LEARNING TO SEARCH BEYOND LINES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study derivative-free nonconvex optimization through reusable conditional responses rather than isolated search directions. Building on the conditional restrictions of convolutional optimization with convex kernels and power lifting, we construct search paths from approximate conditional minimizers. Ideally these paths are conjugate curves: on quadratic objectives they reduce to the conjugate lines of Powell's parallel-subspace construction, and elsewhere they follow the curved valley floor. In one scalar instantiation, power lifting and convolution with a convex kernel give a convex surrogate for nonlocal proposals. We couple this construction to adaptive sampling and error-aware reuse of previously sampled profiles, separating surrogate accuracy from improvement in the original objective. Under explicit regularity assumptions, we derive approximation bounds for profile reuse, evaluation guarantees for safeguarded one-dimensional search, query bounds that charge response learning together with path search, and an exact price for straightening a curved response, and show that on flat curved valleys segment-descending line searches need unboundedly many steps, at a sharp rate, where a conjugate-curve search in an aligned chart needs a constant number of evaluations. Controlled experiments isolate the effects of path construction, convexification, and reuse under matched budgets; on structured curved valleys, curved paths need less than half the evaluations of secant paths that estimate conjugate lines. At larger budgets on the standard COCO/BBOB suite, affine responses beat straight search within the same algorithm, with no resolved further gain from curvature; in value-only training of a diagonal linear network, conditional responses beat straight coordinate search at large budgets in the flattest valleys created by weight decay. The resulting framework connects nonlocal path search with reusable function information while accounting for the cost of acquiring and maintaining that information.

1 INTRODUCTION

Many objectives contain coordinates that compensate for one another. Variable projection exploits this in separable least squares by eliminating linear coefficients conditional on the nonlinear parameters (Golub & Pereyra, 1973; van Leeuwen & Aravkin, 2021). Curved valleys show the same structure geometrically: when one coordinate moves, others must adjust to stay near the valley floor, and a straight search direction soon leaves the valley. We ask whether a search that sees only function values can learn this compensation cheaply enough to profit from it. Conditional-minimizer curves and their implicit derivatives are classical (Section 2). What we add is to charge for learning them from function values, to carry their errors into guarantees on the original objective, and to identify when acquiring a response, searching along it, and bending it each repay their cost.

Search along conditional responses. A chart splits a low-dimensional search space into a baseline coordinate t and navigation coordinates s. Our search learns the conditional minimizer $s^*(t)$ from paid evaluations, freezes the fitted response as a path, and searches the objective along that path with a scalar method. The ideal path is a *conjugate curve*: its tangent is Hessian-conjugate to the navigation subspace (Section 3). Conditional optimal-value restrictions and their response derivatives motivate the scalar COCP schemes of Lu (2025, Section 2.5); here they define a search geometry.

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.

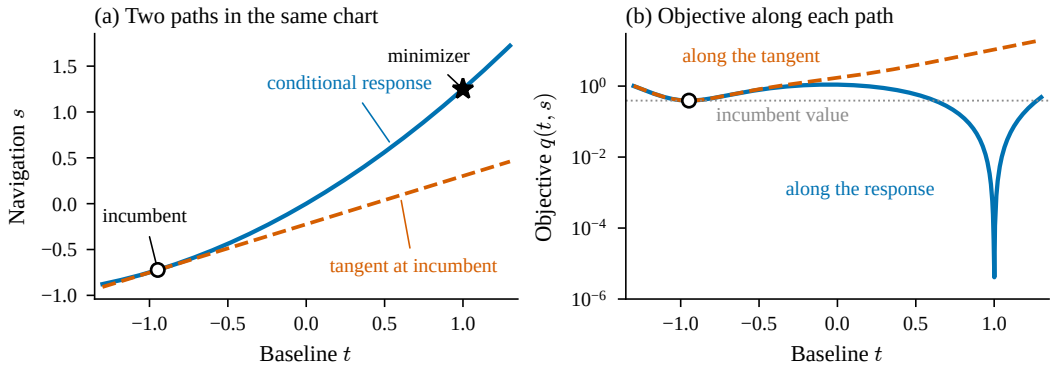

Figure 1: $q(t, s) = (t^2 - 1)^2 + 0.1(t - 1)^2 + 12(s - p(t))^2$ with $p(t) = t + t^2/4$: along the exact tangent at $t_0 = (-1 - \sqrt{0.8})/2$ the objective stays at or above the incumbent value, while the response curve reaches zero (Proposition A.17).

