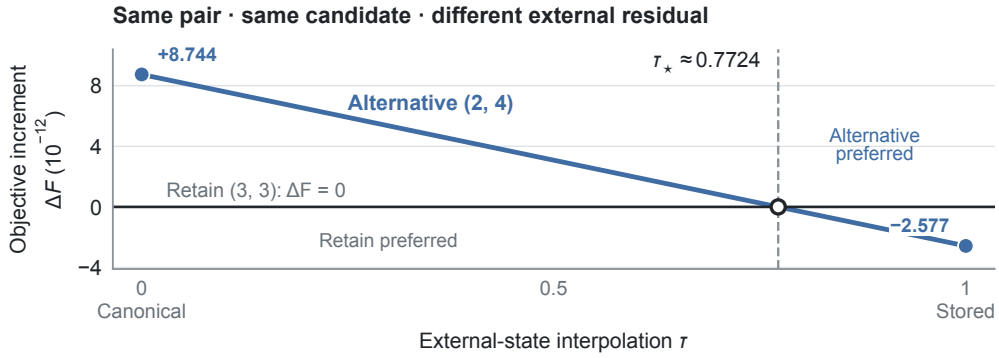



Figure 8: External residual drift can flip a conditional decision for a fixed pair and candidate set. Along an interpolation of the two saved external states, the target and metric remain fixed. The alternative's analytic cost crosses the retained incumbent at $\tau_\star$; negative cost favors the move. Intermediate states are diagnostic perturbations, not asserted feasible quantizer states or executed trajectory samples.

**A quantified boundary on the observed perturbation direction.** Let $z_{-B}(\tau) = (1 - \tau) z^c_{-B} + \tau z^s_{-B}$, holding $z_B$ fixed. Every pair candidate's objective increment is affine in $\tau$. Exhaustive endpoint enumeration verifies that all other candidates have positive increment at the canonical endpoint and nonnegative increment at the stored endpoint. They therefore cannot beat the zero-cost incumbent along the segment. Only the displayed alternative can overtake it. Its exact crossing is $\tau_\star \approx 0.7723996071$ (Figure 8); both codes tie there. The exact-cost crossing of the nominal $-10^{-15}$ acceptance threshold is $\tau \approx 0.7724879407$. This locates sensitivity along one captured perturbation direction; it is not a minimum adversarial radius or a measured machine-trajectory transition. The complete sweep finishes with 68 differing codes across 56 pair-row decisions out of 2,359,296 weights. This captured replay reproduces the historical stored endpoint, but does not recover unsaved historical intermediate tensors. It establishes the first local mechanism in the faithful replay; no model is evaluated here, so it neither quantifies the downstream perplexity difference nor

estimates the frequency of this effect. The example connects the external-update analysis to an observed decision while preserving its conditional scope.

## V PACKED QUIP# FULL-MODEL CONTROL

We extend the released-function QuIP# control to all 48 GPT-2 projections, using three-bit E8P plus residual vector quantization and all three existing training draws. Zero and ten quantization tuning passes are prescribed in advance. The encoder receives original weights and each case's runtime Hessian, regularized by the author routine; teacher features are diagnostic only. The actual deployment weights are independently decoded on CPU from saved packed integer indices and FP16 scale/sign attributes, then copied to the model in FP32. They are not the encoder's pre-storage reconstruction. Gradient fine-tuning, low-rank correction, weight-Hessian rescaling and packed CUDA inference are disabled. This is a function-level adaptation, not a replication of the released full-model pipeline.

Table 17: All six packed-vector model outcomes. Tuning counts quantization passes, not gradient fine-tuning. Differences and contextual intervals compare ten passes with zero on the same draw, conditional on these fitted models.

| Draw | Zero-pass PPL | Ten-pass PPL | $\Delta$NLL $(10-0)$ | Contextual 95% interval |
|---|---|---|---|---|
| Discovery | 83.798 | 105.863 | $+0.2337$ | $[+0.2203, +0.2473]$ |
| Draw 1 | 74.606 | 106.264 | $+0.3537$ | $[+0.3401, +0.3670]$ |
| Draw 2 | 64.680 | 84.413 | $+0.2663$ | $[+0.2544, +0.2785]$ |

