COMPUTABLE SUPPORT AND CONVERGENCE CERTIFICATES FOR CYCLIC COORDINATE DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

When may a sparse optimizer regard its support as final and report a parameter-error bound? We give a finite-candidate test based on uniform curvature of a sign-preserving active neighborhood and strict inactive-gradient margins. The test proves the complete support and signs, uniqueness of the global minimizer, a solution enclosure, and an invariant region with an exact-cycle rate. It requires neither global strong convexity nor absolute diagonal dominance. A rank-deficient quadratic separates this primal test from every standard Gap Safe sphere at the same candidate, regardless of the feasible dual center. We also obtain a descent-safeguarded inexact-update guarantee and an exact-arithmetic certificate test for rational squared loss. Experiments on 30 non-ridge, high-dimensional objectives compare parameter-error decisions with screening followed by the same face analysis. Dual extrapolation reduces the primal test's earlier decisions from 28 to 17, and screened routes remain cheaper for the requested parameter accuracy. The analytic cycle bounds are conservative in the experiments. On 217 compact rational inputs, the test accepts 37 candidates, and exact rational trajectories satisfy the safeguard and cycle budgets.

1 INTRODUCTION

Sparse empirical risk minimization couples support selection with numerical optimization. A coordinate update can set a coefficient exactly to zero and immediately expose that value to the next update, a pattern used in regularization-path solvers (Friedman et al., 2010). But neither a stable support nor a small stopping residual proves that the nonzero coefficients have their final signs. For a solver that will discard coordinates, reuse a local model, or report a convergence bound, the relevant question is what follows from the point already computed.

We consider

$$F(x) = h(x) + \lambda\|x\|_1, \qquad x \in \mathbb{R}^n, \quad \lambda > 0, \tag{1}$$

with h convex and continuously differentiable. The main applications are sparse logistic and least-squares objectives, including fixed nonlinear feature maps with more features than observations. No ridge term is required. Penalized Huber regression is also covered through uniform almost-everywhere curvature bounds. The desired output is a *joint decision*: the candidate support and signs equal those of the unique global minimizer, the minimizer lies in a specified neighborhood, and subsequent *exact* coordinate updates remain in a region with an explicit convergence rate. The enclosure also supports a requested parameter-error threshold. Neither the support statement nor that threshold establishes statistical variable-selection validity.

Our principal test combines three quantities available at a candidate: its active first-order residual, a positive lower bound on curvature throughout a sign-preserving ball, and a strict bound on every inactive gradient throughout that ball. A boundary argument encloses the solution and traps the candidate's sublevel set. This route permits signed cancellations in the active Hessian even when the usual nonnegative influence matrix fails to contract. A stronger coordinate-box test provides componentwise error bounds when its additional contraction condition holds. Figure 1 separates these finite-neighborhood conclusions from the signed Jacobian used only to predict asymptotic behavior.

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.

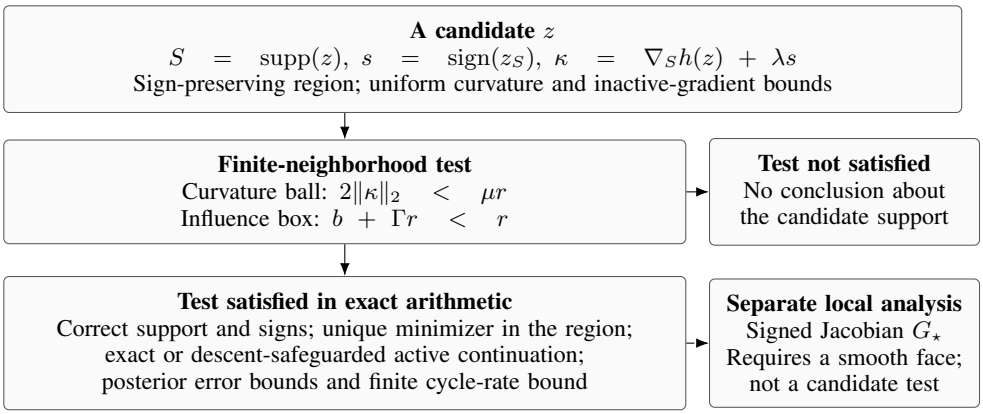


Figure 1: The finite decision and its limits. Exact inequalities give the indicated mathematical conclusions. General binary64 evaluations are numerical witnesses; the rational squared-loss backend checks the conditions exactly. A failed test is inconclusive. The signed local matrix belongs to a separate, asymptotic analysis.

Contributions.

- **C1. A finite support-and-error decision.** Theorem 2 gives a primal curvature test, a solution enclosure, an invariant sublevel set, and an objective-gap rate. Theorem 4 adds componentwise bounds under a stronger hypothesis. Neither test uses an unknown optimum.
- **C2. A precise separation from spherical screening.** Proposition 3 exhibits a rank-deficient lasso for which the primal test succeeds and no feasible dual center makes the standard Gap Safe sphere decisive at that candidate. This is a limitation of that spherical family, not of all dual certificates.
- **C3. Implementable guarantees with an arithmetic boundary.** We derive uniform GLM bounds, a support-preserving inexact-update rule, and an exact rational certificate test for squared loss. Dense active-matrix costs and the additional requirements of validated arithmetic are explicit.
- **C4. Decision-level evaluation.** Matched trajectories compare residual and extrapolated dual points. Cost-complete runs measure the time to a requested parameter-error bound, not just a small objective gap. The comparisons include failed precision requests, conservative cycle budgets, and faster screening baselines.

2 Relation to existing theory

The triangular cycle operator is classical Gauss–Seidel machinery (Varga, 2000), and exact coordinate minimization already has substantial theory. Nonasymptotic comparisons under isotonicity appear in Saha & Tewari (2013); exact block-minimization complexity without per-block strong convexity in Hong et al. (2017); and cyclic or alternating-minimization rates in Beck & Tetruashvili (2013); Sun & Hong (2015); Li et al. (2018). The exact-update question is motivated by Diakonikolas (2025), but do not claim that previous convergence theory is absent or that our conditional rate uniformly improves these results.

Klopfenstein et al. (2024) establish finite identification and local linear convergence of proximal-coordinate methods. The distinction here is a *test at a finite candidate*, not priority for identification or local linear rates. Gap Safe screening (Ndiaye et al., 2017) can safely remove variables and can be combined with additional information about a restricted problem. Accordingly, our screening-plus-face baselines first certify every proposed zero and then apply the *same* active curvature, residual, and sign test. Each baseline also returns a complete support-and-rate conclusion for the reduced solver (Appendix I). We also use improved feasible dual points from prediction extrapolation (Massias et al., 2020), the dual component underlying Celer (Massias et al., 2018). The comparator

Table 1: Positioning of the finite-candidate task. Existing convergence and screening guarantees are not claimed as new.

Line of work	Information supplied	Additional requirement here
Varga (2000)	Splitting-based convergence	A solution-containing candidate region
Hong et al. (2017); Li et al. (2018)	Global exact/cyclic complexity	A posteriori support and local bounds
Klopfenstein et al. (2024)	Identification and local rates	A finite test not using $x^\star$
Ndiaye et al. (2017); Massias et al. (2020)	Safe elimination and improved dual regions	Same active-face test; compare the full decision
This work	Primal joint decision	Uniform local curvature and margins; dense active-matrix cost

retains the best dual objective among the residual, extrapolated, and previous points. It is not the complete Celer working-set solver.

Table 1 summarizes these distinctions. The difference is the source of the enclosure. Our theorem requires a uniform primal inactive-gradient margin and proves invariance even for the original full-coordinate method. The screening baseline instead requires a suitable dual region and invariance only after the screened coordinates are removed. Neither sufficient test dominates the other. The cost of our stronger neighborhood information includes active matrix construction and curvature bounds, not just residual evaluation. As in validated nonlinear computation (Rump, 2010), an exact theorem and a floating-point implementation are different objects; the large learning experiments report *numerical sufficient-condition witnesses*. For rational squared-loss problems, the certificate conditions are evaluated exactly.

3 Cycle comparisons and their limitation

Notation and update maps. For a vector z, let $z^{(i\leftarrow u)}$ replace its ith coordinate by u. Define the *scalar* minimizer $t_i(z_{-i}) = \arg\min_u F(z^{(i\leftarrow u)})$ and the *vector* update $\mathcal{T}_i(z) = z^{(i\leftarrow t_i(z_{-i}))}$. Whenever used, the minimizer is assumed unique on the stated domain. The full cycle is $\mathcal{P} = \mathcal{T}_n \circ \cdots \circ \mathcal{T}_1$. Write $x^{k,0} = x^k$, $x^{k,i} = \mathcal{T}_i(x^{k,i-1})$, and $x^{k+1} = x^{k,n}$. All vector inequalities below are componentwise. The symbol $\succeq$ is reserved for the Loewner order on symmetric matrices. For $w > 0$, $\|v\|_{\infty,w} = \max_i |v_i|/w_i$ and $\|B\|_{\infty,w}$ is its induced matrix norm. The spectral radius is $\rho(B)$.

A nonnegative, zero-diagonal matrix $\Gamma = (\alpha_{ij})$ is a minimizer-influence bound when

$$|t_i(z_{-i}) - t_i(y_{-i})| \le \sum_{j\neq i} \alpha_{ij}|z_j - y_j|. \tag{2}$$

An edge $j \to i$ therefore represents influence *on* coordinate i from coordinate j. Let $\Gamma = L + U$ be its strict lower/upper split in the chosen order. If $x^\star$ is a minimizer and (2) holds for its comparisons with every within-cycle iterate, then, with $e_i^k = |x_i^k - x_i^\star|$,

$$e^{k+1} \le Le^{k+1} + Ue^k, \qquad e^{k+1} \le Ke^k, \qquad K = (I-L)^{-1}U. \tag{3}$$

The lower term uses already updated coordinates. Forward substitution resolves this dependence because $(I-L)^{-1} = I + L + \cdots + L^{n-1} \ge 0$. This is a comparison envelope, not generally the signed derivative of the cycle.

Obtaining and using the coefficients. Suppose each scalar subproblem is μ_i-strongly convex and, for points with $x_i = y_i$,

$$|\nabla_i h(x) - \nabla_i h(y)| \le \sum_{j\neq i} \beta_{ij}|x_j - y_j|. \tag{4}$$

Strong monotonicity of the scalar optimality condition gives $\alpha_{ij} = \beta_{ij}/\mu_i$; Appendix B proves this reduction. The induced weighted norm can be made arbitrarily close to $\rho(K)$ by a positive resolvent

weight. Explicit formulas and the row-mass certificate are in Appendix B. The key limitation is order-independent.

Proposition 1 (Ordering cannot repair failed absolute contraction)**.** *For every nonnegative zero-diagonal Γ and every permutation π,*

$$\rho(K_\pi) < 1 \quad \Longleftrightarrow \quad \rho(\Gamma) < 1. \tag{5}$$

The value of $\rho(K_\pi)$ can depend on π, but existence of this contraction certificate does not.

This classical splitting consequence rules out repairing an intrinsically noncontracting nonnegative bound merely by reordering. The curvature certificate below avoids that condition. For sharpness, symmetric $\Gamma \geq 0$ with $\rho(\Gamma) < 1$ gives the convex quadratic $F(x) = x^\top(I - \Gamma)x/2$. Exact cyclic minimization satisfies $x^+ = Kx$, and the absolute comparison is equality for $x \geq 0$. This exact vector identity does *not* make every finite-step norm ratio equal to $\rho(K)$.

4 Certificates centered at a numerical candidate

Let z be any candidate for (1), $S = \operatorname{supp}(z)$, $Z = [n] \setminus S$, and $s = \operatorname{sign}(z_S)$. Define

$$\kappa = \nabla_S h(z) + \lambda s. \tag{6}$$

This is the residual of the smooth objective restricted to the candidate face. It is computable even though $x^\star$ is unknown. We assume $S \neq \varnothing$; if S is empty, $\|\nabla h(0)\|_\infty < \lambda$ already proves that zero is the unique minimizer.

4.1 A curvature certificate without absolute contraction

For $0 < r < \min_{i \in S} |z_i|$, define the sign-preserving ball

$$\mathcal{B} = \{x : x_Z = 0,\ \|x_S - z_S\|_2 \leq r\}. \tag{7}$$

All bounds in the next result must hold *throughout* $\mathcal{B}$, not only at z.

Theorem 2 (Curvature certificate)**.** *Assume that h restricted to $\mathcal{B}$ is μ-strongly convex, with $\mu > 0$, and*

$$\sup_{x \in \mathcal{B}} |\nabla_j h(x)| < \lambda \quad (j \in Z), \qquad 2\|\kappa\|_2 < \mu r. \tag{8}$$

Then F has a unique global minimizer $x^\star$ in the relative interior of $\mathcal{B}$, with support S and signs s. The set $\mathcal{C} = \{x \in \mathcal{B} : F(x) \leq F(z)\}$ is invariant under every exact coordinate update, and

$$\|z - x^\star\|_2 \leq \frac{\|\kappa\|_2}{\mu}, \qquad F(z) - F^\star \leq \frac{\|\kappa\|_2^2}{2\mu}. \tag{9}$$

Suppose additionally that active scalar curvature is bounded below by $\mu_i > 0$ and (4) holds on $\mathcal{B}$ for $i, j \in S$. In the inherited cyclic order, let $D = \operatorname{diag}(\mu_i)$ and let U_β be the strict upper part of (β_{ij}). Set

$$C_\beta = \|U_\beta D^{-1/2}\|_2^2, \qquad q_E = \frac{C_\beta}{\mu + C_\beta}. \tag{10}$$

Every exact cycle starting in $\mathcal{C}$ satisfies $F(x^{k+1}) - F^\star \leq q_E(F(x^k) - F^\star)$.

Why the test works. On the face, the penalty is affine. Strong convexity gives $F(z + d) - F(z) \geq -\|\kappa\|_2\|d\|_2 + \mu\|d\|_2^2/2$. Condition (8) makes every boundary point more expensive than z. Convexity then prevents an exact coordinate minimizer from crossing that boundary from $\mathcal{C}$. The inactive margins force every inactive update to equal zero. At the interior minimizer, active stationarity and the inactive margins are the full optimality conditions.

For the rate, let Δ_i be the active change during a cycle. Exact scalar minimization decreases F by at least $\sum_i \mu_i \Delta_i^2/2$. The gradient of the restricted objective at the cycle end is bounded componentwise by $U_\beta|\Delta|$, since only later updates can invalidate the optimality equation of coordinate i. Restricted strong convexity bounds the remaining gap by the squared gradient norm divided by 2μ. Combining these statements gives (10). The proof in Appendix C includes the invariance and global-uniqueness arguments. The constant can be conservative, but it remains strictly below one even when $\rho(\Gamma) \geq 1$.

Two distinct sources of inconclusiveness. Positive restricted curvature need not imply absolute contraction. For $H = 0.2I + 0.8\mathbf{1}\mathbf{1}^\top$ in three active coordinates, $\lambda_{\min}(H) = 0.2$, but the absolute minimizer-influence matrix has radius 1.6. No order passes that test, although Theorem 2 applies near a strictly complementary solution. Appendix I.5 gives the full construction.

Screening can fail for a different reason, even when its dual point is improved. Consider

$$F(x) = \tfrac{1}{2}(x_1 - 2)^2 + \tfrac{1}{2}(x_2 + x_3 - 1 + \delta)^2 + \|x\|_1, \quad 0 < \delta < 1. \tag{11}$$

At $z = (1 + \varepsilon, 0, 0)$, $0 < \varepsilon < 1/2$, the primal test passes with $r = 3\varepsilon$, $\mu = 1$, active residual ε, and inactive margin δ. The minimizer is $(1, 0, 0)$.

Proposition 3 (Separation from every feasible spherical center)**.** *For* (11)*, the minimum standard Gap Safe bound on either inactive correlation, over* all *feasible dual centers at this fixed candidate, is*

$$\inf_{\nu \text{ feasible}} \left\{ |\nu_2| + \sqrt{2(F(z) - D(\nu))} \right\} = \sqrt{(1 - \delta)^2 + \varepsilon^2}. \tag{12}$$

Thus no such sphere screens either inactive coordinate when $\varepsilon^2 \geq 2\delta - \delta^2$, although the primal test passes. In particular, take $\varepsilon = 10^{-3}$ and $\delta = 10^{-9}$.

Here the unaveraged dual uses $|\nu_1|, |\nu_2| \leq 1$. Its gap is at least $[\varepsilon^2 + (\nu_2 - 1 + \delta)^2]/2$. The triangle inequality gives (12), with equality at $(\nu_1, \nu_2) = (1, 0)$. Appendix M gives the full derivation. Intersections, nonspherical regions, or other dual arguments are not covered by this separation.

4.2 A SHARPER COORDINATE-BOX CERTIFICATE

Let $r \in \mathbb{R}^{|S|}_{++}$ satisfy $r_i < |z_i|$, and replace (7) by $\mathcal{B}_\square = \{x : x_Z = 0,\ |x_S - z_S| \leq r\}$. Assume uniform inactive margins as in (8), active coordinate curvature $\mu_i > 0$, and cross-gradient bounds β_{ij} on this box. Define $b_i = |\kappa_i|/\mu_i$ and $\Gamma_{ij} = \beta_{ij}/\mu_i$ for $i \neq j$, with diagonal zero.

Theorem 4 (Influence-box certificate)**.** *If $b + \Gamma r < r$, then $\mathcal{B}_\square$ is invariant under every exact coordinate update and contains the unique global minimizer, with support S and signs s. Moreover $\rho(\Gamma) < 1$, the cyclic comparison* (3) *applies on the box, and*

$$|z_S - x^\star_S| \leq a := (I - \Gamma)^{-1} b, \qquad F(z) - F^\star \leq |\kappa|^\top a. \tag{13}$$

A second posterior bound is $|z_S - x^\star_S| \leq (I - K)^{-1} |\mathcal{P}(z)_S - z_S|$.

The strict inequalities ensure opposite derivative signs at the two endpoints of every active coordinate interval. Thus the exact minimizer stays inside the box. They also imply $\Gamma r < r$, which supplies the contraction needed for existence and uniqueness. Appendix D gives the complete proof. A candidate box obtained from a center Hessian is only a *proposal*: the bounds must be recomputed on the proposed box before the test is accepted.

4.3 ENTRY INTO THE CERTIFIED REGION

For a nondegenerate optimum, certification is not restricted to globally weak coupling. Suppose h is C^2 near $x^\star$, $|\nabla_j h(x^\star)| < \lambda$ for $j \notin \operatorname{supp}(x^\star)$, and $H_\star = \nabla^2_{SS} h(x^\star) \succ 0$. Exact cyclic minimization then has an identification basin and a stable local cycle derivative

$$G_\star = -(D_\star + L_\star)^{-1} U_\star, \qquad H_\star = D_\star + L_\star + U_\star, \quad \rho(G_\star) < 1. \tag{14}$$

Here $L_\star, U_\star$ are the signed Hessian parts, unlike L, U in (3). Appendix E proves the local statement and the following additional consequence.

Corollary 5 (Eventual curvature certification)**.** *For candidates $z \to x^\star$ on its identified face, sufficiently accurate uniform curvature and inactive-gradient bounds make Theorem 2 applicable eventually. For $\kappa \neq 0$, the proposal $r = 3\|\kappa\|_2 / \lambda_{\min}(\nabla^2_{SS} h(z))$ suffices for all sufficiently close candidates. No condition on $\rho(|D_\star^{-1}(H_\star - D_\star)|)$ is required.*

This is an asymptotic completeness statement under strict complementarity and positive restricted curvature, not a finite-time guarantee for arbitrary candidates. With a positive ridge penalty, standard global convergence of exact cyclic minimization supplies eventual entry; we give a self-contained proof. A local signed radius remains a predictor rather than a finite-neighborhood certificate unless uniform bounds have also been checked.