One construction, three geometries. For a quadratic objective the conjugate curve is a straight line, and its secants are the directions given by the parallel-subspace property behind Powell's conjugate-direction method (Powell, 1964; Proposition 1). Away from quadratic behavior the response bends, and the second derivative of the objective along a learned path splits exactly into Schur curvature, a conjugacy-defect cost, and residual acceleration (Equation (4)), and an exact identity prices the straightening of a curved response, whose worst case no affine response lowers by more than a factor of four (Proposition 2). Straight lines, conjugate lines, and conjugate curves are thus nested cases of one construction, and the decomposition states what each of them ignores. Bending pays where valleys are flat and curved: there, line searches whose accepted segments stay in the current sublevel set need a number of steps growing as the fourth root of the flatness, a sharp rate for this class, whereas a conjugate-curve search in an aligned chart needs at most 13 evaluations (Theorem A.16).

Contributions. Our main contribution is the conditional-conjugate framework, which turns paid conditional responses into search paths. An instance chooses a chart policy, a response model, a scalar kernel, a descent safeguard, and a reuse rule; ours combine Brent's method, golden-section bracketing, a branch-and-bound envelope, and a coordinate poll with the COCP kernel of Lu (2025), whose power-lifted convex convolution proposes nonlocal points. Specifically:

- *A unifying geometry* that contains Powell's parallel-subspace construction as its quadratic case, decomposes departures from it exactly, prices straightening, and separates curves from segment-descending line searches (Section 3).
- *An error–cost analysis.* Response and slope errors bound conjugacy defects and path curvature, paid conditional gaps yield chart certificates, fully charged budgets include response learning (Theorem 5), and a poll gives stationarity for any learned path (Sections 4–5).
- *Evidence at several levels.* On structured curved valleys, curved paths need less than half the evaluations of secant paths that estimate conjugate lines, and profile reuse cuts certification cost; COCP proposals improve on matched bisection on multimodal profiles. In the flattest valleys of a diagonal linear network, conditional responses, affine and curved alike, beat straight search at large budgets; on COCO/BBOB affine responses do, with no resolved further gain from curvature (Section 6).

2 Related Work

Line and model-based search from function values. Directional and model-based derivative-free optimization (DFO) are surveyed by Conn et al. (2009) and Larson et al. (2019). Powell's method and random pursuit search along lines (Powell, 1964; Stich et al., 2013); BOBYQA learns local quadratic models (Powell, 2009; Cartis et al., 2019); PRAXIS already searches quadratic curves through past iterates (Brent, 1973).

Elimination and curved paths. Variable projection studies conditional elimination and reduced curvature (Golub & Pereyra, 2003; Ruhe & Wedin, 1980), and inexact projection couples inner accuracy to outer progress (van Leeuwen & Aravkin, 2021). Block-coordinate methods minimize one block with the others fixed (Wright, 2015). Conditional-minimizer curves are branches of Newton trajectories of the chart restriction q, and of F itself when the navigators span the complement $r^{\perp}$ of a direction r ($P_r \nabla F = 0$ with induced tangent $P_r H x' = 0$, P_r the projector onto $r^{\perp}$); such trajectories trace reaction paths and valley floors with derivatives (Quapp et al., 1998; Hirsch & Quapp, 2004; Quapp, 2003), profile traces in likelihood inference (Bates & Watts, 1988; Chen & Jennrich, 2002), and best-response maps in hyperparameter learning (MacKay et al., 2019; Lorraine et al., 2020). Curved search paths also appear in negative-curvature methods (Moré & Sorensen, 1979; Goldfarb, 1980), in geodesic acceleration for least squares (Transtrum et al., 2011), and in recent DFO tensor and projection-based pattern searches (Wang & Zhu, 2024; Jia et al., 2026). We learn conditional-minimizer curves from paid function values and search along them, which links curved search to conjugate directions and variable projection.