Ten passes improve final validation NLL on 0 of the three draws. Later runtime Hessians differ between these sequential runs, so this is not a fixed-Hessian tuning ablation. The historical scalar GPTQ/CD10/pair controls share the model, training draws and evaluation tokens but use different targets, grids and representation costs. Their pairwise contextual contrasts are retained in the summary; no equal-cost or isolated-optimizer comparison is inferred. All outcomes remain reported and no final tuning configuration is selected from these reused validation contexts. As a post-run explanatory diagnostic, block 0 `attn.c_attn` has matching original-weight and Hessian hashes between zero and ten passes on each draw. Its recorded runtime reconstruction loss falls by 6.69–6.83%, and its regularized proxy also decreases, while final NLL worsens. Thus local improvement is present at this shared initial state. Later Hessians differ; the diagnostic neither supplies a common objective across every layer nor isolates the cause of the final NLL change.

Table 18: Observed instrumented costs in seconds. Calibration, reference encoding and checkpoint plus CPU decoding are subintervals of the complete case. Validation and codebook construction are separate.

| Draw | Passes | Calibration | Encoding | Save/decode | Case |
|---|---|---|---|---|---|
| Discovery | 0 | 12.10 | 71.56 | 15.16 | 112.90 |
| Discovery | 10 | 12.61 | 316.71 | 14.80 | 358.80 |
| Draw 1 | 0 | 14.15 | 72.61 | 15.79 | 116.93 |
| Draw 1 | 10 | 12.83 | 316.57 | 14.94 | 358.76 |
| Draw 2 | 0 | 13.89 | 72.95 | 15.01 | 116.35 |
| Draw 2 | 10 | 13.14 | 318.37 | 15.31 | 361.34 |

Table 19: Exact tensor payload counts, in bytes. Shared reference buffers are counted once and separately. Compact tables are an alternative decoder-table count, not an additional implemented inference allocation.

| Component | Bytes |
|---|---|
| Packed indices | 31,850,496 |
| FP16 scale/sign attributes | 294,912 |
| Retained unique FP32 parameters | 158,020,608 |
| Per-model tensor payload, excluding shared buffers | 190,166,016 |
| Shared expanded reference buffers | 2,429,960 |
| Compact author decoder tables (informational) | 9,216 |

The per-model tensor payload is 12.22541 bits per unique model parameter before shared buffers. Serialized per-model checkpoints instead occupy 190,310,442 bytes, including container overhead and the retained-parameter file. Tied embeddings/output weights are counted once. These quantities are distinct from dense FP32 evaluation residency; this runner measures no CPU or CUDA peak memory. All 288 saved projection packets are reloaded and validated on CPU, including exact author repacking, independently decoded weight hashes, and retained/shared tensor accounting. Bitwise decoder replay uses the original four CPU threads; a two-thread replay of the first packet changes FP32 reconstruction values despite identical packed attributes. The corrected versioned validator retains all exact hash checks. A cached decode is reused only after checking identical attribute bytes, shapes and dtypes. NLL and Hessian values are validated recorded metrics/provenance, not a second model evaluation. Case clocks include teacher copies, diagnostics, hashing, propagation and checkpoint writes, with desktop activity and CPU verification during parts of the campaign. They exclude initial restore/embeddings, validation and final case save. These observations do not establish competitive encoder latency or packed inference speed.

## W BATCH ACCEPTANCE ON CAPTURED REAL PROJECTIONS

A shared candidate bank exposes a second optimization question: individually improving pair proposals may interact when accepted together. We compare three batch policies with their best-single reference on each output row. All or none accepts the entire eligible set only if its joint update passes the full-row nonincrease check. Prefix halving shortens a fixed ordered prefix until that check passes. The guarded variant explicitly evaluates the best single proposal and checks its returned canonical row against this reference. These are acceptance controls around the same finite search, not distinct codebooks or additional model evaluations.

The four first-block GPT-2 projection states have output-by-input shapes $2304 \times 768$ (QKV), $768 \times 768$ (attention output), $3072 \times 768$ (MLP up) and $768 \times 3072$ (MLP down). Each starts from the same captured CD2 stored weights, aligned center, damped Hessian, three-bit scale/zero attributes and disjoint correlation matching. Every call canonicalizes the same seed. The 11 prescribed cells per projection include budgets 1, 2 and 8 for sequential lazy and prefix halving, the three other budget-8 batch policies, a full conditional sweep, and one scalar sweep. All 44 full FP32 outputs are saved and independently reloaded on CPU; the complete quadratics are recomputed in NumPy float64. The eight full-budget sequential/conditional endpoints match their captured conditional reference tensors bitwise.