4.4 PARAMETER ACCURACY AND SAFEGUARDED INEXACT UPDATES

Once Theorem 2 or the screened face test has located $x^\star$ in $\mathcal{B}$, every $x \in \mathcal{B}$ has the two parameter-error bounds

$$\|x - x^\star\|_2 \le b_P(x) := \frac{\|\nabla_S h(x) + \lambda s\|_2}{\mu}, \qquad b_D(x) := \sqrt{\frac{2(F(x) - D(\nu))}{\mu}}, \tag{15}$$

Both b_P and b_D bound $\|x - x^\star\|_2$; the latter uses any feasible dual lower bound. Neither requires that x was produced by exact updates. Membership and signs must be checked if the path is not validated. Unlike an uncalibrated KKT tolerance, $b_P(x) \le \tau$ states the requested coefficient accuracy directly. With $\bar{g}_0 = \|\kappa\|_2^2/(2\mu)$, the exact-cycle assurance for this accuracy is

$$N_\tau = \left\lceil \frac{[\log(2\bar{g}_0/(\mu\tau^2))]_+}{-\log q_E} \right\rceil \quad (0 < q_E < 1). \tag{16}$$

The zero-rate case needs at most one further cycle. These are worst-case budgets, not signed local predictions.

There is also a direct way to preserve the region under inexact scalar solves. Freeze Z at zero and require objective nonincrease at each active update. Convexity and the strict boundary gap then keep every intermediate point in $\mathcal{C}$. If its post-update scalar residual ξ_i and displacement Δ_i satisfy $|\xi_i| \le \theta\mu_i|\Delta_i|$ with $0 \le \theta < 1/2$, the objective contracts per cycle with

$$q_\theta = \frac{(\sqrt{C_\beta} + \theta\sqrt{\mu_{\max}})^2}{\mu(1 - 2\theta) + (\sqrt{C_\beta} + \theta\sqrt{\mu_{\max}})^2} < 1, \quad \mu_{\max} = \max_{i \in S} \mu_i. \tag{17}$$

Appendix N proves both statements, including an additive-error version. A fixed scalar tolerance alone does not enforce these conditions. The large numerical study uses its parameter enclosures, not a claimed validated inexact trajectory.

5 SPARSE-LEARNING IMPLEMENTATION AND EXACTNESS

Consider

$$h(x) = \frac{1}{m}\sum_{t=1}^{m} \phi_t(X_t x) + \frac{\lambda_2}{2}\|x\|_2^2, \qquad \lambda_2 \ge 0. \tag{18}$$

For a candidate z, the prediction $v_t = X_t z$ varies by at most $d_t = \|X_{tS}\|_2 r$ on the ball and $d_t = \sum_{i \in S} |X_{ti}| r_i$ on the box. Let ℓ_t, u_t bound the scalar curvature on $[v_t - d_t, v_t + d_t]$. With $c_t = (\ell_t + u_t)/2$ and $a_t = (u_t - \ell_t)/2$, define

$$\begin{aligned} H_{\mathrm{lo}} &= m^{-1} X_S^\top \operatorname{diag}(\ell) X_S + \lambda_2 I, & H_{\mathrm{hi}} &= m^{-1} X_S^\top \operatorname{diag}(u) X_S + \lambda_2 I, \\ H_c &= m^{-1} X_S^\top \operatorname{diag}(c) X_S + \lambda_2 I, & V &= m^{-1} |X_S|^\top \operatorname{diag}(a) |X_S|. \end{aligned} \tag{19}$$

Then $H_{\mathrm{lo}} \preceq \nabla^2_{SS} h(x) \preceq H_{\mathrm{hi}}$ and $|\nabla^2_{SS} h(x) - H_c| \le V$ throughout the neighborhood, wherever derivatives exist. Thus we may use

$$\mu = \lambda_{\min}(H_{\mathrm{lo}}), \quad \mu_i = (H_{\mathrm{lo}})_{ii}, \quad \beta_{ij} = |(H_c)_{ij}| + V_{ij} \quad (i \ne j). \tag{20}$$

Unlike $m^{-1}\sum_t M_t |X_{ti} X_{tj}|$, these cross bounds retain cancellations in the center matrix. The inactive test is

$$|\nabla_j h(z)| + \frac{1}{m}\sum_t |X_{tj}| u_t d_t < \lambda \quad (j \in Z). \tag{21}$$

For logistic loss, curvature is $\sigma(v)\sigma(-v)$, so interval endpoints and the point nearest zero determine its extrema. Squared loss has constant curvature one. Huber loss with threshold one has a Lipschitz derivative; interval overlap with $[-1, 1]$ gives curvature bounds in $[0, 1]$. The integrated gradient inequalities remain valid at its kinks. The differentiable local formula (14) requires a neighborhood avoiding those kinks.

Table 2: First joint decision on the same trajectory. Entries are median epochs over five objectives per row. GS–DE improves the dual point without changing the primal path. Every terminal candidate passes the primal curvature test.

Problem	m/n	P	GS+F	GS–DE+F
Lasso, $c = .10$	128/1024	3	4	4
Lasso, $c = .65$	128/1024	14	17	12
Logistic, $c = .10$	256/1024	10	15	13
Logistic, $c = .65$	256/1024	34	46	46
Digits–ReLU	300/768	22	28	19
Diabetes–ReLU	320/767–768	103	105	84

Computational cost. Let $s = |S|$ and let s_t count active nonzeros in row t. Each radius trial costs $O(\mathrm{nnz}(X)+\sum_t s_t^2+s^3)$ for the dense cancellation-preserving certificate, with $O(ms+m+n+s^2)$ workspace beyond stored data. Dense data give $O(mn+ms^2+s^3)$. Candidate predictions, gradients, active row norms, and the center Hessian may be reused across trials; interval-dependent matrices must be rebuilt. The raw magnitude operator instead admits Kv in $O(\mathrm{nnz}(X) + m + n)$ time and $O(m + n)$ workspace. That product neither computes a restricted eigenvalue nor implements the full certificate at the same cost. Appendix L details storage, caching, and path accounting.

Numerical and exact implementations. The GLM evaluator uses binary64 arithmetic with scale-aware 10^{-12} guards. These guards are not outward error bounds, so its accepted outputs remain numerical witnesses. For rational squared-loss inputs, the certificate test uses exact arithmetic. It forms $H = X^\top X/m$ and $b = X^\top y/m$ rationally, establishes $H_{SS} - \mu I \succ 0$ by rational $LDL^\top$, and tests the sign radius together with

$$4\|\kappa\|_2^2 < \mu^2 r^2, \qquad \delta_j := \lambda - |(Hz - b)_j| > 0, \qquad r^2\|H_{jS}\|_2^2 < \delta_j^2 \ (j \in Z). \tag{22}$$

All acceptance comparisons use integers and fractions. The returned squared-distance and gap bounds are rational, as is the rate upper bound obtained by replacing C_β by $\sum_{i<j} H_{ij}^2/H_{jj}$. Appendix O proves soundness for the represented rational inputs. The guarantee does not account for uncertainty in the underlying measurements.

Use in a solver. Radius proposals use the candidate residual and center Hessian, but every proposed region receives fresh uniform bounds before acceptance. Failure is inconclusive. After a joint decision, remove Z and stop when the requested bound in (15) is met. A reduced numerical iterate is checked against its accepted ball and signs before its bound is used. No reference solution or observed convergence ratio enters these decisions. The full test is Algorithm 1 in Appendix R.

6 Experiments

We study 30 non-ridge objectives with $n > m$: lasso and logistic regression at two correlations, five seeds each, and fixed random-ReLU readouts for digit parity and diabetes regression, five splits each. All objectives use training-only standardization, omit the intercept, and keep the feature maps and regularization choices fixed across methods. The ambient Hessians are singular, while final supports have 5–28 coordinates. Appendix J specifies the instances. All large-instance acceptances below are numerical witnesses, independently checked against polished reference solutions.

Stronger dual points on identical trajectories. We compare the primal test (P), residual Gap Safe plus the same face test (GS+F), and Gap Safe with dual extrapolation plus that test (GS–DE+F). Extrapolation uses at most five consecutive prediction differences, followed by feasible rescaling and monotone dual-objective selection (Massias et al., 2020). The same unscreened primal trajectory supplies every candidate in Table 2. Both screening baselines must establish all proposed zeros before the face test; neither is restricted to a partial-support output.

Against residual screening, P is earlier in 28 cases, tied in one, and later in one. Against extrapolated screening, these counts become **17, 3, and 10**, with median paired epoch ratio 0.91 rather than 0.75. Thus dual-point quality accounts for part of the advantage over residual screening. Extrapolation is

Table 3: Total time to a joint support-and-parameter decision, relative to the primal route on the same objective. Values are median [interquartile range] after taking the median of three repetitions. The last row's fine-target time uses 28 paired completions; its two failures at 5,000 cycles remain in the success count and records.

Route	$\tau = 10^{-4}$		$\tau = 10^{-6}$	
	Met	Relative time	Met	Relative time
Primal + residual	30/30	1.00 [1.00, 1.00]	30/30	1.00 [1.00, 1.00]
GS + face residual	30/30	0.42 [0.32, 0.55]	30/30	0.44 [0.32, 0.56]
GS–DE + face residual	30/30	0.45 [0.35, 0.61]	30/30	0.48 [0.37, 0.62]
GS–DE + gap/curvature	30/30	0.66 [0.48, 0.93]	28/30	0.69 [0.52, 1.06]

particularly useful on the nonlinear-feature objectives. All accepted support/sign decisions match the independent references. Appendix P describes the dual construction and cost comparison; complete results are in the supplement.

A requested parameter-error decision. The operational task is to return the complete support and a bound $\|x - x^\star\|_2 \le \tau$, for $\tau \in \{10^{-2}, 10^{-4}, 10^{-6}\}$. The primal and two screening-plus-face routes use b_P. A fourth route uses the extrapolated dual gap and the same established face curvature in b_D. Screening acts online before the joint decision, then all four methods use reduced cyclic updates. Failed radii, full scans, matrix construction, eigenvalue calculations, successful checks, and subsequent stopping work are charged. We run six policies, including two accuracy-only controls, on each of 30 objectives, with three timing repetitions: **540 runs**. The nested thresholds give 1,260 recorded endpoints, not independent problems.

In Table 3, the residual-based parameter bounds meet all 30 requests at each tolerance, with no observed enclosure violation. Screening remains cheaper: at 10^{-6} the median ratios are 0.44 for GS+F and 0.48 for GS–DE+F. The dual-gap parameter route meets 28/30 fine targets; two objectives stall above the numerical gap threshold in all three repetitions despite small KKT residuals. Their median capped costs are 1.65 and 1.61 seconds. These failures are not mathematical failures of the bound. The accuracy-only controls illustrate the distinct contract: KKT stopping at 10^{-7} leaves reference parameter error above 10^{-6} in 13/30 cases; extrapolated-gap stopping at 10^{-8} does so in 11/30. Neither control was designed to guarantee that parameter tolerance.

How conservative is the remaining-cycle assurance? At the first primal decision, median q_E is 0.946. For parameter targets 10^{-4} and 10^{-6}, (16) gives median additional budgets 7.5 and 238, while the numerical paths take 1 and 14 cycles to the corresponding reference distances. No tested count exceeds its budget, but the gap is substantial. We therefore treat the rate as an assurance, not an acceleration heuristic. Appendix Q includes all tolerances and a separate exact-rational comparison in which both the updates and the stopping comparisons are exact.

Exact arithmetic and numerical limits. The rational certificate test evaluates 217 compact squared-loss inputs, including ill-conditioned active blocks, small inactive margins, and incorrect supports or curvature proposals. It accepts 37, all with the stated exact posterior bounds. On three rational problems, exact cycles require 2–6 passes for relative parameter reductions 10^{-2}–10^{-6}, against bounds of 4–15. Seventy-two exact-rational cycle checks also establish descent, support preservation, and the inexact inequality for full and three-quarter scalar steps. This is a delimited implementation result, not an interval guarantee for the floating-point GLM suite.

In the 1,044-case rational boundary study, the unguarded binary64 evaluator has 11 false *condition* acceptances, not demonstrated wrong-support outputs; guards refuse 72 otherwise valid cases. Input rounding is included. In the online and regularization-path studies, ordinary Gap Safe has median solve-time ratios 0.44 and 0.54, respectively, relative to unscreened cyclic descent. One online objective reaches its KKT target before certification, explaining 29/30 online versus 30/30 terminal acceptances. Those timing repetitions and dependent path points are not additional independent objectives. The appendices give the complete comparisons, including scalar-tolerance failures and unfavorable ordering results.

7 DISCUSSION, LIMITATIONS, AND CONCLUSION

The certificate requires sign separation, a strict inactive margin, and positive *uniform* curvature on the proposed active region. Failure is inconclusive. The support belongs to the minimizer of the empirical objective, not a population model. Dense active matrices limit scale, and the experiments stop at 1,024 features with small supports. Fixed random features do not establish results for end-to-end neural-network training. Exact arithmetic currently covers rational squared loss; general GLM outputs remain numerical witnesses. Ordinary scalar tolerances do not imply the safeguarded inexact theorem.

Decision value versus solver speed. The parameter-request study changes what is being timed. A residual threshold is a stopping rule in first-order units, whereas a bound on $\|x - x^\star\|_2$ specifies an error in the coefficients. The restricted curvature converts between these quantities and explains why the same KKT tolerance has different parameter consequences across instances. Screening followed by the same face analysis supplies that conversion as well. Its lower cost in our experiments is substantive evidence that the new test should complement, rather than replace, a good screening implementation. The advantage established by Proposition 3 is an additional route to a decision at a particular candidate, not a claim about which implementation will produce that candidate sooner.

Combining primal and dual information. There is no need for a solver to commit permanently to one source of information. A feasible dual point supplies a global objective lower bound and potentially safe zero coordinates. An accepted primal neighborhood supplies uniform restricted curvature, active signs, and a parameter enclosure. Once both are available, the smaller posterior bound in (15) may be used. The separate experimental arms deliberately isolate these sources of information, including the two fine-accuracy failures of the gap-only arm. A hybrid that chooses the smaller bound would avoid those particular failures, but its overall timing advantage over the tested residual-based routes has not been established here.

What an arithmetic certificate guarantees. The rational backend provides a complete finite check for its represented squared-loss input, including singular ambient problems. It does not validate the measurements that produced the input, select a statistically appropriate penalty, or certify a population support. Likewise, a correct optimizer enclosure does not alone validate a floating-point trajectory. The descent safeguard supplies the missing path condition, and the relative scalar-residual inequality supplies a rate for that path. These distinctions allow the theorem, numerical witnesses, and exact rational results to coexist without assigning the strongest guarantee to every experiment.

Conclusion. The contribution is a finite decision that turns a candidate into a complete support statement, a parameter enclosure, and a controlled continuation region. It can succeed beyond absolute contraction and beyond every center of a specified spherical screening family. Improved dual points nevertheless erase part of its advantage in decision epoch, and screened face analysis is faster for the tested parameter requests. The useful role of the primal test is therefore additional assurance when its neighborhood conditions are available, or a complementary route when spherical localization is inconclusive. Combining such finite-region guarantees with stronger screening and validated numerical linear algebra is more promising than treating certification as free solver acceleration.

AI USE STATEMENT

Generative AI assisted in literature search, grammatical and structure revision, and figure construction. The authors assume responsibility for the content.

REFERENCES

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. URL https://archive.ics.uci.edu/dataset/80/optical+recognition+of+handwritten+digits. Dataset DOI 10.24432/C50P49. CC BY 4.0.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013. doi: 10.1137/120887679.

Jelena Diakonikolas. Open problems in optimization. *SIAG on Optimization Views and News*, 33(1):1–11, 2025. URL https://siagoptimization.github.io/assets/views/ViewsAndNews-33-1.pdf.

Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

Mingyi Hong, Xiangfeng Wang, Meisam Razaviyayn, and Zhi-Quan Luo. Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, 163:85–114, 2017. doi: 10.1007/s10107-016-1057-8.

Quentin Klopfenstein, Quentin Bertrand, Alexandre Gramfort, Joseph Salmon, and Samuel Vaiter. Local linear convergence of proximal coordinate descent algorithm. *Optimization Letters*, 18: 135–154, 2024. doi: 10.1007/s11590-023-01976-z.

Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Mingyi Hong. On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization. *Journal of Machine Learning Research*, 18(184):1–24, 2018. URL https://jmlr.org/papers/v18/17-157.html.

Mathurin Massias, Alexandre Gramfort, and Joseph Salmon. Celer: a fast solver for the lasso with dual extrapolation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3315–3324. PMLR, 2018. URL https://proceedings.mlr.press/v80/massias18a.html.

Mathurin Massias, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Dual extrapolation for sparse GLMs. *Journal of Machine Learning Research*, 21(234):1–33, 2020. URL https://jmlr.org/papers/v21/19-587.html.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18(128):1–33, 2017. URL https://jmlr.org/papers/v18/16-577.html.

Brendan O'Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15:715–732, 2015. doi: 10.1007/s10208-013-9150-3.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010. doi: 10.1017/S096249291000005X.

Ankan Saha and Ambuj Tewari. On the nonasymptotic convergence of cyclic coordinate descent methods. *SIAM Journal on Optimization*, 23(1):576–601, 2013. doi: 10.1137/110840054.

Scikit-learn developers. Diabetes dataset, load_diabetes. Scikit-learn documentation, 2026. URL https://scikit-learn.org/1.8/modules/generated/sklearn.datasets.load_diabetes.html. Version 1.8 documentation, accessed September 24, 2026. Experiments use scikit-learn 1.8.0.

Ruoyu Sun and Mingyi Hong. Improved iteration complexity bounds of cyclic block coordinate descent for convex problems. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1306–1314, 2015. URL https://papers.nips.cc/paper_files/paper/2015/hash/96b9bff013acedfb1d140579e2fbeb63-Abstract.html.

Rachael Tappenden, Peter Richtárik, and Jacek Gondzio. Inexact coordinate descent: Complexity and preconditioning. *Journal of Optimization Theory and Applications*, 170:144–176, 2016. doi: 10.1007/s10957-016-0867-4.

Richard S. Varga. *Matrix Iterative Analysis*. Springer, second edition, 2000. doi: 10.1007/978-3-642-05156-2.

William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast cancer Wisconsin (diagnostic). UCI Machine Learning Repository dataset, 1993. URL https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. Dataset DOI 10.24432/C5DW2B. CC BY 4.0.

A NOTATION AND LOGICAL DEPENDENCIES

The results concern the deterministic full-cycle map, unless randomization is explicitly introduced in the experimental protocol. We use the following notation consistently.

Symbol	Meaning
$X \in \mathbb{R}^{m\times n}$	Feature matrix, with row X_t and column $X_{:i}$
h, F, λ, λ_2	Smooth convex part, full objective, ℓ_1 penalty, ridge penalty
$t_i, \mathcal{T}_i, \mathcal{P}$	Scalar minimizer, vector replacement, full cyclic composition
$x^{k,i}$	Point after coordinate i in cycle k, with $x^{k,0} = x^k$
z, S, Z, s, κ	Candidate, its support, complement, active signs, and face residual
$\mathcal{B}, \mathcal{C}$	Candidate neighborhood and its trapped sublevel set
μ, μ_i, β_{ij}	Restricted strong convexity, scalar curvature, cross-gradient bounds
Γ, L, U, K	Nonnegative minimizer influence, triangular parts, cycle comparison
$w, \|\cdot\|_{\infty,w}$	Positive weight vector and $\max_i \lvert v_i\rvert/w_i$ norm
$H_\star, G_\star$	Hessian at the optimum and signed local cycle derivative
$H_{\mathrm{lo}}, H_{\mathrm{hi}}, H_c, V$	Uniform lower/upper curvature matrices, interval center and entrywise variation
K_{abs}	Cycle matrix formed from the entrywise absolute local minimizer Jacobian
r_i^k	Scalar output distance from an exact coordinate minimizer, not objective suboptimality

The main dependence chain is

$$\begin{aligned}\text{prediction neighborhood} &\implies (H_{\mathrm{lo}}, H_{\mathrm{hi}}, \beta, \text{inactive margins})\\ &\implies \text{candidate test}\\ &\implies \text{invariance, optimality, and a cycle rate.}\end{aligned}$$

No posterior statement substitutes a center Hessian for a uniform neighborhood bound. The global comparison requires the influence property wherever its within-cycle comparisons are made. The candidate theorems instead establish the required invariant region directly.