Global scalar search, convolution, and tracking. Piyavskii–Shubert methods bound a function with a supplied Lipschitz constant (Piyavskii, 1972; Shubert, 1972), DIRECT avoids fixing it (Jones et al., 1993), and quadratic underestimation connects to αBB (Androulakis et al., 1995) and smooth global search (Breiman & Cutler, 1993; Sergeyev, 1998; Kvasov & Sergeyev, 2009; Dym, 2020); our BB kernel uses an endpoint-chord envelope with a derivative-Lipschitz bound. Gaussian homotopy relates to convex envelopes (Mobahi & Fisher, 2015), and exponential Gaussian smoothing plus a quadratic term gives a convex risk-averse objective under a matrix condition (Dvijotham et al., 2014); COCP integrates a convex kernel against nonnegative, objective-dependent weights (Lu, 2025). Consensus-based optimization moves particles toward Gibbs-weighted centers (Pinnau et al., 2017; Carrillo et al., 2018; Fornasier et al., 2024); for a fixed measure, a squared-distance kernel gives that mean and our absolute-value kernel a median. Time-varying optimization tracks moving solutions (Simonetto et al., 2020; Kamijima et al., 2025), and our reuse rule transports cached value intervals under stated drift bounds (Supplement S4).

3 Conditional-conjugate search

Charts and conditional responses. We minimize $F : \mathbb{R}^d \to \mathbb{R}$ from function values. A chart $\Phi : I \times S \to \mathbb{R}^d$ maps a compact baseline interval I and a compact navigation domain $S \subset \mathbb{R}^m$ into the search space. Define

$$q(t,s) = F(\Phi(t,s)), \quad b(t) = \min_{s \in S} q(t,s), \quad s^*(t) \in \arg\min_s q(t,s), \quad \gamma^*(t) = \Phi(t, s^*(t)). \tag{1}$$

Our charts are affine, $\Phi(t,s) = x + ta + Cs$, with directions that are supplied or chosen from the history of evaluated points; each chart contains the incumbent. The algorithm learns a response $P \approx s^*$ from paid conditional queries (every evaluation is charged to the budget) and searches the frozen path $\widehat{\gamma}(t) = \Phi(t, P(t))$. The path stays in $x + \operatorname{span}\{a, C\}$ but can bend inside it, since $\widehat{\gamma}'' = CP''$.

Conjugate curves. Assume $F \in C^2$ and, locally in t, a unique interior conditional minimizer with $q_{ss} \succ 0$ (Proposition A.1). The implicit function theorem gives

$$s^{*\prime}(t) = -q_{ss}^{-1} q_{st}, \qquad C^{\top} \nabla^2 F(\gamma^*(t))\, \gamma^{*\prime}(t) = 0, \tag{2}$$

so the tangent of the ideal path is Hessian-conjugate to the navigation subspace. We call γ^* a conjugate curve; Equation (4) below measures how far a fitted path departs from it. The reduced function has the standard Schur-complement curvature (Ruhe & Wedin, 1980; van Leeuwen & Aravkin, 2021)

$$b'(t) = q_t, \qquad b''(t) = q_{tt} - q_{ts} q_{ss}^{-1} q_{st}, \tag{3}$$

so following the response lowers q_{tt} by the nonnegative transverse term $q_{ts} q_{ss}^{-1} q_{st}$. The price is the cost of learning the response, which the framework charges explicitly.

Proposition 1 (Conjugate lines are the quadratic case)**.** *Let $F(x) = \frac{1}{2} x^{\top} H x - g^{\top} x$ with $H \succ 0$, and let C have full column rank. For the chart $\Phi(t,s) = x + ta + Cs$, assume that the unconstrained conditional minimizer lies in the interior of S for every $t \in I$. Then the ideal response*

is affine, and γ^* is a straight line with direction $d = a - C(C^\top HC)^{-1}C^\top Ha$ and $C^\top Hd = 0$. Hence $\gamma^*(t_2) - \gamma^*(t_1)$ is H-conjugate to $\mathrm{span}(C)$: the parallel-subspace property behind Powell's conjugate directions (Powell, 1964) is the secant of this linear conjugate curve.

Indeed, $q_s = 0$ gives the affine $s^*(t) = -(C^\top HC)^{-1}C^\top(H(x+ta) - g)$.