Table 20: All 44 captured-state outcomes: percent reduction in returned FP32 quadratic loss relative to the same stored CD2 seed; larger is better. Budget counts finite candidates per pair-row, not complete-stage work. A scalar sweep has no pair budget.

| Acceptance/search | $K$ | QKV | Attn out | MLP up | MLP down |
|---|---|---|---|---|---|
| Sequential lazy | 1 | 1.0215 | 1.8348 | 3.0919 | 5.6241 |
| Sequential lazy | 2 | 1.1030 | 1.9637 | 3.2884 | 6.1245 |
| Sequential lazy | 8 | 1.1031 | 1.9637 | 3.2888 | 6.1706 |
| Prefix halving | 1 | 0.3856 | 0.4629 | 0.8786 | 0.8254 |
| Prefix halving | 2 | 0.3939 | 0.5098 | 0.8963 | 0.9180 |
| Prefix halving | 8 | 0.3957 | 0.5098 | 0.8964 | 0.9750 |
| Best single | 8 | 0.3614 | 0.3812 | 0.6949 | 0.4456 |
| All or none | 8 | 0.0850 | 0.0828 | 0.1562 | 0.0345 |
| Best-single guard | 8 | 0.5209 | 0.6801 | 1.1474 | 1.1285 |
| Conditional scan | 8 | 1.1031 | 1.9637 | 3.2888 | 6.1706 |
| Scalar sweep | – | 0.9047 | 1.7026 | 2.8509 | 3.1026 |

The distinction between aggregate and row-wise quality matters. At budget 8, prefix halving reduces total loss more than best single in every projection, yet is worse on 887, 329, 1264 and 182 output rows, respectively: 2662 of 6912 rows. All or none is worse on 5934 rows. The guarded policy has no such returned-row regression in these four captures, and also lowers each aggregate objective relative to best single. All 44 outputs have no row worse than their stored seed in this audit. These observations concern the saved FP32 outputs; the independent numerical checks use relative tolerance $10^{-9}$ and absolute tolerance $10^{-12}$ for recorded quadratic agreement. They are not an outward-rounded theorem for arbitrary floating-point executions.

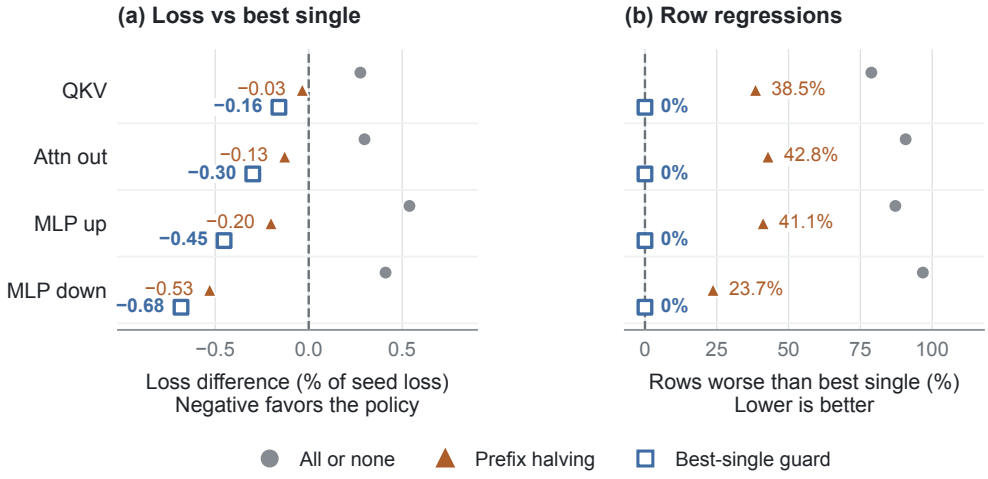



Figure 9: Aggregate improvement can coexist with row-wise regressions relative to best single. At budget 8, prefix halving lowers total loss yet leaves some rows worse than best single. (a) Total returned-loss difference, normalized by stored-seed loss. (b) Fraction of rows with strictly larger returned loss than best single; the guard has zero such rows in these four captures. These are deterministic contrasts for complete captured projections. Table 20 retains all budgets and sequential controls, which recover more loss reduction than any one-shot batch policy here.