B CYCLE COMPARISON, WEIGHTS, AND ORDERING

We first record the comparison for the more general composite objective

$$F(x) = h(x) + \sum_{i=1}^{n} \psi_i(x_i), \tag{23}$$

where h is convex and differentiable and the ψ_i are proper, closed, and convex. The scalar minimizers must exist and be unique on the domain in question. Coordinate strong convexity is a sufficient condition; a tie-breaking convention alone does not establish a Lipschitz influence bound.

Lemma 6 (Cycle comparison)**.** *Let $x^\star$ be a minimizer of* (23)*. Suppose* (2) *is valid for comparisons between $x^\star$ and every within-cycle input generated by the method. Then* (3) *holds. If $\|K\|_{\infty,w} = q < 1$, then*

$$\|x^k - x^\star\|_{\infty,w} \le q^k \|x^0 - x^\star\|_{\infty,w}.$$

If the comparison is valid between every pair of points in a set invariant under the full cycle, the full cycle is a contraction on that set. Uniqueness is then asserted within that domain. Global uniqueness requires a domain covering all minimizers or a separate argument such as the one used in Theorem 2.

Proof. The optimal point is a fixed point of each coordinate minimizer. At coordinate i, earlier coordinates already have their new-cycle values and later coordinates retain their old-cycle values. Thus

$$|x_i^{k+1} - x_i^\star| \le \sum_{j<i} \alpha_{ij} |x_j^{k+1} - x_j^\star| + \sum_{j>i} \alpha_{ij} |x_j^k - x_j^\star|.$$

Stacking yields $(I - L)e^{k+1} \le Ue^k$. A strict lower triangular matrix is nilpotent, so $(I - L)^{-1} = \sum_{a=0}^{n-1} L^a \ge 0$ and multiplication preserves the inequality. The norm estimate follows by induction. Applying the same argument to two trajectories proves contraction of the cycle map when the pairwise hypothesis holds throughout its invariant domain. Two fixed points in that domain would have distance at most q times their own distance and therefore coincide. □

Lemma 7 (Curvature implies minimizer influence)**.** *Assume the ith scalar objective in* (23) *is μ_i-strongly convex and the cross-gradient inequality* (4) *holds at the points used below. Then* (2) *holds with $\alpha_{ij} = \beta_{ij}/\mu_i$.*

Proof. Let $u = t_i(z_{-i})$ and $v = t_i(y_{-i})$. There exist $\zeta_u \in \partial\psi_i(u)$ and $\zeta_v \in \partial\psi_i(v)$ with

$$\nabla_i h(z^{(i\leftarrow u)}) + \zeta_u = 0, \qquad \nabla_i h(y^{(i\leftarrow v)}) + \zeta_v = 0.$$

Strong monotonicity of the subdifferential of the scalar objective with conditioning vector z_{-i} gives

$$\begin{aligned}\mu_i|v-u|^2 &\le \big[\nabla_i h(z^{(i\leftarrow v)}) + \zeta_v - \nabla_i h(z^{(i\leftarrow u)}) - \zeta_u\big](v-u)\\ &= \big[\nabla_i h(z^{(i\leftarrow v)}) - \nabla_i h(y^{(i\leftarrow v)})\big](v-u).\end{aligned}$$

When $u \neq v$, divide the absolute-value bound by $\mu_i|u-v|$ and use (4). When $u = v$, the claim is immediate. For local boxes all the compared points lie in the same box. For more general domains, this membership is part of the hypothesis rather than an implicit assumption. □

B.1 Choosing weights and evaluating a scalar bound

For $K \ge 0$,

$$\|K\|_{\infty,w} = \max_i \frac{(Kw)_i}{w_i}, \qquad \inf_{w>0} \|K\|_{\infty,w} = \rho(K). \tag{24}$$

A positive Perron vector, when available, attains the infimum. Since a cyclic K is often reducible, a robust alternative is $w = (rI - K)^{-1}\mathbf{1}$ for $r > \rho(K)$, yielding $Kw < rw$. A cheaper certificate uses $\delta_i = \sum_{j<i} \alpha_{ij}w_j/w_i$ and $\eta_i = \sum_{j>i} \alpha_{ij}w_j/w_i$:

$$\delta_i < 1 \text{ for all } i \quad\Longrightarrow\quad \|K\|_{\infty,w} \le \max_i \frac{\eta_i}{1-\delta_i}. \tag{25}$$

The proof is a forward induction, provided in the appendix. Small *weighted influence*, not sparsity alone, makes this bound useful.

Lemma 8 (Nonnegative weights and scalar certificate)**.** *For $B \ge 0$, $\|B\|_{\infty,w} = \max_i (Bw)_i/w_i$ and $\inf_{w>0} \|B\|_{\infty,w} = \rho(B)$. If $r > \rho(B)$, then*

$$w_r = (rI - B)^{-1}\mathbf{1} > 0, \qquad Bw_r = rw_r - \mathbf{1} < rw_r.$$

In particular $\rho(B) < 1$ is equivalent to existence of $w > 0$ with $Bw < w$. Equation (25) *also holds.*

Proof. The induced norm is the maximum row sum of $\operatorname{diag}(w)^{-1}B\operatorname{diag}(w)$, and hence bounds its spectral radius from above. The resolvent series $(rI - B)^{-1} = \sum_{k\ge 0} B^k/r^{k+1}$ is nonnegative and has a positive diagonal. It yields the displayed identity and the upper bound $\|B\|_{\infty,w_r} < r$. Letting $r \downarrow \rho(B)$ proves the infimum formula, including reducible matrices. If a positive Perron vector exists it attains equality, but such existence is not needed.

For the scalar certificate, let $y = Kw$ and $s_i = y_i/w_i$. Forward substitution gives

$$s_i = \sum_{j<i} \alpha_{ij}\frac{w_j}{w_i}s_j + \eta_i.$$

Set $q = \max_i \eta_i/(1-\delta_i)$. Since $\delta_i < 1$, induction gives $s_i \le \delta_i q + \eta_i \le q$, starting with the first row. Therefore $\|K\|_{\infty,w} = \max_i s_i \le q$. □

Proof of Proposition 1. Fix an order and suppress π. The splitting satisfies

$$I - \Gamma = (I - L)(I - K).$$

If $\rho(\Gamma) < 1$, choose $w = (I-\Gamma)^{-1}\mathbf{1} > 0$. Then $(I-K)w = (I-L)^{-1}\mathbf{1} > 0$, so $Kw < w$ and Lemma 8 gives $\rho(K) < 1$. Conversely, if $\rho(K) < 1$, then

$$(I-\Gamma)^{-1} = (I-K)^{-1}(I-L)^{-1} \ge 0.$$

The vector $w = (I-\Gamma)^{-1}\mathbf{1}$ is strictly positive, because both factors have nonnegative entries and positive diagonal. It satisfies $\Gamma w = w - \mathbf{1} < w$, giving $\rho(\Gamma) < 1$. Simultaneous permutation preserves the latter spectral radius, so the equivalence holds for every order. This is the regular-splitting argument underlying the proposition, rather than a new characterization of M-matrices (Varga, 2000). □

Proposition 9 (Sharp quadratic realization). *Let $\Gamma = \Gamma^\top \geq 0$ have zero diagonal and $\rho(\Gamma) < 1$. Then $Q = I - \Gamma \succ 0$, and exact cyclic minimization of $x^\top Qx/2$ has cycle map $K = (I - L)^{-1}U$. For $x \geq 0$, the componentwise absolute-error comparison is equality.*

Proof. Every eigenvalue of the symmetric Γ has absolute value at most $\rho(\Gamma) < 1$, so $Q \succ 0$. The scalar optimality equation is $x_i = \sum_{j \neq i} \alpha_{ij} x_j$. Using updated values for earlier coordinates gives $x^+ = Lx^+ + Ux$ and hence $x^+ = Kx$. Since $K \geq 0$, a nonnegative input remains nonnegative and absolute values do not relax the equality. A spectral radius describes asymptotic growth of powers, not a universal finite-step ratio in a prescribed norm. Normalized power iterations used in the experiments are reported separately from this algebraic identity. □

C Proof of the curvature certificate

For $x_Z = 0$ with active signs s, write $f_S(x_S) = h(x_S, 0) + \lambda s^\top x_S$. This function agrees with F on $\mathcal{B}$, is continuously differentiable there, and is μ-strongly convex when the restricted h is. Its gradient at z_S is κ.

Lemma 10 (Strict inactive margin). *If $g : \mathbb{R} \to \mathbb{R}$ is convex and differentiable and $|g'(0)| < \lambda$, then zero is the unique minimizer of $g(u) + \lambda|u|$.*

Proof. Convexity gives $g(u) \geq g(0) + g'(0)u$. Consequently

$$g(u) + \lambda|u| - g(0) \geq (\lambda - |g'(0)|)|u| > 0$$

for every $u \neq 0$. Strong convexity is not required. □

Proof of Theorem 2. **Step 1. An interior minimizer.** For any active displacement d with $\|d\|_2 = r$, strong convexity gives

$$F(z + (d, 0)) - F(z) \geq \kappa^\top d + \tfrac{\mu}{2} r^2 \geq -\|\kappa\|_2 r + \tfrac{\mu}{2} r^2 > 0.$$

The continuous objective attains a minimum on the compact ball $\mathcal{B}$. Every boundary value is larger than $F(z)$, so the restricted minimizer $x^\star$ is in the relative interior. Therefore $\nabla_S h(x^\star) + \lambda s = 0$. The inactive margin implies $|\nabla_j h(x^\star)| < \lambda$ for all $j \in Z$. These are precisely the global convex optimality conditions $0 \in \nabla h(x^\star) + \lambda \partial \|x^\star\|_1$. Thus $x^\star$ is a global minimizer, with support S and signs s.

Step 2. Global uniqueness. If y is another minimizer, convexity of h and the subgradient formula for the norm give

$$0 = F(y) - F(x^\star) \geq \sum_{j \in Z} (\lambda - |\nabla_j h(x^\star)|)|y_j|.$$

Every coefficient is positive, so $y_Z = 0$. If $y_S \neq x^\star_S$, points on the segment joining $x^\star$ to y sufficiently close to $x^\star$ lie inside $\mathcal{B}$. Convexity would make them minimizers, whereas restricted strong convexity makes the minimizer inside $\mathcal{B}$ unique. This contradiction proves uniqueness in the full space, not merely on the candidate face.

Step 3. Invariance of the sublevel set. Let $x \in \mathcal{C}$. For an active coordinate i, the intersection of its coordinate line with $\mathcal{B}$ is a closed interval with two boundary endpoints. Both endpoint objective values exceed $F(z) \geq F(x)$. A convex scalar function with an interior point below both endpoints has all its minimizers inside the interval: its one-sided slopes at the left and right endpoints are strictly negative and positive, respectively. A minimum exists in the interval, and the coordinate curvature lower bound implied by restricted μ-strong convexity makes it unique there. Convexity excludes an additional minimizer outside. The update remains in $\mathcal{B}$ and does not increase F, hence remains in $\mathcal{C}$. For $j \in Z$, the strict inactive margin and Lemma 10 force its update to be exactly zero. It follows by induction over coordinates and cycles that every within-cycle iterate remains in $\mathcal{C}$ and has the same support and signs. This argument establishes the domain needed by the subsequent rate proof.

Step 4. Posterior error bounds. Strong monotonicity of ∇f_S between z_S and $x^\star_S$ gives

$$\mu \|z_S - x^\star_S\|_2^2 \leq \kappa^\top (z_S - x^\star_S) \leq \|\kappa\|_2 \|z_S - x^\star_S\|_2.$$

Division proves the distance estimate. Strong convexity in the opposite direction gives

$$F(x^\star) \geq F(z) + \kappa^\top (x_S^\star - z_S) + \tfrac{\mu}{2}\|x_S^\star - z_S\|_2^2 \geq F(z) - \frac{\|\kappa\|_2^2}{2\mu},$$

which proves the gap estimate.

Step 5. Objective contraction. Relabel the active coordinates in their inherited order as $1, \ldots, s$. Let x^+ be the end of a cycle starting at $x \in \mathcal{C}$, and let $\Delta_i = x_i^+ - x_i$ be the change at the unique visit to coordinate i. Exact minimization and scalar μ_i-strong convexity imply

$$F(x) - F(x^+) \geq \frac{1}{2}\sum_i \mu_i \Delta_i^2. \tag{26}$$

Immediately after coordinate i is updated, its face gradient is zero. At the cycle end only coordinates $j > i$ have subsequently changed. Both comparison points lie in the convex ball $\mathcal{B}$ and share their ith coordinate, so (4) gives

$$|(\nabla f_S(x_S^+))_i| \leq \sum_{j>i} \beta_{ij}|\Delta_j|.$$

Therefore

$$\|\nabla f_S(x_S^+)\|_2^2 \leq \|U_\beta D^{-1/2}\|_2^2 \, \|D^{1/2}\Delta\|_2^2 = C_\beta \sum_i \mu_i \Delta_i^2.$$

Strong convexity on $\mathcal{B}$, applied between x^+ and the interior minimizer, also gives $F(x^+) - F^\star \leq \|\nabla f_S(x_S^+)\|_2^2/(2\mu)$. Combining with (26),

$$F(x^+) - F^\star \leq \frac{C_\beta}{\mu}\big(F(x) - F(x^+)\big).$$

Rearranging proves $q_E = C_\beta/(\mu + C_\beta)$. When $C_\beta = 0$, the ending face gradient is zero and a single cycle reaches the optimum. No inverse of $I - \Gamma$ is used in this proof. □

Complexity interpretation. For $0 < q_E < 1$, a sufficient number of subsequent cycles to make the gap at most ϵ is

$$k \geq \frac{\log((F(z) - F^\star)/\epsilon)}{-\log q_E}.$$

The unknown initial gap may be replaced by (9). If $q_E = 0$, one cycle suffices. This is a local, certified complexity bound from z, not a bound on the number of cycles needed to produce a candidate that passes the test.

D Proof of the influence-box certificate

Proof of Theorem 4. Fix an active coordinate i and the values of all other active coordinates in the box. At $u = z_i$, its scalar face derivative differs from κ_i by at most $\sum_{j\neq i} \beta_{ij} r_j$. Uniform scalar strong convexity then implies that the derivative at $u = z_i + r_i$ is at least

$$-|\kappa_i| - \sum_{j\neq i} \beta_{ij} r_j + \mu_i r_i > 0,$$

and the derivative at $u = z_i - r_i$ is at most

$$|\kappa_i| + \sum_{j\neq i} \beta_{ij} r_j - \mu_i r_i < 0.$$

These are exactly the strict inequalities $b + \Gamma r < r$. Since the coordinate interval does not cross zero, the face derivative is the derivative of the original scalar objective on this interval. Continuity produces a root in its interior and strong convexity makes it unique. Global scalar convexity makes this root the unique scalar minimizer. Inactive coordinates remain zero by Lemma 10. Thus each vector update maps the compact box into itself.

The same inequalities give $\Gamma r < r$, hence $\rho(\Gamma) < 1$ by Lemma 8. Proposition 1 gives $\rho(K) < 1$. Lemma 7 applies within the box, and Lemma 6 shows that its full cycle is a contraction in some

weighted maximum norm. The box is closed in that norm and therefore complete. Banach's theorem gives a unique fixed point of the cycle. A fixed point of a cyclic exact-update composition must be coordinatewise fixed: each coordinate is updated only once, so its final unchanged value was also the value at its update. Consequently all scalar optimality conditions hold. For the separable ℓ_1 penalty they combine into global convex optimality. The point is interior to every active coordinate interval and satisfies strict inactive margins. For global uniqueness, strict inactive margins force any other global minimizer to have zero inactive coordinates, as in the proof of Theorem 2. A sufficiently short segment from the interior fixed point toward another such minimizer would stay inside the box. Convexity would make each point on that segment a minimizer and hence a fixed point, contradicting uniqueness of the fixed point in the box.

At the candidate, the distance from z_i to $t_i(z_{-i})$ is at most b_i. To see this, strong monotonicity of the scalar face gradient between z_i and its root gives $\mu_i|z_i - t_i(z_{-i})| \le |\kappa_i|$. The influence comparison against $x^\star$ therefore yields

$$|z_S - x^\star_S| \le b + \Gamma|z_S - x^\star_S|.$$

Multiplication by $(I - \Gamma)^{-1} \ge 0$ gives the first posterior bound. Convexity of the face objective implies

$$F(z) - F^\star \le \kappa^\top (z_S - x^\star_S) \le |\kappa|^\top (I - \Gamma)^{-1} b.$$

Finally apply the cycle comparison to z and $x^\star$, add $|z_S - \mathcal{P}(z)_S|$, and solve the resulting inequality

$$|z_S - x^\star_S| \le |z_S - \mathcal{P}(z)_S| + K|z_S - x^\star_S|.$$

This proves the second posterior estimate. □

Why the box proposal is not circular. The implementation first computes a pointwise minimizer Jacobian J_0 from the candidate Hessian. When $\rho(|J_0|) < 1$, it proposes a radius

$$r^{\mathrm{trial}} = (I - |J_0|)^{-1}(b_0 + \epsilon \mathbf{1})$$

with $\epsilon > 0$. Neither J_0 nor b_0 is treated as a neighborhood guarantee. The algorithm recomputes scalar curvature, cross-gradient, and inactive bounds throughout the proposed box and then evaluates the strict test in Theorem 4. A failed recomputation is a rejected proposal, not a certificate. The curvature-ball procedure follows the same proposal–check separation.

E Local stability, identification, and eventual certification

Theorem 11 (Local exact-coordinate dynamics)**.** *Suppose h is convex on $\mathbb{R}^n$ and C^2 in a neighborhood of a minimizer $x^\star$ of* (1)*. Set $S = \mathrm{supp}(x^\star)$. Assume $|\nabla_j h(x^\star)| < \lambda$ for $j \notin S$ and $H_\star = \nabla^2_{SS} h(x^\star) \succ 0$. Then there is a neighborhood $\mathcal{N}$ of $x^\star$ such that an exact cyclic sweep starting in $\mathcal{N}$ sets every inactive coordinate to zero and has the correct active signs by its end. All subsequent partial iterates remain on that face and converge locally linearly. On the face the full-cycle derivative is* (14)*. For every $q \in (\rho(G_\star), 1)$, sufficiently close cycle starts satisfy an R-linear Euclidean distance bound with factor q and an objective bound of order q^{2k}.*

Proof. For inactive j, continuity provides a neighborhood of $x^\star_{-j}$ in which $|\nabla_j h(x^{(j \leftarrow 0)})| < \lambda$. Lemma 10 then makes the inactive scalar minimizer identically zero there. For an active coordinate i, consider the smooth equation

$$\nabla_i h(x^{(i \leftarrow u)}) + \lambda \,\mathrm{sign}(x^\star_i) = 0.$$

Its derivative in u at the solution is $(H_\star)_{ii} > 0$. The implicit function theorem supplies a unique C^1 local root depending on the other coordinates. Its sign remains that of $x^\star_i$. Global convexity and positive local curvature make this root the unique global scalar minimizer, so the actual exact update is the local root. All vector updates and their partial compositions are therefore continuous at $x^\star$ and fix it.

On the active face, differentiation of each scalar optimality equation gives $J_{ij} = -(H_\star)_{ij}/(H_\star)_{ii}$ for $i \ne j$ and $J_{ii} = 0$. Propagating a perturbation through a cycle gives $d_i^+ = \sum_{j<i} J_{ij} d_j^+ + \sum_{j>i} J_{ij} d_j$, and hence $G_\star = (I - L_J)^{-1} U_J = -(D_\star + L_\star)^{-1} U_\star$.