When bending pays. Along a learned path $P \in C^2$, evaluate derivatives at $(t, P(t))$ and let $Q = q_{ss}$ (invertible), $r_c = q_s$, and $\mathcal{D} = q_{st} + QP' = \frac{d}{dt}q_s(t, P(t)) = C^\top \nabla^2 F(\widehat{\gamma}(t))\,\widehat{\gamma}'(t)$. Completing the square gives the exact decomposition

$$f_P'' = \underbrace{q_{tt} - q_{ts}Q^{-1}q_{st}}_{\text{Schur curvature at } (t,P)} + \underbrace{\mathcal{D}^\top Q^{-1}\mathcal{D}}_{\text{conjugacy defect}} + \underbrace{r_c^\top P''}_{\text{residual acceleration}} . \tag{4}$$

For fixed P'' and $Q \succ 0$, the conjugate slope $P' = -Q^{-1}q_{st}$ minimizes f_P'', and the nonnegative defect is exactly the excess. On the exact response $\mathcal{D} = 0$, $r_c = 0$, and the first term is b'' of (3); the baseline-parallel line ($P' = P'' = 0$) has $f_P'' = q_{tt}$. A fitted curve pays for its own slope and residual errors, which Supplement S2 bounds. Straightening has an explicit price.

Proposition 2 (The cost of straightening)**.** *Let $q(t,s) = b(t) + \frac{1}{2}\|s - s^*(t)\|_Q^2$, where $\|v\|_Q^2 = v^\top Qv$, with constant $Q \succ 0$ and quadratic s^*. The line through the center anchor with the anchor secant slope, $P(t) = s^*(0) + t\,[s^*(h) - s^*(-h)]/(2h)$, equals the tangent at $t = 0$ and satisfies $q(t, P(t)) - b(t) = \frac{1}{8}t^4\|s^{*\prime\prime}\|_Q^2$, whereas the quadratic response through the anchors at $t \in \{-h, 0, h\}$ is exact. Every affine L has $\max_{|t|\le h}[q(t, L(t)) - b(t)] \ge \frac{1}{32}h^4\|s^{*\prime\prime}\|_Q^2$, with equality for $L = P + \frac{1}{4}h^2 s^{*\prime\prime}$. For $q \in C^4$ with a nondegenerate interior response (Proposition A.1), with Q and $s^{*\prime\prime}$ taken at $t = 0$, the identity holds up to $O(h^5)$ for $|t| \le h$.*

Up to the path-search gap, a conjugate line thus provably keeps the reduced decrease $b(0) - \min_{|t|\le h} b$ at step length h when $h^4\|s^{*\prime\prime}\|_Q^2/8$ is small against it, a sufficient condition, and a fitted curve does when its value loss τ is (Appendix A.9). On locally near-quadratic landscapes $s^{*\prime\prime} \approx 0$, so conjugate curves and lines agree to leading order. On flat curved valleys the curve wins. For $F_\mu(t,s) = A(s - t^2)^2 + \mu(1-t)^2$, straightening costs At^4 against a reduced decrease of order μt; from (t_0, t_0^2), $t_0 \le -1$, every method whose accepted segments stay in the current sublevel set (exact local line searches included) needs at least $|t_0|(A/\mu)^{1/4}/(\sqrt{8}(1-t_0)^{1/2})$ iterations to reach $F_\mu \le \mu$, a sharp order, while the aligned-chart curve search needs at most 13 evaluations (Theorem A.16; compare Jarre, 2013).

A structured running example. Our structured family has $r = d/2$ blocks and

$$F(y,z) = \frac{1}{r}\sum_{i=1}^{r}\{\varphi_i(y_i) + \kappa\psi(e_i)\}, \quad e_i = z_i - p_i(y_i) - \lambda y_{i-1}, \quad \psi(e) = e^2 + A_\psi \sin^2(\omega e). \tag{5}$$

Indices are cyclic, with $\lambda = 0$ for one block. The chart for baseline i contains every navigator its baseline affects: $S_i = \{i, i+1\}$ when coupled and $S_i = \{i\}$ otherwise. The exact reduced profile is $b_i(t) = C_i + \varphi_i(t)/r$. The optimizer receives derivative bounds and this dependency structure; the responses p_i stay hidden.

4 Instances: Learning, Searching, and Reusing Paths

Algorithm 1 lists the components an instance supplies, and Table 1 summarizes our three instances. The *certified instance* receives charts, dependency relations, and regularity bounds and reports objective-gap certificates; the *black-box instance* selects charts from the history of evaluated points, fits responses empirically, and safeguards descent with a classical poll; Section 6.3 takes its charts from a model's architecture.