**Why nonincrease is too weak.** An exact four-coordinate example isolates the issue without floating-point effects. Let the seed be zero on binary grids, with pairs $(1, 2)$ and $(3, 4)$, gradient $g = (-1, 0, -1, 0)$ and Hessian equal to the identity except $H_{13} = H_{31} = 1 - \varepsilon/2$, where $0 < \varepsilon < 2$. The Hessian is positive definite. The optimal frozen proposal for each pair changes its first coordinate from zero to one, lowering the quadratic by one. Applying both proposals lowers

it by only $\varepsilon$, since their joint increment is $-2 + 2H_{13} = -\varepsilon$. A joint nonincrease check therefore accepts a gain arbitrarily smaller than the best single gain. Both local searches can be exact while this acceptance failure remains. The guarded policy explicitly retains the best-single reference; it does not derive that comparison from nonincrease alone.

**Separate candidate selection from acceptance.** We additionally freeze the original budget-8 candidate bank and its original eligibility mask, then visit pairs in the same fixed order. Each visit checks the existing candidate against the updated full gradient and accepts only if its increment is below $-10^{-15}$, matching the conditional sweep threshold. It neither searches a new candidate nor activates a previously ineligible proposal. The original best-single FP32 endpoint and mathematical audit are reproduced before this intervention. An independent NumPy implementation replays all acceptance decisions and reproduces both the FP64 and FP32 endpoints bitwise on all four projections.

Table 21: Ordered acceptance of a frozen candidate bank. Eligible/accepted counts refer to pair-rows; the last column counts output rows worse than best single. Ordered and fresh columns give percent loss reduction from the stored seed. Fresh denotes the complete conditional sweep.

| Projection | Eligible | Accepted | Ordered (%) | Fresh (%) | Worse rows |
|---|---|---|---|---|---|
| QKV | 15,467 | 6,919 | 0.5932 | 1.1031 | 517 |
| Attn out | 10,176 | 4,094 | 0.9913 | 1.9637 | 67 |
| MLP up | 34,743 | 13,624 | 1.5942 | 3.2888 | 428 |
| MLP down | 69,207 | 18,187 | 2.7899 | 6.1706 | 2 |

Ordered acceptance improves aggregate loss over each one-shot batch policy in these captures, but still recovers less than fresh conditional search. It is also worse than best single on 1014 output rows, so ordered greedy acceptance does not supply the guard's row-wise reference property. Candidate staleness and acceptance policy are therefore distinct interventions worth testing. Different accepted updates alter subsequent states; these aggregate orderings are observed outcomes, not a theorem that a fresh greedy sweep must dominate every frozen-candidate sweep.

The quality audit is separate from the prescribed isolated full-stage timing benchmark. Desktop GPU activity was allowed for these correctness/quality calls; internal timing fields remain raw diagnostics and are not aggregated into speed evidence. No model is evaluated from these outputs, and no language-model improvement, peak-memory saving or encoder speedup follows from the plot. Four first-block states do not establish coverage across layers, calibration draws, models or codebook families.

## X CANDIDATE REFRESH AND TRAJECTORY DEPENDENCE

The previous controls act on one candidate bank. We next regenerate that bank after each accepted state while preserving the aligned center, damped Hessian, finite grids and ordered matching. Each of the four captured projections receives exactly 16 stages of Guard 1, Guard 2, Guard 8 and Prefix 8, for 256 actual stage outputs. The suffix is the finite-score budget per pair-row per stage. All trajectories start at the same stored CD2 tensor; subsequent stages receive the preceding canonical FP64 output. The saved FP32 casts are diagnostics and are not fed into the next stage. There is no early stopping, best-stage selection or outcome-based extension.