To show stability without absolute diagonal dominance, consider the positive definite quadratic $Q(v) = v^\top H_\star v/2$. Its exact cyclic map is $G_\star$. Each scalar update decreases Q, and the total decrease is $\sum_i (H_\star)_{ii}\Delta_i^2/2$. If a nonzero vector had zero total decrease, every scalar change would vanish, every partial gradient would be zero, and $H_\star v = 0$, a contradiction. Thus $Q(G_\star v) < Q(v)$ for every $v \neq 0$. Compactness of the $H_\star$-unit sphere gives an induced norm contraction strictly below one, implying $\rho(G_\star) < 1$.

For any $q > \rho(G_\star)$, choose q_0 strictly between them and a vector norm whose induced norm of $G_\star$ is at most q_0. One explicit construction uses the positive definite series $P = \sum_{k\geq 0} q_0^{-2k}(G_\star^k)^\top G_\star^k$, which converges when $q_0 > \rho(G_\star)$ and gives $G_\star^\top P G_\star = q_0^2(P - I) \preceq q_0^2 P$. Continuity of the cycle derivative makes the face cycle a contraction with factor q in a sufficiently small convex P-norm ball. Shrink that ball so every partial cycle stays in the sign and inactive-margin neighborhood. The full cycle maps the ball into itself because it fixes $x^\star$.

Finally, continuity of the full and partial cycle maps gives an ambient neighborhood $\mathcal{N}$ whose first-cycle partial iterates stay inside the margin neighborhood and whose endpoint belongs to the invariant face ball. Each inactive coordinate is set to zero when visited; the active coordinates retain their correct signs. After this first cycle the face contraction applies inductively to every subsequent sweep. Norm equivalence gives an R-linear Euclidean distance estimate. A uniform upper Hessian bound on the face ball and zero face gradient at $x^\star$ give

$$F(x) - F^\star \leq \tfrac{1}{2}(x_S - x_S^\star)^\top H_{\mathrm{hi}}(x_S - x_S^\star),$$

and therefore an objective bound of order q^{2k}. The uniform H_{hi} is not being identified with $H_\star$. □

Proof of Corollary 5. Let z approach $x^\star$ on its identified face. Continuity gives $\kappa \to 0$ and $\mu_0(z) := \lambda_{\min}(\nabla^2_{SS} h(z)) \to \mu_\star > 0$. For $\kappa \neq 0$, set $r = 3\|\kappa\|_2/\mu_0(z)$, which tends to zero. The minimum active coefficient magnitude remains bounded away from zero, so eventually $r < \min_{i\in S} |z_i|$. The gradient continuity and strict inactive margins make (21), or any consistent uniform inactive bound, valid on this shrinking ball. A consistent uniform lower-curvature estimate $\mu(z, r)$ converges to $\mu_\star$. Therefore eventually $\mu(z, r)/\mu_0(z) > 2/3$, giving

$$\mu(z, r)r = 3\frac{\mu(z, r)}{\mu_0(z)}\|\kappa\|_2 > 2\|\kappa\|_2.$$

This is the curvature test. If $\kappa = 0$, the candidate already has active stationarity; any sufficiently small positive radius preserves positive curvature and the inactive margins. For GLMs with continuous scalar second derivatives, the interval bounds (19) are consistent as the radius tends to zero. The result also applies to Huber objectives when no residual at the optimum lies at a kink. □

Proposition 12 (Global entry for positive ridge regularization)**.** *For* (18) *with $\lambda_2 > 0$, convex losses with Lipschitz derivatives, and the ℓ_1 penalty, exact cyclic minimization from any finite point converges to the unique minimizer. If the nondegeneracy assumptions of Theorem 11 hold, finite identification and eventual curvature certification follow.*

Proof. The objective is strongly convex and coercive, so its initial sublevel set is compact and contains a unique minimizer. Every scalar subproblem is at least λ_2-strongly convex. Summing its decrease bound over all updates shows that the sum of squared scalar changes is finite. In particular, every within-cycle displacement tends to zero. Let a subsequence of cycle starts converge to a cluster point $\bar{x}$. All partial iterates in those cycles converge to the same $\bar{x}$. At the update of coordinate i,

$$0 \in \nabla_i h(x^{k,i}) + \lambda \partial |x_i^{k,i}|.$$

By continuity of the gradient and closedness of the scalar subdifferential graph, the limiting point satisfies these inclusions for all i. Separability combines them into the full convex optimality condition, so $\bar{x}$ is the unique minimizer. Every cluster point is the same, and compactness implies convergence of the entire sequence, including partial iterates. The local theorem then gives finite entry and identification. Corollary 5 applies to the identified sequence. □

F GLM NEIGHBORHOOD BOUNDS AND COMPUTATION

Proposition 13 (Uniform GLM bounds)**.** *Suppose each scalar derivative is absolutely continuous and its second derivative, wherever defined, lies in* $[\ell_t, u_t]$ *throughout the candidate prediction interval. Then* (19)–(21) *provide the curvature, cross-gradient, and inactive-gradient bounds required by Theorems 2 and 4.*

Proof. On either candidate region, Cauchy–Schwarz or the componentwise radius bound gives $|X_t(x-z)| \le d_t$. Write the diagonal loss curvature at x as $W(x)$. Then $\operatorname{diag}(\ell) \preceq W(x) \preceq \operatorname{diag}(u)$ and $|W_{tt}(x) - c_t| \le a_t$. Congruence with X_S proves the Loewner inequalities. Entrywise,

$$|(\nabla^2_{SS} h(x) - H_c)_{ij}| \le \frac{1}{m} \sum_t a_t |X_{ti} X_{tj}| = V_{ij}.$$

It follows that active scalar curvature is at least $(H_{\mathrm{lo}})_{ii}$ and the off-diagonal magnitude is at most $|(H_c)_{ij}| + V_{ij}$. Integrating the Hessian along the line segment between two points of the convex candidate region with the same ith coordinate proves (4). Integrating the lower Loewner bound gives restricted strong convexity with modulus $\lambda_{\min}(H_{\mathrm{lo}})$.

For an inactive coordinate, the ridge term vanishes because $x_j = z_j = 0$. The scalar derivative variation obeys

$$|\nabla_j h(x) - \nabla_j h(z)| \le \frac{1}{m} \sum_t |X_{tj}|\, u_t\, |X_t(x-z)| \le \frac{1}{m} \sum_t |X_{tj}| u_t d_t.$$

This proves the inactive test. The integrated argument uses almost-everywhere Hessians and absolute continuity, so it remains valid for Huber loss despite its isolated curvature discontinuities. It does not justify differentiating the coordinate map at a Huber kink. □

Loss conventions. Binary logistic loss is $\phi_t(v) = \log(1 + e^v) - y_t v$, with $y_t \in \{0, 1\}$. Let $b_t = \max(|v_t - d_t|, |v_t + d_t|)$ and let ν_t denote the distance of $[v_t - d_t, v_t + d_t]$ from zero. Its curvature bounds are $\ell_t = \sigma(b_t)\sigma(-b_t)$ and $u_t = \sigma(\nu_t)\sigma(-\nu_t)$. Squared loss is $(v - y_t)^2/2$. Huber loss is $(v - y_t)^2/2$ when $|v - y_t| \le 1$ and $|v - y_t| - 1/2$ otherwise. For Huber, $\ell_t = 1$ when the full residual interval lies in $[-1, 1]$ and zero otherwise; $u_t = 1$ when that interval intersects $[-1, 1]$ and zero otherwise.

Candidate procedure. The following procedure specifies what is checked and what is merely proposed.

1. Read the candidate support S and compute κ. If $S = \varnothing$, check the strict global zero-minimizer condition directly.
2. Propose a finite set of radii from the candidate curvature and residual. Reject every radius crossing an active sign boundary.
3. For each surviving radius, recompute prediction intervals, H_{lo}, cross bounds, and all inactive margins. Accept the ball only when (8) holds with a strictly positive numerical margin. Return its distance, gap, and objective-rate bounds.
4. Optionally form a box proposal from the pointwise absolute Jacobian resolvent. Recompute all bounds on the box and accept only when $b + \Gamma r < r$ and the inactive margins hold.
5. If no test succeeds, return "not certified." Neither failed test establishes an incorrect support or divergence.

For ball trials, the experiment uses $b_0 = \max(4\|\kappa\|_2/\lambda_{\min}(H(z)), 10^{-10})$, its multiples $\{0.5, 1, 2, 4, 8\}$, and fractions $\{0.01, 0.05, 0.1, 0.25, 0.5\}$ of the smallest active magnitude. It returns the accepted ball with the smallest q_E. Box proposals use $(I - |J_0|)^{-1}(b + 10^{-9}\mathbf{1})$ and multipliers $\{1.01, 1.2, 2, 5\}$. These are deterministic implementation choices, not assumptions of the theorems. Comparisons use a relative slack 10^{-10} and an inactive absolute slack 10^{-10}. Such slack is a practical tolerance, not an outward-rounded error bound.

Matrix-free raw influence product. Let $B_{ti} = \sqrt{M_t}|X_{ti}|$ and let the raw bound be $\Gamma_{ij} = \sum_t B_{ti}B_{tj}/(m\mu_i)$ for $i \neq j$. To compute $u = Uv$, maintain row accumulators $s_t = \sum_{j>i} B_{tj}v_j$ while traversing coordinates backward. Then $u_i = \sum_t B_{ti}s_t/(m\mu_i)$, after which the current column is added to the accumulators. To solve $y = u + Ly$, traverse forward with row accumulators $p_t = \sum_{j<i} B_{tj}y_j$ and use

$$y_i = u_i + \frac{1}{m\mu_i}\sum_t B_{ti}p_t.$$

Each nonzero is read a constant number of times. The work is $O(\mathrm{nnz}(X) + m + n)$ and extra workspace is $O(m + n)$. This gives matrix-free weighted-norm evaluation or resolvent iteration. It does not remove the cost of estimating a cancellation-preserving active Hessian, nor does an inexpensive product make a vacuous raw bound informative.

G EXACT, PROXIMAL-GRADIENT, AND INEXACT UPDATES

Proposition 14 (A precise exact-versus-gradient comparison)**.** *For* (23)*, assume scalar smooth curvature of h in $[\mu_i, L_i]$ and the cross-gradient bound* (4)*. At a minimizer $x^\star$, the proximal-coordinate step*

$$x_i^+ = \mathrm{prox}_{\eta_i\psi_i}(x_i - \eta_i\nabla_i h(x))$$

satisfies

$$|x_i^+ - x_i^\star| \le \sigma_i|x_i - x_i^\star| + \eta_i\sum_{j\neq i}\beta_{ij}|x_j - x_j^\star|, \quad \sigma_i = \max\{|1-\eta_i\mu_i|, |1-\eta_i L_i|\}.$$

The exact scalar map has zero dependence on the incoming x_i, but its curvature-based cross coefficient is β_{ij}/μ_i. Thus deleting a self coefficient does not imply that all remaining coefficients or all runtimes improve.

Proof. Optimality makes $x_i^\star$ a fixed point of the scalar proximal-gradient map with the other coordinates fixed at their optimal values. The proximal map is nonexpansive. Split the gradient difference into variation along coordinate i, with all other coordinates fixed at x_{-i}, and a remaining difference with coordinate i fixed at $x_i^\star$. The first difference equals $a_i(x_i - x_i^\star)$ for $a_i \in [\mu_i, L_i]$ by integration of scalar curvature. The second is bounded by (4). Nonexpansiveness and the triangle inequality give the assertion. The exact map $t_i(x_{-i})$ simply has no incoming x_i argument. □

Quadratics do not separate the updates. If $h(x) = x^\top Hx/2 - b^\top x$ with $H_{ii} > 0$, the scalar subproblem equals a constant plus $\frac{1}{2}H_{ii}(u - x_i + \nabla_i h(x)/H_{ii})^2 + \psi_i(u)$. Its minimizer is exactly the proximal-coordinate step with $\eta_i = 1/H_{ii}$. This includes lasso and elastic-net least squares. Therefore neither a quadratic experiment nor a separable quadratic example establishes a distinction between exact minimization and a properly scaled coordinate-gradient/proximal-gradient step.

A genuine nonquadratic separable example. For $0 < \mu < L$, let

$$g(u) = \frac{L}{2}u^2 - (L-\mu)\log\cosh u.$$

Then $g''(u) = L - (L-\mu)\,\mathrm{sech}^2 u \in [\mu, L]$, and its unique minimizer is zero. Exact minimization of $\sum_i g(x_i)$ solves it in one cycle. A coordinate-gradient step of length $1/L$ gives $x_i^+ = (1 - \mu/L)\tanh x_i$. Its derivative at zero is $1 - \mu/L$, and nonzero sufficiently small inputs converge geometrically rather than in one step. This is a fixed-stepsize, nonquadratic comparison. It does not preclude line search, Newton updates, acceleration, or another gradient stepsize from improving performance.

Random sampling in the separable limit. Both a deterministic full sweep and any permutation without replacement solve a separable objective with unique scalar minimizers in one epoch. Under independent uniform coordinate sampling with replacement, a coordinate is unvisited after a updates with probability $(1 - 1/n)^a$. Thus for fixed nonnegative weights v_i,

$$\mathbb{E}\sum_i v_i|x_i^{(a)} - x_i^\star| = (1-1/n)^a\sum_i v_i|x_i^0 - x_i^\star|.$$

This coupon-coverage effect is not a general deterministic-versus-reshuffling theorem.

G.1 Inexact scalar solves

Proposition 15 (Perturbed cycle recursion)**.** *Suppose the exact minimizer maps obey* (2) *on a domain containing the actual inexact partial iterates and $x^\star$. If the output at visit (k, i) differs from the exact minimizer at the same input by at most $r_i^k \geq 0$, then*

$$e^{k+1} \leq Ke^k + (I - L)^{-1} r^k. \tag{27}$$

For $q = \|K\|_{\infty,w} < 1$ and $\epsilon_k = \|(I - L)^{-1} r^k\|_{\infty,w}$,

$$\|e^k\|_{\infty,w} \leq q^k \|e^0\|_{\infty,w} + \sum_{t=0}^{k-1} q^{k-1-t} \epsilon_t. \tag{28}$$

If $\epsilon_t \leq C\theta^t$ and $0 < q, \theta < 1$, the forcing term is bounded by $C(q^k - \theta^k)/(q - \theta)$ for $\theta \neq q$ and Ckq^{k-1} for $\theta = q$.

Proof. At coordinate i, add the scalar approximation error to the exact-map comparison. Stacking gives $e^{k+1} \leq Le^{k+1} + Ue^k + r^k$. Multiplication by the nonnegative triangular inverse yields (27). Taking norms and iterating gives (28). The displayed geometric-series formulas follow directly. The cases $q = 0$ or $\theta = 0$ follow by the same convolution without division. □

Uniformly bounded $\epsilon_t \leq \epsilon$ gives a limiting error at most $\epsilon/(1 - q)$. Vanishing errors imply convergence, as is seen by splitting the convolution into a finite prefix and a uniformly small tail; summability is sufficient but not necessary. Errors decaying strictly faster than q^k preserve an $O(q^k)$ rate, equal-rate errors incur the factor k, and slower errors determine the resulting geometric rate.

For a μ_i-strongly convex scalar objective, a scalar subgradient residual of magnitude at most ε_i implies distance at most ε_i/μ_i from its minimizer, by strong monotonicity. Scalar objective suboptimality at most δ_i instead implies distance at most $\sqrt{2\delta_i/\mu_i}$. These are different quantities. The experiments use the first convention. A quadratic approximation to a GLM scalar objective falls under this result only after bounding the residual in the *original* scalar objective.

The candidate theorems state invariance for exact updates. To use a local certificate with inexact updates, one must additionally ensure that the perturbed outputs remain in the certified region, for example by checking membership and the required sublevel condition, and preserve the inactive threshold rule exactly. Small coordinate distance error alone need not produce exact zeros. We do not infer finite identification from an inexact norm bound without these conditions.

G.2 Block comparison

For disjoint coordinate blocks with norms $\|\cdot\|_i$, suppose exact block minimizers satisfy

$$\|t_i(z_{-i}) - t_i(y_{-i})\|_i \leq \sum_{j \neq i} \alpha_{ij} \|z_j - y_j\|_j.$$

The same proof with scalar block errors $e_i^k = \|x_i^k - x_i^\star\|_i$ gives (3) and (27). Block strong monotonicity μ_i and cross-block dual-norm bounds β_{ij} give $\alpha_{ij} = \beta_{ij}/\mu_i$ by the same optimality comparison and dual pairing. This extension does not automatically transfer scalar sign identification to group penalties or coupled constraints; those require their own manifold and nondegeneracy assumptions.

H Experimental protocol and additional results

The 75-objective ridge-regularized study complements the non-ridge decision comparison in Section 6. It examines certificate acceptance, local rate predictions, and the cost of exact scalar minimization under globally positive curvature. The supplement contains the complete results.

H.1 Data, objectives, and preprocessing

The 30 synthetic problems use 700 samples and 60 features with

$$X_{ti}^{\mathrm{raw}} = \sqrt{1-c}\,\xi_{ti} + \sqrt{c}\,\zeta_t, \quad \xi_{ti}, \zeta_t \overset{\mathrm{iid}}{\sim} N(0,1), \quad c \in \{0, 0.3, 0.7\}.$$

The first eight planted coefficients are independent standard normals, and labels are independent Bernoulli variables with probability $\sigma(X_t^{\mathrm{raw}}\beta)$. The first 500 examples are training data and the remaining 200 are held out. Seeds are $0, \ldots, 9$ for each c. Reusing seeds across correlation settings gives paired constructions, not 30 independent datasets.

WDBC contains 569 examples with 30 features. The scikit-learn digits subset contains 1,797 examples with 64 raw features; labels are digit parity. Diabetes contains 442 examples with ten features and uses Huber loss with threshold one. Public data are split 80/20 with seeds $0, \ldots, 4$, stratified by the binary label for WDBC and digit parity. This gives training sample sizes 455, 1,437, and 353, respectively. The digits subset is the original UCI test subset bundled by scikit-learn, but our within-subset splits do not purport to reproduce the official UCI evaluation. The public-data provenance is given in Wolberg et al. (1993); Alpaydin & Kaynak (1998); Scikit-learn developers (2026).

Training-column means and standard deviations are used to standardize both training and held-out features. Columns with training standard deviation at most 10^{-12} are removed. Diabetes responses are centered and scaled using training statistics. The scikit-learn diabetes loader itself distributes preprocessed features; our additional training transformation is the one used in the optimization experiments. All models omit the intercept so every method optimizes precisely the same explicitly specified objective. The experiments evaluate optimization mechanics, not a tuned prediction pipeline.

Let $\lambda_{\max} = \|\nabla \ell(0)\|_\infty$, where ℓ is the average loss. We use $\lambda_2 = 0.02$ throughout, $\lambda = 0.2\lambda_{\max}$ on synthetic data, and $\lambda/\lambda_{\max} \in \{0.1, 0.3, 0.6\}$ on public data. These fixed settings are not selected using held-out scores. The three strengths on each split are correlated observations; error bars across objectives are descriptive, not population confidence intervals.

H.2 Algorithms, accuracy, and timing

All algorithms start from zero. Exact cyclic uses natural feature order. Reshuffling samples an independent random permutation at each epoch; i.i.d. sampling makes n independent uniform coordinate draws per epoch. Random schedules are generated inside the timed loop, only when needed. Each stochastic configuration uses seed $10000+$ the data-split seed. Cyclic proximal-gradient uses $1/L_i$ with $L_i = M\|X_{:i}\|_2^2/m + \lambda_2$, where $M = 1/4$ for logistic loss and $M = 1$ for Huber loss. FISTA uses $L = M\|X\|_2^2/m + \lambda_2$ and the gradient restart test $(z - x^+)^\top(x^+ - x) > 0$ (O'Donoghue & Candès, 2015). Its spectral-norm setup is timed.