Scalar kernels and two error measures. A scalar kernel searches a frozen profile f on an interval. On a cell $[u_j, u_{j+1}]$ of width w_j, let ℓ_j interpolate the midpoints of the endpoint value intervals, whose radii are η_j. With a derivative-Lipschitz bound M,

$$|f(u) - \ell_j(u)| \le e_j := Mw_j^2/8 + \max(\eta_j, \eta_{j+1}),$$
$$B_j(u) = \ell_j(u) - \tfrac{M}{2}(u - u_j)(u_{j+1} - u) - \max(\eta_j, \eta_{j+1}). \tag{6}$$

Table 1: Instances of Algorithm 1 in the experiments.

Instance	Charts	Response model	Scalar kernel	Guarantee
Certified (§6.1)	supplied, with bounds	3 anchors, quadratic	COCP or BB	gap certificate
Black-box (§6.4)	sweep and history	3 Brent anchors	Brent	poll stationarity
Architecture (§6.3)	factor pairs (u_i, v_i)	3 exact parabolas	Brent	poll stationarity

Algorithm 1 Conditional-conjugate search; each instance supplies the bracketed components.

Require: Evaluated incumbent x, budget, optional bounds and safeguard $\delta, \sigma > 0$.

1: **while** budget remains **do**
2: [chart policy] Select a bounded chart $\Phi(t, s) = x + ta + Cs$.
3: [response model] Acquire paid conditional anchors and fit a feasible response P; when a certificate is requested, refine P within the caps.
4: [scalar kernel] Freeze P and search the true restriction $F(\Phi(t, P(t)))$ within a budget. Keep every evaluated point.
5: Let y be the best evaluated point, including x.
6: [safeguard] If enabled and $F(x) - F(y) < \sigma\delta^2$, poll $y \pm \delta e_i$ around this fixed y until sufficient decrease or a complete unsuccessful poll.
7: [reuse rule] Set x to the best observation; update radius, history, errors, and cache, charging every evaluation.
8: **end while**
9: **return** Best evaluated point, full cost, and the certificates actually achieved.

Thus $L_f = \min_j \min B_j \le \min f \le U_f$ for any fresh evaluated value U_f, and the gap $U_f - L_f$ measures accuracy on the original profile. The *BB* (branch-and-bound) kernel samples the cell with the lowest bound. The *COCP* kernel, a variant of Lu (2025) with height $h = \exp(-f)$, power $\beta > 0$, and the absolute-value kernel, alternates these safeguard samples with minimizers of the convex surrogate

$$K_\beta(z) = \int_A |z - u| \exp[-\beta f(u)]\, du, \tag{7}$$

which are weighted medians over the active support A. The envelope brackets this density and hence its median (Proposition A.10), so surrogate accuracy and objective accuracy are measured separately; cells that cannot improve the incumbent leave the proposal support but stay in the certificate.

Learning and refining a path (certified instance). At three baseline anchors per interval, conditional scalar searches return paid navigation coordinates and their achieved gaps, and a quadratic response is fitted. Supplied response regularity turns anchor gaps into a full-interval response-value loss τ, and derivative bounds certify the curvature of the fitted path (Section 5). If the certificate is too weak, the algorithm first tightens anchor accuracy and then splits the interval (Corollary A.8 motivates this order). Each leaf is searched separately, so smoothness is needed only within leaves. A path stays frozen during its scalar search, and recalibration creates a new path. Multiple navigators are handled componentwise, which requires supplied conditional separability; Algorithm S1 gives the complete policy.

History charts (black-box instance). Each cycle first sweeps the coordinates with bounded Brent searches. The chart then takes the sweep displacement as baseline a and the previous cycle's displacement, orthogonalized to a, as navigator. Three anchors, each a nine-call Brent search over s, give paid responses at $t \in \{-r_t, 0, r_t\}$. The *straight* control keeps $P = 0$ and skips the anchors; the *affine* control uses the center response and the endpoint secant; the *curved* control interpolates all three. Each path receives the same 25-call Brent search. If a cycle decreases F by less than $\sigma\delta^2$, a coordinate poll of radius δ around the fixed best point follows, and a failed poll halves δ. A cycle costs at most $27d + 52$ evaluations and $O(d)$ memory for its geometry (Supplement S2.3).