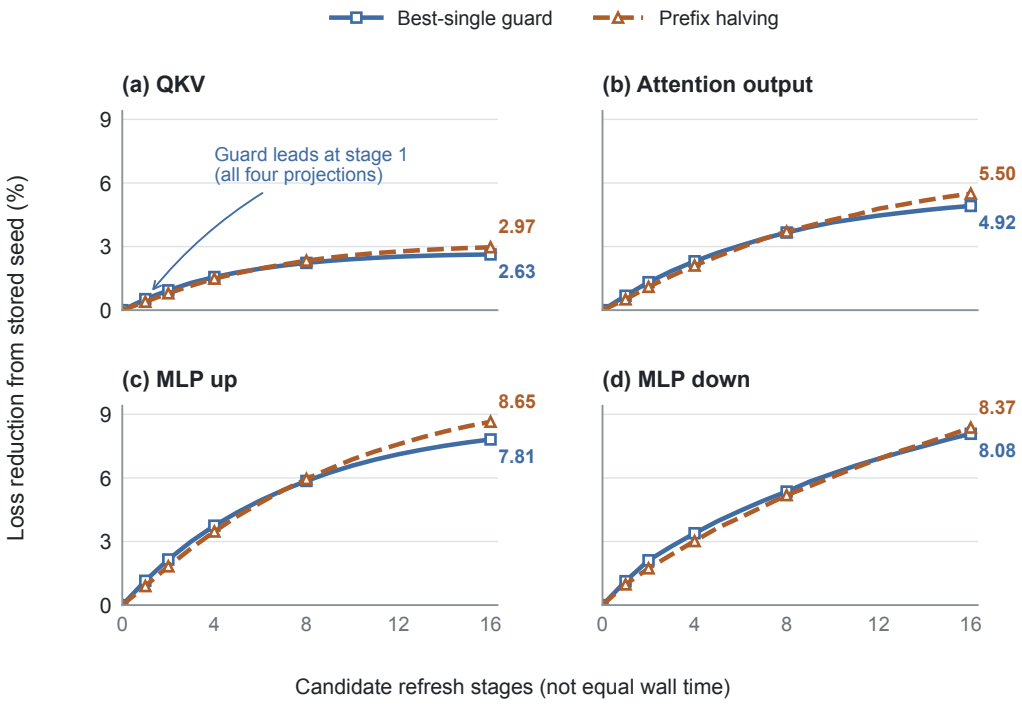



Figure 10: Acceptance rules generate different future candidate banks, so a one-step guarantee does not establish trajectory dominance. Budget-eight traces cover all four captured projections. Every measured stage is shown; markers highlight stages 1, 2, 4, 8 and 16, with the shared seed at stage zero. The vertical axis is reduction in the full quadratic of the diagnostic FP32 cast, relative to that seed. The guard initially improves more, but prefix halving has lower aggregate loss at stage 16 in all four cases. These are deterministic captured-state traces, not model-NLL measurements or a comparison at equal elapsed time.

**Complete prescribed policy coverage.** Table 22 retains all four policies at the five reporting stages prescribed before execution. The released raw records additionally retain every intervening state, its code changes, bounds, objectives and full saved tensors. The two budget-eight trajectories in Figure 10 isolate the acceptance-rule question; the figure does not select the strongest budget for either rule.

Table 22: Fixed-horizon refresh controls on all four projections. Entries under stage count are percentage reductions from the common stored-seed quadratic. The final column is the actual cumulative number of finite-grid scores after 16 stages, in millions; it excludes center solves, bound scores, acceptance checks, diagnostics and memory traffic. All policies perform 16 complete stages.

| Projection | Policy | 1 | 2 | 4 | 8 | 16 | Scores (M) |
|---|---|---|---|---|---|---|---|
| QKV | Guard 1 | 0.510 | 0.906 | 1.484 | 2.081 | 2.346 | 13.950 |
| | Guard 2 | 0.522 | 0.936 | 1.561 | 2.222 | 2.616 | 14.263 |
| | Guard 8 | 0.521 | 0.936 | 1.563 | 2.229 | 2.628 | 14.279 |
| | Prefix 8 | 0.396 | 0.795 | 1.485 | 2.334 | 2.967 | 14.315 |
| Attn out | Guard 1 | 0.666 | 1.292 | 2.258 | 3.593 | 4.766 | 4.652 |
| | Guard 2 | 0.680 | 1.320 | 2.305 | 3.659 | 4.914 | 4.737 |
| | Guard 8 | 0.680 | 1.320 | 2.306 | 3.661 | 4.917 | 4.738 |
| | Prefix 8 | 0.510 | 1.100 | 2.105 | 3.706 | 5.496 | 4.771 |
| MLP up | Guard 1 | 1.129 | 2.118 | 3.692 | 5.710 | 7.435 | 18.604 |
| | Guard 2 | 1.147 | 2.153 | 3.748 | 5.858 | 7.815 | 18.892 |
| | Guard 8 | 1.147 | 2.153 | 3.748 | 5.857 | 7.814 | 18.893 |
| | Prefix 8 | 0.896 | 1.829 | 3.471 | 5.951 | 8.648 | 18.969 |
| MLP down | Guard 1 | 1.062 | 2.063 | 3.315 | 5.263 | 7.837 | 18.574 |
| | Guard 2 | 1.078 | 2.053 | 3.323 | 5.288 | 8.109 | 19.847 |
| | Guard 8 | 1.128 | 2.114 | 3.390 | 5.355 | 8.082 | 19.902 |
| | Prefix 8 | 0.975 | 1.735 | 3.019 | 5.186 | 8.373 | 20.785 |