For nonquadratic exact updates, the scalar solver first tests the derivative at zero against $[-\lambda, \lambda]$. If zero is not optimal, it brackets a root on the appropriate half-line by doubling and then uses safeguarded Newton steps, falling back to bisection. The scalar stationarity tolerance is 10^{-12}, with at most 100 refinement steps. Thus "exact" in the experiments denotes this high-accuracy implementation of the exact-minimization algorithm, not symbolic arithmetic. The inexact comparison explains this distinction. Squared-loss tests use the exact soft-threshold formula.

The common stopping residual is

$$\mathcal{R}(x) = \max_i \begin{cases} |\nabla_i h(x) + \lambda \operatorname{sign}(x_i)|, & x_i \neq 0, \\ (|\nabla_i h(x)| - \lambda)_+, & x_i = 0. \end{cases}$$

It is evaluated after every epoch or FISTA iteration, and all methods stop at $\mathcal{R}(x) \leq 10^{-7}$. The cap is 3,000 such iterations. A separate exact-cyclic reference uses tolerance 2×10^{-12} and cap 6,000. Every reported run converged. The largest reference residual is 1.973×10^{-12}. Objective differences smaller than ordinary floating-point cancellation are not used as stopping tests. Since the reference is approximate, it is not labeled an exact optimum.

Runs use one CPU thread on an Intel Xeon Platinum 8573C, Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0, scikit-learn 1.8.0, and Numba 0.65.1. JIT compilation is warmed up before timings. Public-data configurations use three timing repeats of the same deterministic seed and report their median, with min/max retained in the raw results; synthetic configurations use one timed repeat each.

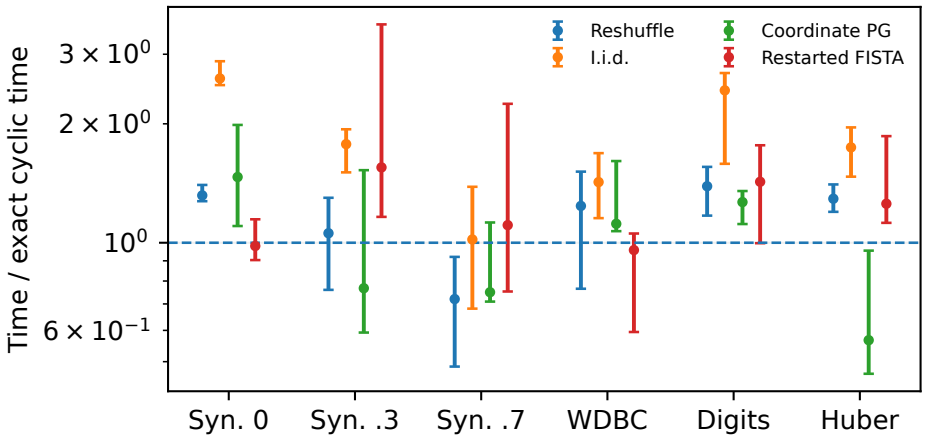


Figure 2: Paired wall-time ratios relative to exact cyclic minimization. Markers are medians and bars are interquartile ranges across objectives. Values below one favor the alternative. Fewer coordinate epochs do not necessarily translate into less time. Certification costs are excluded here and reported separately.

The timing includes solver-specific constants, random schedules, scalar iterations, and full stopping checks. Shared data preparation, JIT compilation, high-accuracy reference solutions, and separate certificate calculations are excluded. Consequently these numbers compare the implementations studied here, not production-optimized packages. FISTA performs full-vector gradient work and its iteration count should not be interpreted as an equal-cost coordinate epoch.

H.3 Retained ridge-regularized comparisons

We test three separate questions: whether a candidate can be certified, whether local operators predict convergence, and whether iteration savings survive their computational costs. We use 30 synthetic logistic instances with 60 features, eight planted coefficients, and common-factor correlation $c \in \{0, 0.3, 0.7\}$, with ten seeds per setting. Public-data experiments use WDBC, binary digit parity, and diabetes regression with Huber loss, each with five train/test splits and three regularization strengths (Wolberg et al., 1993; Alpaydin & Kaynak, 1998; Scikit-learn developers, 2026; Pedregosa et al., 2011). Features are standardized using training statistics, zero-variance columns are removed, and all objectives omit an intercept. The digits data are the 1,797-example subset distributed with scikit-learn, re-split here rather than evaluated on the original UCI partition. This gives 75 objectives, not 75 independent datasets.

We compare fixed-order exact cyclic updates, independently reshuffled exact epochs, i.i.d. exact updates with replacement, cyclic proximal-coordinate gradient, and gradient-restarted FISTA (Beck & Teboulle, 2009; O'Donoghue & Candès, 2015). Each method starts at zero and stops when the same full KKT residual is at most 10^{-7}. An epoch means n coordinate updates; a FISTA iteration is reported separately in the same table. Scalar solves use a safeguarded Newton/bisection procedure with residual tolerance 10^{-12}. All 375 configurations converge within the 3,000-iteration cap. Timings include method-specific setup, order generation, and stopping checks, but exclude common preprocessing and JIT compilation. Certification is timed separately. Appendix H specifies every setting, repeats, and machine details.

Certificate coverage and timing. The numerically evaluated curvature test accepts all 75 exact-cyclic terminal candidates. The box test accepts 29. The local absolute cycle radius is at least one in the other 46, while the raw global magnitude bound fails on all 75. Theorem 2 therefore covers a regime genuinely missed by the original influence criterion (Table 6). Numerically accepted posterior distance bounds are a median factor 1.35 above distance to a high-accuracy numerical reference. This comparison is not itself a proof of the bounds; the reference has a maximum KKT residual 1.98×10^{-12}, and its error allowance is included in the numerical tests.

Table 4: Median iterations and milliseconds, with interquartile ranges across objectives. Each public row aggregates five splits and three strengths. Timing IQRs describe differences among objectives, not only repeated-run noise.

Problem	Method	Iterations [Q1, Q3]	ms [Q1, Q3]
Synthetic, $c = 0$	cyclic	6 [5.25, 7]	3.13 [2.65, 3.35]
Synthetic, $c = 0$	reshuffle	8 [7.25, 9]	4.12 [3.65, 4.64]
Synthetic, $c = 0$	iid	15 [14, 18]	7.69 [6.82, 8.95]
Synthetic, $c = 0$	proxgrad	15 [13, 16]	4.60 [3.93, 4.99]
Synthetic, $c = 0$	fista	20.5 [16.25, 21.75]	3.06 [2.71, 3.36]
Synthetic, $c = 0.3$	cyclic	13.5 [9, 17.5]	6.94 [4.08, 8.95]
Synthetic, $c = 0.3$	reshuffle	12.5 [11.25, 13]	6.87 [5.04, 7.51]
Synthetic, $c = 0.3$	iid	21 [20, 22.75]	10.80 [8.94, 14.82]
Synthetic, $c = 0.3$	proxgrad	16.5 [15.25, 18]	5.71 [4.95, 6.34]
Synthetic, $c = 0.3$	fista	94.5 [82.25, 96]	11.64 [10.70, 12.36]
Synthetic, $c = 0.7$	cyclic	36.5 [22.25, 54]	16.95 [9.65, 25.40]
Synthetic, $c = 0.7$	reshuffle	23 [18.5, 25.75]	11.31 [8.07, 13.49]
Synthetic, $c = 0.7$	iid	35.5 [28.75, 37.75]	17.28 [13.18, 18.67]
Synthetic, $c = 0.7$	proxgrad	39 [35.25, 46]	12.62 [11.86, 15.90]
Synthetic, $c = 0.7$	fista	170 [159.75, 178]	19.62 [19.00, 22.28]
WDBC	cyclic	47 [37.5, 48]	10.50 [10.06, 18.52]
WDBC	reshuffle	46 [32, 70.5]	13.01 [11.48, 16.20]
WDBC	iid	61 [51.5, 73]	17.03 [14.03, 19.43]
WDBC	proxgrad	110 [83, 135]	15.82 [13.19, 19.50]
WDBC	fista	163 [151, 169]	9.91 [9.77, 11.21]
Digits parity	cyclic	8 [7, 15.5]	14.11 [10.64, 38.51]
Digits parity	reshuffle	12 [9.5, 21]	20.43 [14.41, 55.78]
Digits parity	iid	24 [13, 34]	38.72 [18.47, 93.33]
Digits parity	proxgrad	17 [9, 41]	19.27 [12.04, 50.37]
Digits parity	fista	55 [51.5, 83]	20.64 [18.71, 29.29]
Diabetes / Huber	cyclic	15 [11.5, 17]	0.51 [0.32, 0.65]
Diabetes / Huber	reshuffle	18 [14, 21.5]	0.59 [0.38, 0.84]
Diabetes / Huber	iid	26 [20.5, 29]	0.84 [0.56, 1.14]
Diabetes / Huber	proxgrad	18 [17, 19]	0.29 [0.28, 0.31]
Diabetes / Huber	fista	44 [42, 48]	0.69 [0.60, 0.76]

Table 5: Median iterations / milliseconds to the common KKT tolerance. The FISTA column counts full-vector iterations, not coordinate epochs. Public-data rows aggregate 15 objectives each, synthetic rows ten each. Per-objective timing uses three repeats for public data and one for synthetic data. Interquartile ranges are in Appendix H.

Problem	Exact cyclic	Reshuffle	I.i.d.	Coord. PG	FISTA
Synthetic, $c = 0$	6 / 3.13	8 / 4.12	15 / 7.69	15 / 4.60	20.5 / 3.06
Synthetic, $c = 0.3$	13.5 / 6.94	12.5 / 6.87	21 / 10.80	16.5 / 5.71	94.5 / 11.64
Synthetic, $c = 0.7$	36.5 / 16.95	23 / 11.31	35.5 / 17.28	39 / 12.62	170 / 19.62
WDBC	47 / 10.50	46 / 13.01	61 / 17.03	110 / 15.82	163 / 9.91
Digits parity	8 / 14.11	12 / 20.43	24 / 38.72	17 / 19.27	55 / 20.64
Diabetes / Huber	15 / 0.51	18 / 0.59	26 / 0.84	18 / 0.29	44 / 0.69

Table 6: Certificate coverage and local rates. "Curv." and "Box" count accepted terminal candidates. Radii are medians at those candidates; the last column is the median first epoch accepted by the curvature test. K_{abs} uses the entrywise absolute local minimizer Jacobian.

Problem	Cases	Curv.	Box	$\rho(K_{\mathrm{abs}})$	$\rho(G_\star)$	Entry
Synthetic, $c = 0$	10	10	10	0.136	0.083	3
Synthetic, $c = 0.3$	10	10	4	2.801	0.379	6
Synthetic, $c = 0.7$	10	10	0	11.652	0.707	19
WDBC	15	15	0	8.425	0.774	21
Digits parity	15	15	10	0.385	0.147	3
Diabetes / Huber	15	15	5	1.460	0.358	5

When checked after every exact cycle, the curvature certificate first succeeds at a median of eight epochs, or half the iterations required by the 10^{-7} stopping rule. The median one-shot certification cost is 2.60 ms, approximately 41% of the paired solve time. Checking after every epoch costs a median 1.94 times the solve time in total. These figures argue for selective certification, not for treating it as a free acceleration.

Prediction is not an exact finite-step norm ratio. After a small support-preserving perturbation of a high-accuracy reference, we fit the logarithmic distance decay above a specified numerical-noise cutoff. A fit is available in 74 of 75 cases; one is excluded because too few points remain above the cutoff. The median absolute difference from $\rho(G_\star)$ is 0.00221, with a maximum of 0.0567. Figure 3 distinguishes these fitted rates from the often noncontracting absolute envelope. Nonnormal transients and initial directions explain why equality is not required. The scalar-accuracy study also detects the expected error floor: tolerance 10^{-4} fails the 10^{-7} KKT target in all three tested cases, whereas 10^{-8} and 10^{-12} succeed.

Runtime and ordering. Exact cyclic updates use fewer epochs than coordinate PG in the median of every problem group, but they are slower in wall time on several groups. Reshuffling is faster for strongly correlated synthetic designs. We make no universal ordering or exactness claim. To examine the proposed ordering diagnostic directly, we enumerate all 7! orders for 18 positive quadratic instances. The greedy incoming-mass rule is compared with natural, reverse, 20 fixed random orders, and exact best/worst spectral-radius oracles. At coupling 0.92, the greedy rule has median percentile 95.4% among all orders, where a larger percentile is worse. Ordering differences are small in that regime, and a smaller spectral radius need not reduce the finite stopping time. Full search costs about 0.38 seconds per seven-coordinate instance, far more than its solve. Appendix H.5 reports the complete results. The theory supplies a diagnostic, not a validated ordering speedup.

H.4 Certificates, local fits, and cost accounting

The two tests are evaluated at the exact-cyclic 10^{-7} terminal candidate and independently at the high-accuracy reference. The main-text coverage uses only the terminal candidates. In every accepted case, the active indices agree with those of the high-accuracy reference. The distance comparison allows for reference error: strong convexity implies $\|x^{\mathrm{ref}} - x^\star\|_2 \le \sqrt{n}\mathcal{R}(x^{\mathrm{ref}})/\lambda_2$. We

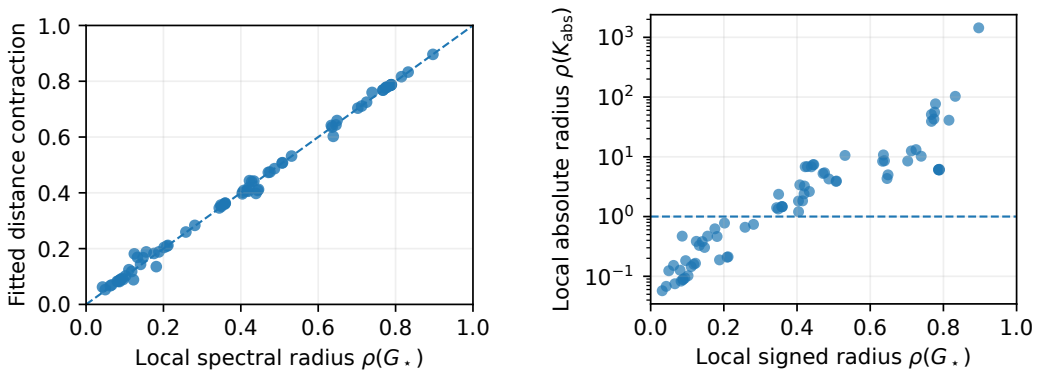


Figure 3: Left: fitted local distance contraction versus the signed cycle radius on 74 valid fits, with the identity line for reference. Right: absolute-influence radii at 75 terminal candidates. The horizontal line is the contraction threshold. Positive restricted curvature permits certification even when the nonnegative criterion fails.

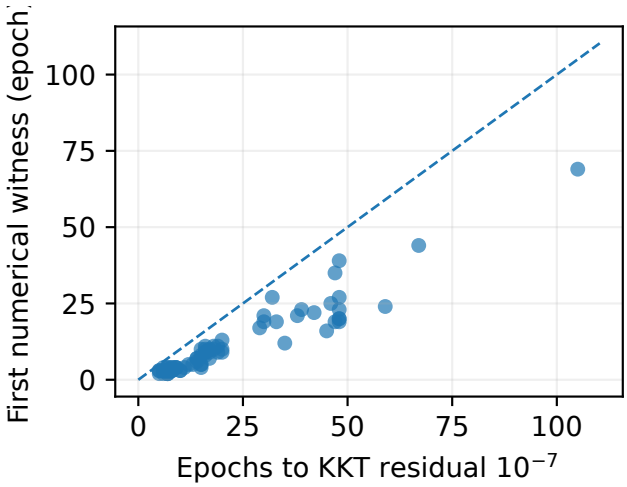


Figure 4: First accepted curvature certificate versus the epoch count required by the common terminal KKT rule. Every point is one objective. The diagonal marks equal counts. The certificate can validate support before high optimization accuracy, but repeated evaluation has a measurable cost.

use this allowance when checking the candidate-distance inequalities numerically. The analytic theorem, rather than reference agreement, is the basis for support certification.

To measure certificate entry, we invoke the ball check after each complete exact cycle until it succeeds, up to 200 cycles. All 75 cases pass within this cap. The median first accepted epoch is eight and the median ratio to the terminal solve's epoch count is 0.5 (Figure 4). The first-success KKT residuals are recorded rather than inferred from terminal results. Repeated checking is expensive in this implementation: cumulative checking costs a median 1.94 times the paired terminal solver time. A single terminal check costs a median 2.60 ms or 0.41 of that solver time. These experiments establish practical evaluability on small active supports, not a net acceleration from checking every sweep.

For the local-rate study, compute $G_\star$ and K_{abs} at the high-accuracy reference. Perturb its active coefficients in a fixed pseudorandom unit direction by magnitude $\min(10^{-3}, 0.05 \min_{i \in S} |x_i^{\mathrm{ref}}|)$. Run 60 exact cycles with scalar tolerance 10^{-13}. Fit $\log \|x_S^k - x_S^{\mathrm{ref}}\|_2$ over the last at most 12 admissible points, keeping only distances below half the initial displacement and above $\max(2 \times 10^{-10}, 20\mathcal{R}(x^{\mathrm{ref}})/\lambda_2)$. A fit requires at least four points. One of 75 cases fails this rule and is

not plotted. No fit-window choice is optimized against the predicted radius. A separate normalized linear power iteration records the asymptotic behavior of $G_\star$ without underflow. The finite nonlinear fits need not exactly equal its spectral radius, especially in the presence of nonnormal transients or unfavorable initial directions.

Raw operator scaling. The matrix-free test uses $m = 2,000$, about 20,000 nonzeros, and $n \in \{100, 500, 2,000, 10,000\}$ with nonnegative feature magnitudes and scalar denominators one. Seven timing batches, each with 100 products, give median times of 0.100, 0.114, 0.141, and 0.321 ms, respectively. At $n = 10,000$, the sparse inputs and workspace occupy 616,004 bytes, versus 800,000,000 bytes for a dense $n \times n$ array. Explicit dense checks at $n = 100$ and 500 have maximum absolute discrepancies 3.33×10^{-16} and 5.55×10^{-17}. These measurements concern applying the *raw* influence operator, not assembling all local signed bounds or training a 10,000-feature model.

Scalar-accuracy sensitivity. On the seed-zero synthetic $c = 0$ problem, seed-zero digit parity at strength 0.3, and seed-zero WDBC at strength 0.3, scalar tolerance 10^{-4} produces terminal KKT residuals 9.59×10^{-5}, 8.82×10^{-5}, and 8.80×10^{-5} after the 1,000-cycle cap, missing the 10^{-7} target. Tolerances 10^{-8} and 10^{-12} both reach the target in 7, 8, and 33 cycles, respectively. This is consistent with the bounded-error floor, but is not a sharp validation of every constant in Proposition 15.

H.5 Exhaustive ordering study

We generate six symmetric, positive, zero-diagonal 7×7 matrices using symmetrized lognormal weights, and rescale each to spectral radius $\tau \in \{0.3, 0.6, 0.92\}$. For a positive diagonal matrix D_s, define $\Gamma = D_s^{-1} W D_s$. This is realized by the positive definite quadratic Hessian $D_s(I - W)D_s$, whose diagonal is D_s^2 and whose coordinate-minimizer influence is Γ. All instances therefore belong to the exact sharpness family up to coordinate scaling.

For each of 18 matrices, enumerate all $7! = 5,040$ permutations. Natural and reverse orders are compared with 20 fixed random permutations, the spectral-radius best and worst orders, and a greedy incoming-mass rule. The greedy rule computes $w = (I - \Gamma)^{-1}\mathbf{1}$ and repeatedly selects, from the remaining set R, an index minimizing $\sum_{j \in R \setminus \{i\}} \alpha_{ij} w_j / w_i$. The orientation is incoming influence from not-yet-placed coordinates. Weight construction is included in greedy setup time; exhaustive search is included in oracle setup time.

Table 7: Median cycle radii over six matrices per coupling level, and the greedy order's median percentile among all orders. Higher percentile is worse. Exact spectral-radius oracles are not finite-time or wall-clock oracles.