Reusing changing profiles. A cached value interval $[l_u, v_u]$ for an unchanged scalar parameterization can be transported to a changed profile. Suppose an absolute drift bound $|f_{\text{new}} - f_{\text{old}}| \le A_0$

and a centered bound

$$|[f_{\text{new}}(u) - f_{\text{new}}(a)] - [f_{\text{old}}(u) - f_{\text{old}}(a)]| \le \rho(u) \tag{8}$$

hold, for example Lipschitz in the context parameters. A fresh paid value at a, with $\delta_a = f_{\text{new}}(a) - f_{\text{old}}(a)$ and both values at a exact, then gives

$$[l_u - A_0, v_u + A_0] \cap [l_u + \delta_a - \rho(u), v_u + \delta_a + \rho(u)]. \tag{9}$$

Each interval contains the new value, hence so does their intersection (Proposition A.12); centered transport removes a vertical offset. Every refresh is charged, and accumulated drift or a change of chart can require fresh queries.

5 Guarantees and their scope

Certified statements assume valid supplied bounds and exact real arithmetic; the float64 implementation, without directed rounding, is only tested numerically. Proofs are in Appendix A and Supplement S1.

The path's own curvature counts. A twice differentiable frozen path has $f_P'' = \widehat{\gamma}'^{\top}\nabla^2 F\,\widehat{\gamma}' + \nabla F^{\top}\widehat{\gamma}''$, so $|f_P''| \le L_F V^2 + G_F A_\gamma$ when $\|\nabla^2 F\| \le L_F$, $\|\nabla F\| \le G_F$, $\|\widehat{\gamma}'\| \le V$, and $\|\widehat{\gamma}''\| \le A_\gamma$: a certified bound includes the path's own acceleration.

Proposition 3 (Paid response and path errors imply a chart certificate)**.** *Suppose leaves I_ℓ cover I, and feasible frozen paths satisfy $0 \le q(t, P_\ell(t)) - b(t) \le \tau_\ell$ on each leaf. If scalar search gives $U_\ell - \eta_\ell \le \min_{I_\ell} q(t, P_\ell(t))$, then*

$$L = \min_{\ell}(U_\ell - \eta_\ell - \tau_\ell) \le \min_{t,s} q(t, s).$$

For the best evaluated value U in this chart, including its initial incumbent F_0, the test $U - L \le \varepsilon + \xi(F_0 - U)$ with $\xi \ge 0$ implies

$$F_0 - U \ge \frac{F_0 - \min_{t,s} q(t, s) - \varepsilon}{1 + \xi}. \tag{10}$$

The certificate combines the response loss τ_ℓ with the path-search gap.

Proposition 4 (Static safeguarded scalar query bound)**.** *On a fixed interval I, assume a valid derivative-Lipschitz bound M, noiseless exact evaluations, a constant-size initialization covering the interval endpoints, retained samples without eviction, and a non-exhausted budget. A scalar routine with at most one extra proposal between middle-half minimum-envelope safeguards establishes objective gap $\epsilon > 0$ in $O(1 + |I|\sqrt{M/\epsilon})$ evaluations.*

BB shares this order, and the safeguards protect the objective search independently of the proposals.

Stationarity for arbitrary learned paths. The black-box instance safeguards descent with a classical coordinate poll with forcing constant $\sigma > 0$ (Kolda et al., 2003; Conn et al., 2009; Vicente, 2013): in exact arithmetic with unlimited iterations, if F is bounded below with an L-Lipschitz gradient, a subsequence of unsuccessful poll centers satisfies $\|\nabla F(y_k)\| \le \sqrt{d}(L/2 + \sigma)\delta_k \to 0$ for any proposed paths (Proposition S7); certified COCP profiles add envelope and mass-integration costs (Supplement S1.3).

Response certificates and a cold-start budget. For the running family, an anchor gap g implies a position error of at most $\sqrt{rg/\kappa}$ because $\psi(e) \ge e^2$. Three paid anchors and a supplied third-derivative bound then control the fitted residual and its first two derivatives by $E_{0,j}, E_{1,j}, E_{2,j}$. While ideal responses and fitted paths stay in the navigator domain, these paid accuracies give valid full-segment bounds on the response loss τ and the path curvature (Proposition A.7). Under a navigator-domain margin and sufficient caps, Theorem S1 bounds the evaluations with which an exact-real first sweep attains every chart certificate, response learning at every refinement depth included; with absolute stopping it returns $F - F^* \le r(\tau + \eta)$, and Theorem A.14 turns chart certificates into a conditional contraction of a complete cyclic sweep. A stronger navigation assumption separates the price of learning a response from that of searching along it.