**What refresh resolves and what it does not.** All trajectories continue to reduce the observed aggregate loss. However, the larger search budget does not impose a trajectory ordering: Guard 2 slightly exceeds Guard 8 at stage 16 on both MLP projections. The best-single guard compares candidates from its *current* bank. Once acceptance rules choose different states, their refreshed proposals and cross-pair interactions also differ. Its local reference property therefore does not imply that it dominates prefix halving after repeated stages. These observations motivate studying the interaction between acceptance and candidate regeneration, rather than treating one-step improvement or a larger budget as sufficient evidence of final-model benefit.

**Validation and cost scope.** Independent CPU reconstruction verifies all 256 complete FP64 outputs, their FP32 casts, state chains and full quadratics. No strict numerical canonical row increase is observed in these replays. All eight applicable stage-one outputs match the earlier single-stage endpoints. Sixteen tiny exact-Fraction transitions were checked on CPU and actual CUDA before the large runs. Per-stage best-single baseline tensors were not saved: their recorded row inequalities are checked for internal consistency, while the reused guard kernel has separate finite-oracle tests. This is not an independent re-execution of every candidate search or a formal certificate for floating arithmetic.

Each trajectory uses $16mn/2$ pair-row center solves for an $m \times n$ projection, in addition to the finite scores in the table and recorded bound/acceptance work. Equal stage count and nominal budget do not imply equal total work or wall time. Sixteen refreshed stages also spend more search effort than the earlier one-sweep sequential control, so their larger loss reductions do not establish a superior compute-matched optimizer. Final ordinary-FP64 conditional-gap upper diagnostics remain positive for every trajectory; the maximum ranges from approximately $0.0181$ to $0.9599$ across the 16 endpoints. No convergence or outward-rounded optimality claim follows from this fixed horizon.

**An exact two-stage reversal.** The trajectory distinction does not require finite-score truncation, floating-point error or a singular metric. Consider the binary grid $\{0,1\}^4$, the zero seed, the fixed pairs $(1,2)$ and $(3,4)$, and

$$\Phi(z) = z^\top H z + 2g^\top z, \qquad H = \begin{pmatrix} 16 & -11 & -6 & 1 \\ -11 & 18 & 9 & 5 \\ -6 & 9 & 15 & 10 \\ 1 & 5 & 10 & 24 \end{pmatrix}, \qquad g = \begin{pmatrix} -25 \\ -5 \\ -10 \\ -2 \end{pmatrix}.$$

The leading principal minors of $H$ are $16$, $167$, $1749$ and $26497$, so $H$ is positive definite. Completing the square with $w = -H^{-1}g$ expresses the same decisions as minimizing $(z-w)^\top H(z-w)$; the omitted additive constant changes no comparison. At each stage, solve both pair subproblems exactly at the current state, retain strictly improving proposals and order them by their individual decreases. Prefix halving tests the full proposal set against the current objective, halving the prefix after rejection. The guard applies the same test against the best-single reference instead.

> **Claim.** The guard makes a strictly larger first-stage improvement but stays at a nonglobal fixed point. Prefix halving reaches the unique global minimum in two stages.

*Verification.* At $0000$, the unique pair minimizers give $1100$ with value $-48$ and $0010$ with value $-5$. Their joint proposal is $1110$, with value $-47$. Prefix halving accepts it because $-47 < 0$. The guard rejects it because $-47 > -48$ and instead accepts $1100$. At $1100$, both exact pair minimizers return the current state; thus every later guard stage is a no-op. At $1110$, the two unique pair proposals are $1010$ with value $-51$ and $1100$ with value $-48$. Their joint proposal, $1000$, has value $-34$ and is rejected. Halving leaves the stronger proposal, $1010$, which is accepted. The complete enumeration below establishes that it is the unique global minimum. Hence the first-stage ordering $-48 < -47$ reverses to $-51 < -48$ after two stages. □

Table 23: Every feasible point of the exact trajectory witness. Binary strings list coordinates in order; all displayed objectives are integers.