$\rho(\Gamma)$	Natural	Greedy	Best oracle	Worst oracle	Greedy percentile
0.30	0.110327	0.107084	0.103614	0.116391	48.5
0.60	0.374650	0.372778	0.369821	0.379165	55.0
0.92	0.847358	0.847476	0.847033	0.847671	95.4

The greedy rule is not consistently near-optimal and is particularly poor by spectral percentile at $\tau = 0.92$. Natural and reverse orders have equal radii in this symmetrizable family, though finite trajectories can differ. At $\tau = 0.92$, median stopping counts from the specified all-ones initial vector to maximum-coordinate error 10^{-8} are 121.5 for natural order, 119.5 for greedy, 122 for the best spectral-radius oracle, and 123 for the worst. These values illustrate transient and norm dependence rather than contradicting spectral asymptotics. The median exhaustive search time is approximately 0.38 seconds per matrix; the matrix-iteration solve itself takes only fractions of a millisecond. This evidence does not justify introducing the greedy rule as a solver improvement. Its useful role here is to test and delimit the ordering interpretation of the comparison theorem.

I FINITE DECISIONS, SCREENING, AND SEPARATING EXAMPLES

I.1 WHAT A STOPPING THRESHOLD DOES NOT ESTABLISH

An accuracy certificate and a complete-support certificate answer different questions. For $a > 0$, $\lambda > 0$, and $\varepsilon > 0$, consider

$$f_\varepsilon(t) = \frac{a}{2}t^2 - (\lambda + \varepsilon)t + \lambda|t|.$$

Its unique minimizer is $t^\star = \varepsilon/a > 0$. At the zero candidate, the minimum-norm subgradient has magnitude ε, and the objective gap is $\varepsilon^2/(2a)$. Both quantities can be arbitrarily small while the candidate support is wrong. This does not invalidate residual or gap stopping for their intended accuracy objectives. It shows why no fixed positive threshold alone certifies support uniformly over this family. The primal test would require additional margin and curvature information, and must refuse the zero candidate here.

I.2 A SCREENING BASELINE WITH THE SAME DOWNSTREAM DECISION

Let F attain its minimum and suppose a valid safe-screening rule establishes that every global minimizer has $x_Z = 0$. This is a statement about all minimizers, not merely a guess based on small coefficients. Consider a candidate z with support $S = [n] \setminus Z$ and the sign-preserving ball $\mathcal{B}$ of (7).

Proposition 16 (Screening followed by a face test)**.** *Assume that every global minimizer is zero on Z, that the restriction of h to $\mathcal{B}$ is μ-strongly convex, and that $\mu > 0$ and $2\|\kappa\|_2 < \mu r$. Then F has a unique global minimizer in the relative interior of $\mathcal{B}$, with support S and signs s. The posterior bounds* (9) *hold. The set $\mathcal{C}$ is invariant under exact updates restricted to S, and the rate* (10) *holds for that reduced method under the same cross-gradient assumptions as Theorem 2. No uniform inactive-gradient bound is needed for this conclusion about the reduced method.*

Proof. The boundary argument in Theorem 2 locates a minimizer $\widehat{x}$ of the face restriction in the relative interior of $\mathcal{B}$. On this ball the penalty is affine, so $\nabla_S h(\widehat{x}) + \lambda s = 0$. Convexity makes $\widehat{x}$ a global minimizer of the objective restricted to $x_Z = 0$, including points on other sign faces. Since a global minimizer of F exists and lies in this subspace by the safe-screening hypothesis, the minimum over the subspace equals the full minimum. Thus $\widehat{x}$ is a full minimizer. Any second global minimizer also lies in the subspace. The segment joining it to $\widehat{x}$ would contain other minimizers arbitrarily near $\widehat{x}$, contradicting positive curvature on $\mathcal{B}$. Hence the minimizer is unique. The radius constraint gives its support and signs.

The distance and objective-gap bounds use only restricted strong convexity and the face residual, exactly as in the proof of Theorem 2. For a reduced update, any attempted passage from $\mathcal{C}$ through the ball boundary would contradict the boundary value inequality and convexity along the coordinate line. This proves reduced-update invariance. The descent and end-of-cycle gradient calculation uses only active coordinates, so it gives the same energy rate. There is no conclusion here that an update of an omitted coordinate at an intermediate point would remain zero. The uniform inactive-gradient condition in Theorem 2 supplies that stronger property for the original full-coordinate method. □

This proposition prevents a misleading comparison. Screening need not be presented as incapable of yielding complete-support information: it can be combined with an active test. In the experiments, "Gap Safe + face" uses the same active curvature bounds, radius proposals, posterior bounds, and rate calculation as the primal test. Its additional requirement is that every proposed zero is safely screened. It also refuses a candidate whose proposed nonzero coordinate has been screened. This last safeguard is conservative and does not weaken the mathematical implication of an accepted decision.

I.3 DUAL CONVENTIONS AND THE IMPLEMENTED SCREENING BALL

For the non-ridge squared-loss problem,

$$F(x) = \frac{\|y - Xx\|_2^2}{2m} + \lambda\|x\|_1,$$

the dual is

$$\max_{\|X^\top \nu\|_\infty \le \lambda} D(\nu), \qquad D(\nu) = y^\top \nu - \frac{m}{2}\|\nu\|_2^2.$$

For binary logistic loss with $y_t \in \{0,1\}$, write $p_t = \sigma(X_t x)$ and $q_t = y_t - m\nu_t$. The dual objective is

$$D(\nu) = -\frac{1}{m}\sum_t \{q_t \log q_t + (1-q_t)\log(1-q_t)\}, \quad 0 \le q_t \le 1, \quad \|X^\top \nu\|_\infty \le \lambda.$$

The conventions $0 \log 0 = 0$ apply. These are the standard Gap Safe constructions under our averaged-loss normalization (Ndiaye et al., 2017). The squared dual is m-strongly concave. The negative entropy Hessian for the logistic dual has magnitude $m/[q_t(1-q_t)] \ge 4m$, giving strong concavity $4m$. If $g = F(x) - D(\nu) \ge 0$, a dual optimizer lies within radius

$$R = \sqrt{2g/m} \quad \text{(squared)}, \qquad R = \sqrt{g/(2m)} \quad \text{(logistic)}.$$

Consequently

$$|X_{:j}^\top \nu| + \|X_{:j}\|_2 R < \lambda \quad \Longrightarrow \quad x_j^\star = 0 \text{ for every primal minimizer.} \tag{29}$$

To see the implication, the left side bounds $|X_{:j}^\top \nu^\star|$ throughout the dual ball. A strict bound below λ forces a zero primal coordinate by the primal-dual optimality conditions. Strong concavity gives the radius inequality by comparing the feasible dual point with a dual maximizer and using weak duality. No support reference is needed.

Our dual point is obtained from $\nu_0 = (y - Xx)/m$ for squared loss or $(y-p)/m$ for logistic loss, scaled by $\max\{1, \|X^\top \nu_0\|_\infty/\lambda\}$. The logistic domain is preserved by this scaling because it interpolates q between y and p. Numerically, an additional factor $1 + 64\epsilon_{\text{mach}}$ moves the point slightly into the linear-constraint interior. Screening uses a $10^{-12}\max(1,\lambda)$ margin. These adjustments do not constitute outward rounding; the baseline, like our test, is a guarded numerical implementation. This is the residual baseline. Appendix P implements and evaluates the stronger extrapolated-dual comparator.

I.4 A FINITE-CANDIDATE SEPARATION FROM THE RESIDUAL-BASED BALL

The objective (11) is a squared-loss problem with $m = 2$,

$$X = \sqrt{2}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \qquad y = \sqrt{2}\begin{pmatrix} 2 \\ 1-\delta \end{pmatrix}, \qquad \lambda = 1.$$

The full Hessian has rank two and null vector $(0,1,-1)$. For $0 < \delta < 1$, the unique solution is $(1,0,0)$. This follows either from Theorem 2 at the solution, or by minimizing first over x_1 and then over $t = x_2 + x_3$: the second objective is at least $\frac{1}{2}(t-1+\delta)^2 + |t|$, whose unique minimizing sum is zero, and the minimum penalty then forces $x_2 = x_3 = 0$.

At $z = (1+\varepsilon, 0, 0)$, $0 < \varepsilon < 1/2$, the active residual is ε, the active curvature is one, and $r = 3\varepsilon < 1 + \varepsilon$. Both inactive gradients equal $-(1-\delta)$ for every point on the candidate face, so the strict inactive margin is δ. The curvature conditions all hold. The posterior distance bound is ε and the objective-gap bound is $\varepsilon^2/2$; both are attained in this example. Subsequent updates of coordinates 2 and 3 remain zero, and one active exact update sets $x_1 = 1$.

The residual dual point is already feasible in exact arithmetic,

$$\nu = \frac{1}{\sqrt{2}}(1-\varepsilon, 1-\delta)^\top, \qquad X^\top \nu = (1-\varepsilon, 1-\delta, 1-\delta)^\top.$$

Direct substitution gives $F(z) - D(\nu) = \varepsilon(1+\varepsilon)$. Since each inactive column has norm $\sqrt{2}$, the left side of (29) is

$$1 - \delta + \sqrt{2\varepsilon(1+\varepsilon)}.$$

Thus neither inactive coordinate is screened whenever $\delta \le \sqrt{2\varepsilon(1+\varepsilon)}$. For $\varepsilon = 10^{-3}$ and $\delta = 10^{-6}$, the gap is 0.001001, whereas the primal test already licenses the joint decision. The example does not establish that no dual region could succeed: changing the active dual correlation to its optimal value would dramatically tighten the region. It isolates a specific loss of information in the residual-derived isotropic ball, namely variation in an active residual being charged to an inactive direction whose gradient is constant on the primal candidate face.

I.5 Separation from absolute contraction

Let $S = \{1, 2, 3\}$ and $H = 0.2I + 0.8\mathbf{1}\mathbf{1}^\top$. For any $\lambda > 0$, define

$$F(x) = \tfrac{1}{2}(x_S - \mathbf{1})^\top H(x_S - \mathbf{1}) - \lambda \mathbf{1}^\top x_S + \tfrac{1}{2}x_4^2 + \lambda\|x\|_1.$$

The unique minimizer is $(1, 1, 1, 0)$. The active Hessian has eigenvalues $0.2, 0.2, 2.6$, unit diagonal, and positive off-diagonal entries 0.8. The absolute minimizer-influence matrix therefore has spectral radius 1.6. Proposition 1 rules out a contracting absolute cycle envelope for every order. The signed cycle radius is approximately 0.716, while the energy bound in Theorem 2 is approximately 0.893.

For $z = (1 + \varepsilon, 1 - \varepsilon, 1, 0)$ the face residual is $0.2\varepsilon(1, -1, 0)$. The radius $r = 3\sqrt{2}\varepsilon$ satisfies $2\|\kappa\|_2 < 0.2r$ and is sign-preserving for $0 < \varepsilon < (1 + 3\sqrt{2})^{-1}$. The inactive gradient is identically zero on the face. Thus the curvature test applies despite failure of every absolute ordering test. This example and (11) address different obstructions; no single test is asserted to dominate all screening and comparison methods.

J Non-ridge experiments and common-decision protocol

J.1 Objectives and frozen nonlinear features

The primary suite contains six groups of five objectives. Synthetic lasso uses $m = 128$, synthetic logistic regression uses $m = 256$, and both use $n = 1024$ before standardization. There are 200 independent held-out rows. For correlation $c \in \{0.1, 0.65\}$, unstandardized rows are

$$A_{ti} = \sqrt{1 - c}\,\xi_{ti} + \sqrt{c}\,u_t, \qquad \xi_{ti}, u_t \overset{\text{iid}}{\sim} N(0, 1).$$

The first six true coefficients are $(1.5, -1.3, 1.1, -1, 0.8, -0.6)$ and the rest are zero. Lasso responses add independent $N(0, 0.1^2)$ noise. Logistic responses are independent Bernoulli variables with parameter $\sigma(A_t\beta)$. Synthetic seeds are $33000 + j$, $j = 0, \ldots, 4$, with the same seed rule across correlation groups. Responses for squared loss are centered and standardized using training statistics. Set $\lambda = 0.2\lambda_{\max}$ for lasso and $0.3\lambda_{\max}$ for logistic loss, where $\lambda_{\max} = \|\nabla h(0)\|_\infty$. All ridge penalties are zero.

The other ten objectives use the scikit-learn digits and diabetes data (Alpaydin & Kaynak, 1998; Scikit-learn developers, 2026; Pedregosa et al., 2011). Digits are re-split into 300 training examples and the remaining test examples, stratified by the parity label. Diabetes uses 320 training examples and the remaining test examples, with a squared-loss response rather than the Huber loss used in the ridge-regularized study. Splits use random states $j = 0, \ldots, 4$. The original features are standardized from training observations and transformed as

$$\Phi(a) = \max\{aW + b, 0\}, \qquad W_{ij} \sim N(0, 1/d), \quad b_j \sim N(0, 0.5^2),$$

where d is the number of nonconstant original features and the requested width is 768. Transformation seeds are $44000 + j$. The transformed features are again standardized using training statistics, with constant columns removed, leaving 767–768 features. Set $\lambda = 0.3\lambda_{\max}$. The random map is fixed throughout optimization and is not a trained representation. Its role is to create a nonlinear prediction problem with an overcomplete convex readout, not to establish state-of-the-art representation learning.

All 30 full Hessians are rank deficient since $n > m$ and $\lambda_2 = 0$. The numerical design ranks are retained with the result records. Active support sizes range from 5 to 28; the smallest reference restricted eigenvalue is about 0.02455. These choices deliberately exercise restricted rather than global curvature, but do not establish robustness for arbitrary high-dimensional learning distributions or weak regularization. Held-out metrics are retained as consistency diagnostics and are not used to choose hyperparameters, candidates, radii, or decision times.

J.2 Scalar updates, references, and trial radii

Cyclic updates use the shared safeguarded Newton/bisection routine with scalar residual tolerance 10^{-13} and double-precision arithmetic. The high-accuracy reference uses tolerance 3×10^{-12} on

the full KKT residual and 10^{-14} for scalar solves, with a 10,000-cycle cap. A separate linear solve on the reference face for squared loss, or Newton refinement on that face for logistic loss, polishes the reference. Full-space KKT conditions and all inactive margins are checked again after polishing; no support reference enters an acceptance rule. The independent checks compare support and signs, candidate-to-reference distance, the posterior gap bound, and dual feasibility. Tiny objective differences are compared with an absolute roundoff allowance rather than interpreted as exact arithmetic.

For each candidate, the evaluator first obtains the active center Hessian H_0 and κ. It rejects empty-curvature cases and proposals with more than m active coordinates. For nonempty support and $\lambda_{\min}(H_0) > 0$, the six radius proposals before deduplication are

$$b\{1, 2, 4\} \cup \min_{i \in S} |z_i|\{0.02, 0.1, 0.3\}, \qquad b = \max\left\{\frac{3(\|\kappa\|_2 + 10^{-12})}{\lambda_{\min}(H_0)}, 10^{-13}\right\}.$$

Invalid sign-preserving radii are skipped and the remaining proposals are tested in increasing order. Uniform curvature bounds are recomputed at each radius. The first successful radius is retained, and the finite cycle-rate bound is then constructed. A failed center-curvature, radius, residual, or inactive-margin test returns only an inconclusive result. The special empty-support test uses a strict full gradient margin at zero.

J.3 Four different endpoints

The common-trajectory study distinguishes four endpoints. The primal test returns the joint decision of Theorem 2. Gap Safe + face returns the complete-support and reduced-update conclusions of Proposition 16; it does not compute the unnecessary uniform inactive-gradient scan after screening has succeeded. KKT stopping uses

$$R_\infty(x) = \max_i \begin{cases} |\nabla_i h(x) + \lambda \operatorname{sign}(x_i)|, & x_i \neq 0, \\ (|\nabla_i h(x)| - \lambda)_+, & x_i = 0, \end{cases} \qquad R_\infty(x) \leq 10^{-7}.$$

Dual-gap stopping uses the feasible point above and gap at most 10^{-8}. The latter two endpoints guarantee their respective accuracy criteria in exact arithmetic, not a complete support uniformly over problems.

Each joint test is attempted under three schedules: every epoch; powers of two; and a stable-support schedule that waits for three consecutive unchanged supports, then checks at powers of two or at the first three-epoch stabilization. A final check is forced when the common trajectory reaches KKT tolerance 10^{-10}. Trajectories stop after every endpoint has been observed at that accuracy, or at the stronger 10^{-12} tolerance, subject to a 5,000-epoch cap. The forced final check explains non-power-of-two terminal entries in the geometric schedule. No endpoint in this suite is censored.

The same trajectory is replayed to avoid attributing an ordering or path change to a different acceptance rule. The charged time for a policy is the accumulated epoch and full stopping-check time plus all checks that policy would have performed, including unsuccessful trials and the final successful rate construction. Necessary dual work is charged to both the dual-gap endpoint and the screening-plus-face endpoint. If several schedules ask for the same test on the same epoch, one measured call is used to attribute the same standalone cost to each schedule; these are *charged replay times*, not independently timed online solves. Independent online timings are reported separately below.

The primal test succeeds earlier than the screening-plus-face test on 28 trajectories, ties on one, and is later on one. Its median paired epoch ratio is 0.75 relative to that joint baseline, 0.588 relative to KKT stopping, and 0.496 relative to the dual-gap endpoint. These ratios compare different kinds of stopping information in the latter two cases. They must not be read as a common-accuracy solver speedup.

J.4 Quality of the first accepted output

For each primary objective we replay the unscreened trajectory to its first accepted primal witness, then recompute the candidate bounds and compare with the independently polished reference. The energy rate has median 0.946 and interquartile range [0.788, 0.973]. The posterior distance bound divided by the observed candidate-to-reference distance has median 1.85 [1.51, 2.13]. For objective

Table 8: Decision endpoints on matched trajectories. Times are charged replay times relative to that objective's KKT endpoint. Entries are medians [interquartile ranges] across 30 objectives. For KKT and dual-gap endpoints, "Calls" counts stopping checks; the full KKT check is already included in epoch time.

Endpoint / schedule	Epoch	Relative time	Calls	Check fraction
KKT stopping	35.0 [12.0, 64.8]	1.00 [1.00, 1.00]	35.0 [12.0, 64.8]	0.00 [0.00, 0.00]
Dual-gap stopping	42.0 [15.0, 81.5]	1.25 [1.19, 1.30]	42.0 [15.0, 81.5]	0.03 [0.02, 0.06]
Primal, every epoch	17.5 [9.2, 33.8]	0.94 [0.83, 1.19]	17.5 [9.2, 33.8]	0.38 [0.24, 0.54]
GS+F, every epoch	23.5 [12.0, 46.8]	0.84 [0.64, 1.04]	23.5 [12.0, 46.8]	0.06 [0.04, 0.12]
Primal, geometric	32.0 [16.0, 60.8]	0.96 [0.81, 1.24]	6.0 [5.0, 7.0]	0.08 [0.04, 0.20]
GS+F, geometric	32.0 [16.0, 61.5]	0.97 [0.77, 1.42]	6.0 [5.0, 7.8]	0.02 [0.01, 0.04]
Primal, stable support	32.0 [16.0, 60.8]	0.93 [0.63, 1.18]	4.0 [2.0, 5.0]	0.06 [0.02, 0.18]
GS+F, stable support	32.0 [16.0, 61.5]	0.97 [0.76, 1.42]	4.0 [3.0, 6.0]	0.01 [0.01, 0.03]

gaps above 10^{-13}, the analogous bound-to-gap ratio is 2.11 [1.78, 3.13]. These comparisons assess conservatism and do not establish exact arithmetic for the references.