Theorem 5 (Paid response acquisition and path search). *Let $q(t,s) = b(t) + \kappa\psi(s - p(t))$ on $I \times J$ with one navigator, $|I| = \ell$, $|J| \le W$, $|b''| \le M_b$, $|p'''| \le R$, $\psi(0) = \psi'(0) = 0$, $|\psi''| \le M_\psi$, and $K = \kappa M_\psi > 0$. Suppose every conditional profile is strictly unimodal with minimizer $p(t)$ at distance at least $\rho > 0$ from the ends of J. With exact values and these bounds, golden-section anchors, a piecewise quadratic response, and safeguarded COCP or BB search along it return an evaluated value within $\epsilon \in (0, K\rho^2]$ of $\min q$ using*

$$O\Big[\ell\sqrt{M_b/\epsilon} + \big(1 + \ell(KR^2/\epsilon)^{1/6}\big)\big(1 + \log_+(W\sqrt{K/\epsilon})\big)\Big] \tag{11}$$

queries: the first term pays for path search, the second for response acquisition.

The factor $(KR^2/\epsilon)^{1/6}$ counts the leaves this construction uses; it vanishes for quadratic responses, and the acquisition bound is then logarithmic in the stiffness K. The theorem is an exact-real golden-section instance (proof in Appendix A.11), not a bound for the nine-call Brent anchors of the black-box instance.

6 Experiments

Each confirmation run used source and a protocol frozen by file hash beforehand; follow-up cohorts are marked. Every objective evaluation (FE), including response learning and refreshes, is charged.

6.1 Structured problems: geometry, kernel, and reuse

Protocol. Six settings of (5) in 2D and 8D, with quadratic or sinusoidal responses and rippled conditional penalties, compare variants of the certified instance: COCP with absolute ($\xi = 0$) or relative ($\xi = 1$) stopping, relative without reuse, relative with fitted tangents or, in a follow-up cohort, anchor secants in place of curves, and BB with absolute or relative stopping. All receive the same charts, dependency closure, and valid bounds; exact responses only place a common inferior well and score results. Code fixed on a four-seed pilot is confirmed on 32 new seeds per setting at budgets 256 and 1024, with target $F \le 10^{-4}$ and misses costing the cap; intervals resample the 32 seed clusters (95% percentile, unadjusted; Supplement S3).

Curves beat secants and tangents; BB beats COCP under tight bounds. At budget 1024, COCP relative reaches the target in 189/192 cases, its secant control in 185/192, and its tangent control in 91/192 (Table 2). The secant, re-centered on each subinterval, passes through its center anchor with the slope of the outer anchors and estimates the conjugate line. The curve lowers the paired capped cost by 279.58 FE [260.89, 300.81] against the secant and 427.77 FE [408.29, 448.34] against the tangent, and with BB by 247.87 FE [229.74, 268.50] against the secant: curved paths need less than half the evaluations of secant paths. These are effects of complete certified strategies: as Proposition 2 predicts, the secant carries a larger certified response loss (median 1.38 against 9×10^{-6}) and so needs more splits (15.6 against 3.3 per run); only on the sine 2D setting is it not slower. Under these tight bounds BB is cheaper than COCP by 32.70 FE [31.16, 34.22]. At 10^{-6}, tangents hit 54/192 cases, curves 22/192, and secants 7/192 (joint-chart search and external solvers: Supplements S3.7 and S3.9).

Table 2: Held-out confirmation on 192 paired setting/seed combinations. Success means a real 10^{-4} observation. Capped first-hit means and seed-cluster intervals use the 1024-budget runs, retaining all misses. Secant rows: follow-up cohort (Supplement S3.6).