| $z$ | $\Phi(z)$ | $z$ | $\Phi(z)$ |
|---|---|---|---|
| 0000 | 0 | 1000 | −34 |
| 0001 | 20 | 1001 | −12 |
| 0010 | −5 | 1010 | −51 |
| 0011 | 35 | 1011 | −9 |
| 0100 | 8 | 1100 | −48 |
| 0101 | 38 | 1101 | −16 |
| 0110 | 21 | 1110 | −47 |
| 0111 | 71 | 1111 | 5 |

All visited pair minima are unique, and both policies have full local search budgets. The guard's fixed-point additive gap is three; scaling $(H, g)$ by any positive constant scales that gap while preserving every decision. This observation supplies no scale-independent additive guarantee from the current-bank guard alone, and is not a multiplicative approximation lower bound. It also does not imply that prefix halving usually wins. The witness was found after the captured-state observations by an exploratory integer search, then checked independently by exact enumeration and against four transitions of the unchanged CPU production kernels. Its role is to disprove a trajectory implication, not to estimate its prevalence.

## Y COMPLETE CONVEX PAIR PROCEDURE

Table 24: Lazy pair search. The score budget excludes the known feasible incumbent; center and lower-bound work are accounted separately.

Input: fixed pair grids, coefficients $(a, b, c, p, q)$, budget $K_b$.
Set $U = 0$ and retain the current pair; compute $\bar{x}$ from Equation (6).
Initialize the two frontier heads around $\bar{x}$; set score count $k = 0$.
While $k < K_b$ and the lesser head bound $r < U$:
  Consume head $x$; compute $y_f(x)$ and $v(x)$; increment $k$.
  If $v(x) < U$, retain $(x, y_f(x))$ and set $U = v(x)$.
  Advance the consumed frontier and compute its next bound, if present.
Return the retained pair and $[\min(U, r), U]$, with $r = +\infty$ at exhaustion.

## Z CODEBOOK CONTROLS AND COMPLETE COST ACCOUNTING

We adapt released QuIP# Hadamard, LDLQ and E8P/RVQ routines, with independently checked packing and stored FP16 attributes. Twelve single-layer runs cross three layers, 3/4 bits and zero/ten quantization tuning passes. Gradient fine-tuning and packed CUDA inference are disabled.

Table 25: Matched QuIP# single-layer controls, averaged across layers 0, 5 and 11. Entries are ratios, not perplexities. Iterations are quantization tuning passes; gradient fine-tuning is disabled.

| Bits | Iterations | Cal. / GPTQ | Val. RMSE / GPTQ | Cal. / pair |
|---|---|---|---|---|
| 3 | 0 | 0.480 | 0.685 | 0.532 |
| 3 | 10 | 0.453 | 0.688 | 0.502 |
| 4 | 0 | 0.717 | 0.839 | 0.795 |
| 4 | 10 | 0.780 | 0.927 | 0.865 |

At 3 bits and ten passes, mean calibration error is 0.453 times GPTQ and 0.502 times scalar pair repair. Four-bit tuning is not monotone; all twelve outcomes remain in Table 25. Appendix V extends the control to all 48 projections. Across the three draws, zero-pass PPL is 83.80/74.61/64.68; ten passes worsen it to 105.86/106.26/84.41. These controls motivate evaluating representation, optimization and sequential targets jointly; more reconstruction tuning does not uniformly improve quality.

Matched index precision is not matched deployment cost. The vector index payload is exactly 3 or 4 bits per weight; matrix scale/sign attributes add 0.02604 bits per weight in the single-layer control, with shared lookup storage additional. Scalar group-128 FP32 scales alone add 0.25 bits per weight. For the initial symmetric scalar runs, only 68.25% of unique model parameters are quantized; embeddings, the tied output head, norms and biases remain FP32. The resulting nominal whole-model totals are 12.38 and 13.06 bits per parameter under the recorded accounting convention. They are not serialized checkpoint sizes or measured inference memory. Forward evaluation uses reconstructed FP32 weights, so no packed inference speedup is inferred.

Warm subproblem timing excludes substantial fixed costs. Archive decompression and packing take 62.5/54.2 ms for the three-/four-bit batches, and the first JIT/cache dispatch takes 262.6 ms. An integrated encoder must also maintain full residuals and move coefficients between memory spaces. These measurements support removing the bound table when coefficients are available, but do not establish a speedup for offloading live GPU reconstruction to the CPU. Resident-data integration must be measured at the encoder boundary.