We also take one additional numerical cyclic pass. Among 27 cases with initial gap above 10^{-12}, the observed next-gap ratio has median 0.501 [0.086, 0.751]. The remaining three are not used to estimate a ratio from roundoff-scale objective differences. Every computed next gap obeys the energy bound within the stated 2×10^{-13} absolute comparison allowance, and all candidate-to-reference distances satisfy the posterior enclosure check with a 5×10^{-12} allowance. The next iterate also lies inside the accepted ball to 2×10^{-12}. These are checks of a particular numerical trajectory, not a replacement for the exact-update proof or an assertion that all the rate bounds are tight.

J.5 Independent online runs

Each of the 30 objectives is solved under five policies: unscreened cyclic updates with the stated scalar tolerance; Gap Safe; and primal support fixing under the three schedules above. Every policy stops at the same *full-space* KKT tolerance 10^{-7}. Full stopping checks continue after coordinates have been removed. The run cap is 5,000 epochs. Each objective/policy pair is timed three times after JIT warmup, giving 450 timing runs. The median within each pair is used before taking ratios and interquartile ranges across objectives. Ratios are descriptive; the objectives reuse datasets and some random seeds and are not 30 independent draws from a specified task population.

The Gap Safe policy checks at powers of two and at termination. It removes a coordinate only when that coordinate is both currently zero and safely screened. Removal is conservative; no claim is made that this is the fastest screened working-set strategy. The primal policies fix the full proposed zero set after the first accepted witness. Before acceptance, every attempted radius and rate computation is charged. At termination a final test is attempted if none has yet succeeded. Timings include prediction initialization, column norms, scalar updates, order-independent setup, full stopping checks, failed and successful tests, and subsequent reduced updates. Training-data preprocessing and compilation are excluded equally. The full candidate-space gradient used by stopping is not omitted from the reported costs.

We monitor the Euclidean distance from every post-acceptance cycle iterate to the accepted ball center. All observed distances stay within the recorded radius in these runs. Each primal schedule accepts on 29 of 30 objectives before or at the common KKT endpoint, giving 261 accepted runs among the 270 primal timing runs. The remaining objective, low-correlation logistic seed 3, reaches KKT tolerance before acceptance and stays unscreened. It is not counted as a false support decision. On the more accurate matched trajectory the same case eventually passes. Agreement with the reference support and this trajectory observation do not turn an approximate scalar solve into an exact minimizer or prove invariance for all possible numerical implementations. No result depends on assuming such a transfer.

Table 9: Online paired time ratios by group, medians [interquartile ranges] over five objectives. The unscreened time is one for each objective.

Group	Gap Safe	Primal every	Primal geometric	Primal stable
Lasso, $c = .10$	0.59 [0.57, 0.65]	1.18 [0.98, 1.21]	1.02 [0.96, 1.43]	0.74 [0.67, 0.77]
Lasso, $c = .65$	0.44 [0.43, 0.49]	1.02 [0.83, 1.04]	0.74 [0.64, 0.77]	0.67 [0.65, 0.74]
Logistic, $c = .10$	0.49 [0.44, 0.51]	1.15 [0.86, 1.49]	1.14 [0.95, 1.15]	1.08 [0.96, 1.10]
Logistic, $c = .65$	0.34 [0.29, 0.36]	0.80 [0.63, 1.01]	0.95 [0.94, 0.96]	0.94 [0.94, 0.94]
Digits–ReLU	0.44 [0.38, 0.46]	0.96 [0.93, 1.05]	1.04 [1.02, 1.07]	1.03 [1.00, 1.03]
Diabetes–ReLU	0.31 [0.27, 0.35]	1.54 [1.33, 1.73]	1.04 [0.84, 1.04]	1.03 [0.84, 1.05]

J.6 Regularization-path costs

For each of the five low-correlation lasso and five low-correlation logistic instances, we solve a six-point path with $\lambda/\lambda_{\max} \in \{0.60, 0.45, 0.35, 0.25, 0.18, 0.12\}$, warm-starting only the coefficient vector from the previous solution. Unscreened, Gap Safe, and primal-geometric policies use the same full KKT target 10^{-7} at each point. Three repeats give $10 \cdot 6 \cdot 3 \cdot 3 = 540$ point-level timed runs. We sum times along each path before taking the median over repeats and forming paired ratios.

Screening and certificate state are reset at every new regularization parameter. Previously screened coordinates are restored before the next solve. Inactive margins, residuals, radii, and local rate bounds are not transferred as valid certificates to another objective. The standardized data remain fixed; all candidate-dependent checking costs, including failures and terminal checks, are charged anew. The implementation also recomputes the starting predictions and column norms within each timed point. Thus the experiment includes warm-start amortization without assuming free certificate reuse. All point-level runs reach the common KKT target. The primal policy accepts on 57 of 60 path points, giving 171 acceptances over 180 timing runs. At the final strength 0.12 for logistic seeds 0, 2, and 4, the KKT target is met before the test accepts; these nine runs are retained with all failed checking costs.

Table 10: Six-point warm-started paths. Ratios compare total timed path cost, not isolated final checks. Entries are medians [interquartile ranges] over ten paths, after three timing repeats per path/policy.

Policy	Relative path time	Calls	Failed	Check fraction
Unscreened cyclic	1.00 [1.00, 1.00]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.00 [0.00, 0.00]
Gap Safe	0.54 [0.48, 0.59]	28.0 [24.0, 31.8]	7.5 [3.5, 9.8]	0.05 [0.01, 0.10]
Primal, geometric	1.05 [1.02, 1.08]	20.0 [14.2, 27.0]	14.5 [8.2, 21.8]	0.27 [0.12, 0.48]

Gap Safe is substantially faster in this controlled implementation. The joint primal test should be invoked for its additional support, enclosure, and dynamics information, not advertised as the best generic solve-time policy. More elaborate dual points, screening schedules, or compiled production solvers could widen or change the observed differences.

K Numerical boundary tests with rational ground truth

We test the distinction between an analytic certificate and its numerical evaluation on objectives specified in exact rational arithmetic. Let $\lambda > 0$, $a > 0$, and $0 \le \mu \le 1$, and define

$$H = \begin{pmatrix} (1+\mu)/2 & (1-\mu)/2 & 0 \\ (1-\mu)/2 & (1+\mu)/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad x^0 = (a, 1, 0), \quad b = Hx^0 + (\lambda, \lambda, \lambda - \delta).$$

The objective is $x^\top Hx/2 - b^\top x + \lambda\|x\|_1$. For $\mu > 0$, its unique minimizer is $(a, 1, \max\{0, -\delta\})$ for the tested parameter ranges. For $\mu = 0$, the positive first-two-coordinate face contains multiple minimizers, so a claimed unique-support certificate is invalid. The ordinary candidate is $(a + e, 1 - e, 0)$. We also set its first coordinate to zero, reverse its sign, or add a third-coordinate coefficient e to probe erroneous supports.

Three penalty scales $\lambda \in \{0.1, 1, 10\}$ are crossed with four families. Small-coefficient tests take $a \in \{10^{-1}, 10^{-3}, 10^{-6}, 10^{-9}, 10^{-12}, 10^{-15}\}$ and $e/a \in \{0.001, 0.05, 0.2\}$. Inactive-margin tests take $\delta/\lambda \in \{10^{-1}, 10^{-4}, 10^{-8}, 10^{-12}, 10^{-15}, 0, -10^{-12}, -10^{-6}\}$ and $e \in \{10^{-6}, 10^{-4}, 10^{-2}\}$. Curvature tests take $\mu \in \{10^{-1}, 10^{-4}, 10^{-8}, 10^{-12}, 10^{-15}, 0\}$ with the same three perturbations. Residual-boundary tests take radii

$$r = 2\sqrt{2}\, e(1+\tau), \quad \tau \in \{-10^{-2}, -10^{-6}, -10^{-10}, -10^{-14}, 0, 10^{-14}, 10^{-10}, 10^{-6}, 10^{-2}\}.$$

Here the radius is first evaluated to 100 decimal digits and then treated as that exact rational truncation; no transcendental comparison is delegated to double precision. The four support/sign variants and three penalty scales yield 1,044 cases.

For any proposed support, its restricted eigenvalue is known rationally: it is μ when both first coordinates are active, and otherwise the smallest selected diagonal entry. With $\kappa = H_{SS} z_S - b_S + \lambda \operatorname{sign}(z_S)$, the strict residual test is evaluated rationally as $4\|\kappa\|_2^2 < \mu_S^2 r^2$. The exact inactive condition uses $\lambda - |\nabla_j h(z)| > 0$ and

$$(\lambda - |\nabla_j h(z)|)^2 > r^2 \|H_{jS}\|_2^2.$$

Together with the sign radius and positive curvature, these give the exact truth value of the selected sufficient-condition test. All model coefficients, candidates, and radii used for ground truth are rational. The float evaluator receives their rounded binary64 values; the study therefore includes input-rounding effects as well as gradient and eigenvalue calculations.

The guarded evaluation requires

$$\begin{aligned} \widehat{\mu} &> \gamma \max\{1, \widehat{\lambda}_{\max}(H_{SS})\}, \\ \widehat{\mu} r - 2\|\widehat{\kappa}\|_2 &> \gamma \max\{1, \widehat{\mu} r, 2\|\widehat{\kappa}\|_2\}, \\ \widehat{\delta}_{\text{inactive}} &> \gamma \max\{1, \lambda\}, \qquad \gamma = 10^{-12}. \end{aligned}$$

The hatted quantities are numerical values, not certified upper or lower enclosures. The unguarded comparison sets $\gamma = 0$. This boundary evaluator uses constant quadratic curvature directly, rather than a GLM prediction-interval surrogate, so it tests the numerical decisions in those inequalities, including rounding of the represented inputs.

Table 11: Boundary tests. "Exact pass" means that the rational sufficient conditions hold. "Guarded" uses the scale-aware 10^{-12} margins. A false condition acceptance need not imply an incorrect support conclusion.

Family	Cases	Exact pass	Guarded pass	Guarded false	Unguarded false
Small active coefficient	216	36	24	0	2
Inactive threshold	288	45	27	0	0
Restricted curvature	216	45	24	0	0
Residual–radius boundary	324	36	15	0	9

Of 162 exact passes, the guarded evaluator accepts 90 and refuses 72. It has no false condition, support/sign, or uniqueness acceptances in this grid. The unguarded evaluator accepts 160 cases, 11 of which fail the exact sufficient conditions; none of those 11 is reported as a demonstrated wrong-support solution. These counts exhibit both conservatism and a real distinction between exact and floating-point tests. They do not prove that a fixed guard controls all dimensions, scalings, matrix conditioning, or input errors. A general floating-point implementation requires outward-rounded enclosures through gradients, curvature bounds, eigenvalue lower bounds, and strict comparisons (Rump, 2010). The separate rational squared-loss implementation in Appendix O avoids these rounding operations instead.

L Construction cost, memory, and what is cached

The raw comparison operator and the full neighborhood certificate have different costs. For a fixed candidate support of size s and a_t active nonzeros in row t, each radius trial forms weighted active Gram matrices at cost $O(\sum_t a_t^2)$, plus a full inactive scan of cost $O(\operatorname{nnz}(X))$ for the primal route. A

dense eigenvalue calculation, rate-matrix norm, or box linear solve has cost $O(s^3)$. With q attempted radii, the worst-case trial work is $O(q[\mathrm{nnz}(X) + \sum_t a_t^2 + s^3])$, in addition to candidate setup. Screening plus a face test replaces the inactive scan by the dual-screening computation; it does not need both.

The new evaluator stores the active design block, its center Hessian, predictions, face residual, and active row norms. Beyond the stored input design, dense workspace is $O(ms + m + n + s^2)$. The inactive scan processes 32 columns at a time to avoid materializing a full absolute inactive-design copy. For sparse inputs, active storage can be reduced to the active nonzeros. The older retained ridge-study implementation materializes larger dense temporaries; its timings are not used to assert the smaller memory bound of the streamed evaluator. Dense support growth remains a limitation of both implementations.

Within one call, the prediction vector, gradient, support, signs, row norms, and center Hessian are cached across radius proposals. The interval endpoints and weights depend on the proposed radius, so their curvature matrices and acceptance conditions are recomputed. Upon changing the candidate, predictions, residuals, and center curvature must be refreshed. Along a regularization path, the data and standardization may be shared, but the gradient-plus-penalty residual, sign margins, and certificate must be reevaluated. Our timings charge the implemented refreshes rather than assuming an unimplemented cache.

By contrast, applying the *raw magnitude* cycle operator uses rowwise suffix and prefix accumulations at cost $O(\mathrm{nnz}(X) + m + n)$ and workspace $O(m + n)$, as derived in Appendix F. That operation supports a matrix-free product or a Perron iteration for the raw comparison. It does not produce cancellation-preserving entries, a restricted minimum eigenvalue, a complete-support witness, or the finite curvature rate at that same cost. The study's matrix-free scaling numbers therefore concern only that operator product.

M Separation from all feasible Gap Safe spherical centers

The residual-based example in Appendix I compares one particular dual construction. Proposition 3 makes a different, stronger statement for a smaller inactive margin: it minimizes the screening bound over the entire feasible dual set. The conclusion concerns the usual individual Gap Safe spheres at a fixed primal candidate. It does not exclude an intersection of regions, a nonspherical localization, or a certificate based on further primal or dual structure.

Proof of Proposition 3. Use the unaveraged least-squares representation of (11), with

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad y = (2, a)^\top, \quad a = 1 - \delta, \quad \lambda = 1.$$

The dual problem is

$$\max_{|u| \le 1,\ |v| \le 1} D(u, v), \qquad D(u, v) = 2u + av - \tfrac{1}{2}(u^2 + v^2).$$

The constraints arise from $A^\top(u, v) = (u, v, v)$. A maximizer is $(1, a)$, since $0 < a < 1$. The unique primal minimizer is $(1, 0, 0)$, as proved in Appendix I. Its value and the candidate value are

$$F^\star = \tfrac{3}{2} + \tfrac{1}{2}a^2, \qquad F(z) = \tfrac{3}{2} + \tfrac{1}{2}a^2 + \tfrac{1}{2}\varepsilon^2.$$

For any feasible (u, v), direct expansion yields the exact identity

$$2\big(F(z) - D(u, v)\big) = \varepsilon^2 + (u - 1)^2 + 2(1 - u) + (v - a)^2. \tag{30}$$

The first-coordinate contribution is nonnegative because $u \le 1$, and vanishes at $u = 1$. Both inactive feature columns have norm one and correlation v. Consequently their standard spherical screening expression obeys

$$\begin{aligned} |v| + \sqrt{2\big(F(z) - D(u, v)\big)} &\ge |v| + \sqrt{\varepsilon^2 + (v - a)^2} \\ &\ge \sqrt{\varepsilon^2 + a^2}. \end{aligned}$$

The second inequality is the triangle inequality applied to $(\varepsilon, a) = (\varepsilon, a - v) + (0, v)$. The feasible center $(u, v) = (1, 0)$ attains equality. This proves the infimum formula, including attainment. Screening needs a strict bound below one, so no feasible center succeeds if $\varepsilon^2 + a^2 \geq 1$, equivalently $\varepsilon^2 \geq 2\delta - \delta^2$.

The primal neighborhood has $S = \{1\}$, residual $\kappa = \varepsilon$, and restricted curvature one. With $r = 3\varepsilon$ and $0 < \varepsilon < 1/2$, the ball is sign preserving and $2\varepsilon < r$. The two inactive gradients are the constant $-a$ throughout this face, so their strict margin is $\delta > 0$. All hypotheses of Theorem 2 hold regardless of the relative size of δ and ε. This proves the claimed separation. □

At $\varepsilon = 10^{-3}$ and $\delta = 10^{-9}$, the squared minimum spherical bound is

$$(1 - 10^{-9})^2 + 10^{-6} = 1 + 998 \cdot 10^{-9} + 10^{-18} > 1.$$

Thus the failure is not a floating-point tie. Notice that the center minimizing the screening expression, $(1, 0)$, need not maximize the dual objective. Better objective values always tighten the radius at a fixed primal point, but the center also matters. This is one reason that a dual-gap comparison alone is not a complete ordering of screening regions.

N Parameter decisions and safeguarded inexact continuation

Write $f_S(u) = h(u, 0) + \lambda s^\top u$ on the sign-preserving ball. We assume either the hypotheses of Theorem 2, or those of Proposition 16, so the unique full minimizer lies in the ball and satisfies $\nabla f_S(x_S^\star) = 0$. All curvature and cross-gradient constants below are uniform on that ball.

N.1 A posteriori parameter bounds

For every $x \in \mathcal{B}$, strong monotonicity gives

$$\mu \|x_S - x_S^\star\|_2^2 \leq \langle \nabla f_S(x_S), x_S - x_S^\star \rangle \leq \|\nabla f_S(x_S)\|_2 \|x_S - x_S^\star\|_2.$$

Division proves the b_P bound in (15); the coincident-point case is immediate. The other direction of strong convexity, based at $x_S^\star$, gives

$$\frac{\mu}{2} \|x - x^\star\|_2^2 \leq F(x) - F^\star \leq F(x) - D(\nu)$$

for any feasible dual point. This proves the b_D bound. Either bound can be used, or their minimum if both are already available. The separate gap-based experimental arm deliberately isolates b_D rather than paying for both and choosing the smaller result.

These posterior statements require only membership in the certified face neighborhood, not an exact trajectory. An approximate iterate outside that neighborhood cannot reuse its curvature constant without a new test. A practical implementation therefore rechecks the support, signs, and distance to the certified center before using a posterior bound. This membership check does not turn numerically estimated curvature or margins into rigorous real-arithmetic bounds.

For exact cycles from the certified candidate, $g_k := F(x^k) - F^\star \leq q_E^k \bar{g}_0$, where $\bar{g}_0 = \|\kappa\|_2^2/(2\mu)$. Thus $q_E^k \bar{g}_0 \leq \mu\tau^2/2$ implies the parameter request. Solving for the smallest sufficient integer k gives (16). If the initial bound already meets the target, take $k = 0$. If $q_E = 0$ and it does not, the next cycle is optimal. The formula is an assurance based on a uniform neighborhood bound. A fitted signed local radius cannot replace q_E in this guarantee.

N.2 An invariant region for approximate active updates

Controlling scalar-solve errors is central in inexact coordinate descent (Tappenden et al., 2016). The additional issue here is to preserve a region whose support and curvature have already been certified.

Proposition 17 (Descent and relative scalar residuals)**.** *Let $\mathcal{C} = \{x \in \mathcal{B} : F(x) \leq F(z)\}$ be the certified sublevel set. Start in $\mathcal{C}$, keep $x_Z = 0$, and change active coordinates in their inherited cyclic order. Suppose each update from x to y satisfies $F(y) \leq F(x)$. Then every within-cycle iterate lies in $\mathcal{C}$ and has signs s.*

At the update of active coordinate i, let $\Delta_i = y_i - x_i$ and $\xi_i = \nabla_i h(y) + \lambda s_i$. Assume the same fixed $0 \le \theta < 1/2$ satisfies

$$|\xi_i| \le \theta \mu_i |\Delta_i| \quad \textit{at every active update}. \tag{31}$$

Then each complete active cycle contracts the objective gap by at most q_θ from (17)*. For $\theta = 0$, it reduces to the exact-cycle bound q_E.*

Proof. The boundary estimate in the curvature theorem gives $F(b) > F(z)$ at every relative boundary point of $\mathcal{B}$. If an update from $x \in \mathcal{C}$ reached $y \notin \mathcal{B}$ with $y_Z = 0$ and $F(y) \le F(x)$, the segment from x to y would meet such a boundary point b. Convexity implies $F(b) \le F(x) \le F(z)$, a contradiction. Thus $y \in \mathcal{B}$, and its objective nonincrease places it in $\mathcal{C}$. Induction proves invariance for all updates. The ball's radius condition fixes the signs.

Within the ball, the regularizer is affine. Applying scalar strong convexity at the post-update point gives

$$F(x) - F(y) \ge -\xi_i \Delta_i + \frac{\mu_i}{2}\Delta_i^2 \ge \left(\frac{1}{2} - \theta\right) \mu_i \Delta_i^2.$$

Let $T = \sum_i \mu_i \Delta_i^2$ over one cycle. Summing yields

$$g_k - g_{k+1} \ge \frac{1 - 2\theta}{2} T. \tag{32}$$

At cycle end, later updates may change coordinate i's derivative. The uniform cross-gradient bounds give

$$|\nabla f_S(x_S^{k+1})| \le |\xi| + U_\beta |\Delta|.$$

Since $\xi_i^2 \le \theta^2 \mu_i^2 \Delta_i^2$ and $\mu_i \le \mu_{\max}$,

$$\|\xi\|_2 \le \theta \sqrt{\mu_{\max}} \sqrt{T}, \qquad \|U_\beta |\Delta|\|_2 \le \sqrt{C_\beta}\sqrt{T}.$$

Restricted strong convexity and completion of the square imply $g_{k+1} \le \|\nabla f_S(x_S^{k+1})\|_2^2/(2\mu)$. Combining with (32) gives

$$g_{k+1} \le \frac{(\sqrt{C_\beta} + \theta\sqrt{\mu_{\max}})^2}{\mu(1 - 2\theta)} (g_k - g_{k+1}).$$

Rearrangement proves (17). The denominator is positive and the resulting factor is strictly below one. At $\theta = 0$ the expression is $C_\beta/(\mu + C_\beta)$. □

Objective nonincrease is a genuine additional safeguard. The relative-residual check alone is not used to infer that an arbitrary proposed point lies in the region where the curvature bounds hold. One implementation first accepts only objective-nonincreasing proposals and then refines the scalar solve until (31) holds. A validated implementation must check those inequalities exactly or with rigorous enclosures. The general binary64 GLM runs do not claim such a pathwise guarantee.

N.3 Absolute residual tolerances

Without (31), descent still preserves $\mathcal{C}$. The same argument gives a useful perturbation bound. Scalar strong convexity and Young's inequality imply

$$F(x) - F(y) \ge \frac{\mu_i}{4}\Delta_i^2 - \frac{\xi_i^2}{\mu_i}, \qquad T \le 4\left(g_k - g_{k+1} + \sum_i \frac{\xi_i^2}{\mu_i}\right).$$

Also $\|\nabla f_S(x_S^{k+1})\|_2^2 \le 2\|\xi\|_2^2 + 2C_\beta T$. Therefore

$$g_{k+1} \le \frac{4C_\beta}{\mu + 4C_\beta} g_k + \frac{\|\xi\|_2^2 + 4C_\beta \sum_i \xi_i^2/\mu_i}{\mu + 4C_\beta}. \tag{33}$$

The additive term tends to zero if the scalar residuals tend to zero, and the stable scalar recurrence then gives $g_k \to 0$. A fixed residual tolerance instead permits a nonzero error floor. This inequality concerns the safeguarded region, unlike the generic influence recursion whose valid domain has to be assumed separately. It need not recover the sharp exact-cycle factor when the residuals vanish; Proposition 17 does.

O EXACT-ARITHMETIC CERTIFICATES FOR SQUARED LOSS

O.1 RATIONAL INPUTS AND SUFFICIENT CONDITIONS

The input consists of rational $X \in \mathbb{Q}^{m\times n}$, $y \in \mathbb{Q}^m$, $\lambda > 0$, a rational candidate z, and rational proposals $r, \mu > 0$. The represented objective is exactly

$$F(x) = \frac{\|Xx - y\|_2^2}{2m} + \lambda\|x\|_1.$$

A decimal string is interpreted as the corresponding rational number, not an uncertain real measurement. A binary floating-point value can likewise be converted to its exact dyadic rational, but this proves a claim only for that represented input. No assertion is made about unmodeled input uncertainty. Integer sizes may grow, so the compact experiments are not a bit-complexity or large-scale performance claim.

The checker first computes $H = X^\top X/m$ and $b = X^\top y/m$ exactly. Thus the full smooth part is convex even when H is singular. With $S = \operatorname{supp}(z)$, it checks $r < \min_{i\in S} |z_i|$ and performs unpivoted rational $LDL^\top$ on $H_{SS} - \mu I$. It refuses unless every pivot is strictly positive. For an empty support it instead checks $\|b\|_\infty < \lambda$. All gradient products, squared norms, and comparisons in (22) use arbitrary-precision integer fractions. Floating point is used in the experiment only to propose μ, which is subsequently checked independently.

Proposition 18 (Soundness of rational acceptance)**.** *Every accepted nonempty-support input satisfies the hypotheses of Theorem 2. The exact rational outputs*

$$E_2 = \frac{\kappa^\top\kappa}{\mu^2}, \qquad E_F = \frac{\kappa^\top\kappa}{2\mu}, \qquad \bar{q} = \frac{C_F}{\mu + C_F}, \quad C_F = \sum_{i<j,\ i,j\in S} \frac{H_{ij}^2}{H_{jj}},$$

bound the squared parameter error, objective gap, and exact-cycle contraction, respectively. The rational comparison $E_2 \le \tau^2$ therefore implements a parameter request without taking a numerical square root. The same representation permits exact iteration of the rate budget.

Proof. Positive rational $LDL^\top$ pivots prove $H_{SS} - \mu I \succ 0$ by congruence, hence uniform active strong convexity with modulus μ. Since $\mu, r > 0$, the squared active-residual inequality is equivalent to $2\|\kappa\|_2 < \mu r$. For an inactive coordinate j and $x \in \mathcal{B}$,

$$|\nabla_j h(x)| \le |(Hz - b)_j| + \|H_{jS}\|_2 r.$$

The positive δ_j and its squared comparison in (22) make this strictly less than λ. The sign radius is checked directly. Theorem 2 now gives uniqueness, support, signs, invariance, and the two posterior outputs. Here $\mu_i = H_{ii} > 0$ and $\beta_{ij} = |H_{ij}|$. The squared Frobenius norm C_F bounds the squared spectral norm C_β. Monotonicity of $C \mapsto C/(\mu + C)$ proves the claimed rate upper bound. Every premise of this argument is a rational equality or inequality decided by the checker. □

These sufficient conditions are evaluated in exact arithmetic for rational squared-loss inputs. They do not extend the implementation to interval arithmetic for logistic or Huber loss. A proposed radius or curvature bound may fail even when another would succeed, so an unsuccessful test remains inconclusive about the true support.

O.2 COMPACT RATIONAL TEST FAMILY

For $m \in \{2, 4, 6\}$, form an invertible rational $m \times m$ matrix $M = I + E$ with $E_{ij} \in \{-2, -1, 0, 1, 2\}/(8m)$. Three fixed random seeds are used for each dimension. The first m feature columns equal M; another $3m$ columns equal Mc_j with small rational c_j. Optional ill-conditioning divides the last column of M by 10^3. An optional small-margin column takes $c_j = (1 - 10^{-8})e_1$. With $\lambda = 1/10$, positive active coefficients $x_i^\star = 1 + (i-1)/10$, and zero inactive coordinates, choose

$$y = Mx_S^\star + m\lambda M^{-\top}\mathbf{1}.$$

Then the active optimality equations hold exactly and the inactive correlations are strictly below λ, because $|\mathbf{1}^\top c_j| < 1$. This gives a known unique optimizer with $n = 4m > m$. The known optimizer is used only to assess the resulting bounds, never to make the certificate decision.

The ordinary candidate is a small alternating-sign rational perturbation of this optimizer. Its amplitude is chosen relative to the proposed radius, lower curvature, and row masses. For each of the 36 base cases we also test a large residual, a missing active coordinate, a sign reversal, an extra inactive coordinate, and an invalid curvature proposal. One additional case is the rank-deficient separation with $\varepsilon = 10^{-3}$ and $\delta = 10^{-9}$. A rational rescaling of its two least-squares rows represents twice (11), preserving support and all conclusions. These choices yield 217 inputs and 37 acceptances. Every accepted squared-distance bound exceeds or equals the exactly evaluated error to the known optimizer. Median checker time is approximately 2.0 milliseconds and the largest observed time is 6.7 milliseconds on these compact inputs. These timings include the full rational Gram construction and all checks, but not experiment generation.

For one well-conditioned, ordinary-margin case at each dimension, we perform 12 full cycles with exact scalar minimization and 12 with a three-quarter step toward the exact scalar minimizer. Every objective comparison, support test, region-membership test, and inequality (33) is evaluated rationally, giving 72 complete-cycle checks. For the damped quadratic step,

$$|\xi_i| = \tfrac{1}{3} H_{ii} |\Delta_i|,$$

so Proposition 17 applies with $\theta = 1/3$, in addition to the directly checked objective nonincrease. The exact-update trajectories also satisfy the rational rate bound. These trajectory conclusions are exact for the rational inputs; the larger floating-point studies provide numerical observations.

P Stronger dual points and matched finite decisions

P.1 Prediction extrapolation and safeguards

We use the non-ridge objectives defined in Appendix J. Each common-trajectory run starts at zero and applies the same natural-order cyclic updates at scalar derivative tolerance 10^{-13}. All methods see each completed epoch. Let $v^k = Xx^k$. For $K = \min\{5, k-1\} \geq 1$, let $V = [v^{k-K}, \ldots, v^k]$ contain the most recent $K+1$ available predictions and form

$$U = [v^{k-K+1} - v^{k-K}, \ldots, v^k - v^{k-1}], \qquad c = \frac{(U^\top U)^{-1}\mathbf{1}}{\mathbf{1}^\top (U^\top U)^{-1}\mathbf{1}}, \qquad \widehat{v} = \sum_{j=1}^{K} c_j v^{k-K+j}.$$

The use of consecutive post-difference predictions is the one-step-shifted fixed-point extrapolate. The coefficient formula and prediction-to-dual construction follow the extrapolation principle in Massias et al. (2020); the implementation does not include Celer's working-set or proximal-Newton strategy (Massias et al., 2018).

The Gram matrix is rescaled by its largest absolute entry. Extrapolation is skipped if it is zero, singular, or has numerical condition number at least 10^{12}, or if normalizing the coefficients is unstable. These refusals retain the ordinary residual point. The proposed dual point is

$$\widehat{\nu} = \frac{-\nabla \ell(\widehat{v})}{m \max\{1, \|X^\top \nabla \ell(\widehat{v})/m\|_\infty / \lambda\}},$$

where $\nabla \ell(v) = v - y$ for squared loss and $\sigma(v) - y$ for logistic loss. An additional factor $1 + 64\epsilon_{\mathrm{mach}}$ is applied to the denominator, as in the residual baseline. For logistic loss, $q = y - m\widehat{\nu}$ remains in $[0,1]^m$ because the rescaling interpolates y and $\sigma(\widehat{v})$. The chosen point maximizes the numerical dual objective among the current residual point, the extrapolated point if available, and the previous chosen point. Linear feasibility and the logistic conjugate domain are checked. The chosen dual objective is therefore no worse numerically than the current residual objective; this does not imply inclusion of every previous spherical screening region.

Both GS+F and GS–DE+F apply the same active radius, curvature, residual, and sign test as the primal route after screening every currently proposed zero. Their primal information is identical. Reference solutions, statistical labels of the synthetic support, local fitted rates, and future iterates do not enter any acceptance test. The output records include the first accepted epoch, full checking time, unsuccessful checks, attempted radii, dual feasibility, extrapolation attempts, and accepted extrapolates.

P.2 Results and interpretation

On the fixed trajectory with residual-based screening, P is earlier in 28 of 30 cases, tied in one, and later in one. With improved dual points, P is earlier in 17, tied in three, and later in ten; its median paired epoch ratio is approximately 0.909. All 90 joint decisions agree in support and signs with the independent references. Thus improved dual points materially reduce the primal test's advantage in decision epoch. There is no basis for describing the residual-only comparison as a universal advantage over screening.

The separation in Proposition 3 remains relevant despite these reversals. It isolates a specific information loss of isotropic dual localization, while the learning study shows that improved dual centers can be decisive on other instances. A proof of one separation and a benchmark win count answer different questions. Neither is used to claim that all screening methods or all primal neighborhood tests are ordered by strength.

Q Online parameter requests and remaining-cycle budgets

Q.1 One downstream task and six policies

The task is to return the complete support together with the requested Euclidean parameter-error bound. Three nested tolerances $10^{-2}, 10^{-4}, 10^{-6}$ are evaluated along each run. Four methods actually target that joint task. The primal route checks Theorem 2 after each epoch. GS+F and GS–DE+F use residual and extrapolated dual points, respectively, and remove currently zero variables once they have been screened. Accumulated safe removals are retained. After the complete face test succeeds, each freezes the certified support and checks b_P. The fourth method, GS–DE with gap/curvature, uses the same support decision but checks b_D instead. It continues paying for the improved dual point and full dual correlations after support fixing.

The two controls stop only at the original accuracy targets: unscreened cyclic descent uses a full KKT tolerance 10^{-7}, and screened extrapolated-dual descent uses gap 10^{-8}. They are not said to certify the parameter request. Their actual parameter errors are compared with the polished reference only after stopping. In particular, the lack of a global positive curvature constant makes a raw dual gap insufficient to translate directly into a global parameter bound; the face test supplies the missing local information.

The decision experiments use one BLAS/OpenMP thread on an AMD EPYC 9V74, with the same Python and numerical-library versions as the ridge-regularized study. Each policy starts at zero, uses natural-order updates and the same 10^{-13} scalar derivative tolerance, and has a 5,000-cycle cap. Data generation, JIT compilation, reference polishing, and comparisons with the reference iterate are outside the timing boundary. Initialization, norm calculations, prediction maintenance, gradient scans, unsuccessful and successful radius trials, eigenvalues, rate construction, dual extrapolation, screening, membership checks, and stopping checks are inside. Policy order is permuted by a fixed seed in each timing repetition. There are six policies, 30 objectives, and three repetitions, giving 540 runs. The nested parameter observations and single control endpoints give 1,260 records. Repeated timings and nested thresholds are not independent optimization problems. The table takes a median within each objective/policy configuration before forming paired ratios across objectives.

The final online parameter requests are all met by the three b_P routes. At $\tau = 10^{-6}$, their median checking-inclusive times are approximately 48.4 milliseconds for P, 16.2 for GS+F, and 18.5 for GS–DE+F; these are separate medians, so their ratios need not equal medians of paired time ratios. The paired ratios relative to P are 0.438 and 0.476 for the two screened routes. These outcomes favor screening for cost even on a task that requires the additional face information.

Q.2 Unmet precision requests and accuracy-only controls

The gap/curvature route meets all coarse and intermediate requests but only 28 of 30 fine requests. The two instances are logistic correlation 0.1, seed 2, and digit-ReLU seed 4. All three repetitions of each reach the 5,000-cycle cap. Their median total times are 1.65 and 1.61 seconds, respectively. The stored final gaps, lower curvature bounds through the accepted certificate, parameter bounds, KKT residuals, and reference errors distinguish this numerical precision limit from a failure to find

the optimizer. The fine-target runtime summary is restricted to 28 paired completions, with the 28/30 denominator and capped cases explicitly retained. No failed endpoint is silently dropped or counted as success.

A production implementation can switch to the smaller available posterior bound, and the successful residual routes demonstrate why that can help. The gap-only parameter arm is an ablation of the source of the error bound, not a claim that screening inherently fails on those two instances. All large-instance certificate evaluations remain numerical.

At the KKT control's stopping point, 13 of 30 reference parameter errors exceed 10^{-6}, with maximum 4.60×10^{-6}. At the extrapolated dual-gap control's stopping point, 11 of 30 exceed it, with maximum 3.16×10^{-4}. These are not incorrect outputs for those controls' stated objectives. They show why a requested coefficient accuracy needs a problem-specific conversion rather than an uncalibrated stopping threshold.

Q.3 Remaining cycles at the first primal decision

For each unscreened trajectory, we retain the first accepted primal ball and compute N_τ from its μ, q_E, and $\bar{g}_0$. The observed count is the number of later numerical cycles until the independent reference distance meets the same tolerance. It is never used to construct the bound. The main text reports the two tighter tolerances; the full set is below. Median first-candidate q_E is 0.946. The large

Table 12: Additional cycles after the first primal decision. Medians [interquartile ranges] across 30 objectives. The analytic budget assumes exact cycles; observed counts use the stated numerical scalar solver. The last column records numerical count exceedances, not a rigorous guarantee for the approximate trajectory.

Parameter target	Budget N_τ	Observed cycles	Exceeded
0.01	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0/30
0.0001	7.5 [0.0, 100.0]	1.0 [0.0, 4.2]	0/30
$1e-06$	238.0 [22.5, 546.2]	14.0 [5.0, 31.2]	0/30

difference between 238 assured cycles and 14 observed cycles at 10^{-6} shows that the bound is not a useful point predictor in this suite. A signed local fit may explain observed behavior, but substituting it would remove the finite-neighborhood guarantee.

The rational experiments remove the numerical-trajectory qualification on a compact class. Relative to the initial exact parameter error, reductions $10^{-2}, 10^{-4}, 10^{-6}$ require respectively $(2, 4, 6)$ exact cycles at $m = 2$, $(2, 3, 5)$ at $m = 4$, and $(2, 4, 5)$ at $m = 6$. The corresponding rational upper budgets are $(6, 11, 15)$, $(5, 9, 13)$, and $(4, 7, 10)$. Thus the budget has a concrete exact operational meaning, although it remains conservative and is not presented as a wall-clock improvement.

R Complete candidate-test procedure

The following analytic procedure separates proposal generation from acceptance. Its rational backend replaces numerical evaluations by the exact tests in Appendix O. For the GLM backend, every accepted output has the numerical-witness interpretation stated in the main text.

Algorithm 1 Analytic candidate test for a sparse GLM

Require: Candidate z, objective parameters, finite sets of trial radii
Set $S = \mathrm{supp}(z)$, $Z = [n] \setminus S$, $s = \mathrm{sign}(z_S)$.
if $S = \varnothing$ **then**
    Return the zero-solution certificate if $\|\nabla h(0)\|_\infty < \lambda$.
    Otherwise return NOTCERTIFIED.
end if
Compute $\kappa = \nabla_S h(z) + \lambda s$ and the center predictions.
for each trial ball radius $0 < r < \min_{i \in S} |z_i|$ **do**
    Compute prediction intervals and uniform bounds (19)–(21).
    if $\mu > 0$, $2\|\kappa\|_2 < \mu r$, and every inactive bound is $< \lambda$ **then**
        Retain the support, invariant set, posterior bounds, and q_E.
    end if
end for
for each trial box radius $0 < r < |z_S|$ **do**
    Recompute the uniform bounds; skip the trial if any $\mu_i \le 0$.
    Set $b_i = |\kappa_i|/\mu_i$.
    if $\mu_i > 0$, $b + \Gamma r < r$, and every inactive bound is $< \lambda$ **then**
        Retain the box certificate, its componentwise bounds, and K.
    end if
end for
Return all retained certificates, or NOTCERTIFIED if none passed.

S DATA AND SCOPE

The supplement contains the implementations and complete experimental results.

Data and broader considerations. The experiments use synthetic data and existing public datasets. No new human-subject data were collected, and no trained model is proposed for clinical or other high-stakes use. Classification accuracy and regression error serve only as consistency diagnostics, not as claims of predictive or clinical validity. WDBC and the original digits data are distributed under CC BY 4.0 by UCI; the diabetes data are distributed through scikit-learn (Wolberg et al., 1993; Alpaydin & Kaynak, 1998; Scikit-learn developers, 2026). These optimization guarantees do not establish fairness, statistical variable-selection validity, or suitability for a downstream decision.

Scope of the contribution. The cycle comparison, Gauss–Seidel local derivative, and generic inexact recursion are established mechanisms specialized here. The contribution is a candidate-centered support enclosure, a curvature test that remains applicable beyond absolute contraction, and finite-candidate sufficient conditions for sparse learning. We do not claim priority for finite identification, global exact-coordinate convergence, or local linear convergence. The controlled screening comparisons favor Gap Safe for solve time; they are not benchmarks of tuned production solvers. Exact-arithmetic conclusions apply to the represented rational squared-loss inputs. Outward-rounded evaluation for general GLMs remains separate work.