Method	Success_{256}	Success_{1024}	Capped hit FE	95% interval
COCP absolute	159/192	192/192	246.59	[243.92, 249.13]
COCP relative	151/192	189/192	236.98	[229.70, 245.22]
COCP relative, cold	151/192	190/192	236.96	[230.03, 244.71]
COCP relative, tangent	44/192	91/192	664.76	[644.01, 686.95]
COCP relative, secant	78/192	185/192	516.56	[496.52, 539.15]
BB absolute	160/192	192/192	212.46	[210.20, 214.65]
BB relative	152/192	192/192	204.29	[197.47, 212.06]
BB relative, secant	77/192	185/192	452.16	[433.58, 473.53]

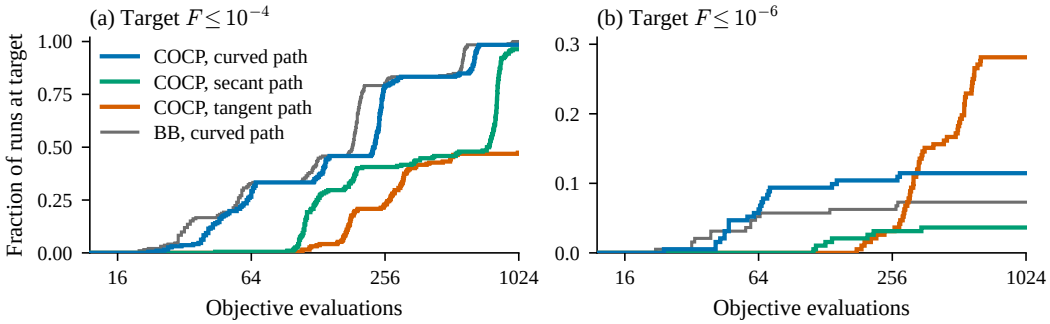

Figure 2: Absolute-target data profiles (Moré & Wild, 2009) over six settings and 32 held-out seeds: fraction of runs with a paid value at or below the target (secant: follow-up cohort, same instances). Absolute-stopping and no-reuse variants track their relative counterparts (Supplementary Figure S2).

Reuse under changing profiles. On 12-stage moving rippled profiles, full-profile reuse and a warm scalar-certificate control both certify all 192 stages per regime. Under slow, drifting, and shock changes, full profiles cut the total cost by 68.6%, 30.8%, and 61.0% with BB and by 66.1%, 27.2%, and 60.3% with COCP (Supplement S3.8); under shocks, first-hit progress improves by 3.5 to 4.1%, so there the savings lie in certification.

6.2 The COCP kernel on multimodal profiles

Each COCP method is paired with a control that lacks only the COCP proposal (Supplement S7). On 600 one-dimensional restrictions of the multimodal BBOB functions f15–f24, sampled COCP with the weighted-median proposals of (7) lowers the mean $\log_{10}$ normalized gap by 0.16 to 0.35 at 8 to 64 evaluations (intervals excluding zero after Bonferroni correction), while DIRECT is stronger from 32 evaluations on. With valid curvature bounds loosened up to 64-fold, BB stays cheaper by 2.8 to 6.8 evaluations, a margin that grows with looseness against the prediction fixed in advance.

6.3 Value-only training of a diagonal linear network

A diagonal linear network with weight decay (Woodworth et al., 2020) fits $w = u \odot v$ by minimizing $F(u, v) = \frac{1}{2n}\|y - X(u \odot v)\|^2 + \frac{\lambda}{2}(\|u\|^2 + \|v\|^2)$. Its minimum is the Lasso optimum (Hoff, 2017), and the gap splits exactly into the Lasso gap of $u \odot v$ and the imbalance $\frac{\lambda}{2}\sum_i(|u_i| - |v_i|)^2$ along the curved valleys $u_i v_i = w_i$, which only the weight decay tilts. Charts pair u_i with v_i; as F is quadratic in v_i given the rest, three values give each conditional parabola and one paid evaluation at its minimizer gives an anchor (Theorem A.16(c)). On eight problems (four scikit-learn datasets with $d = 2p \in \{20, 26, 60, 122\}$ parameters, $\lambda = c\|X^\top y\|_\infty / n$, $c \in \{0.03, 0.003\}$; Table 3, Supplement S8), curved and affine response paths improve on the straight path of the same instance by 1.0 in mean $\log_{10}$ gap at budget $1000d$ (problem-cluster interval excluding zero) and 0.3 at $300d$ (curve interval reaching $+0.03$), after trailing at $100d$ while acquisition is repaid. The gain comes from the flattest valleys, where the straight path leaves a median 98% of its gap in the imbalance. A follow-up control that pays the anchors and skips the path search improves on the straight path by 0.77 at $1000d$; against it, the path search adds 0.30 at $300d$ and 0.23 at $1000d$ and removes the imbalance, and affine paths add as much, so this gain does not require curvature. Only the 0.77 and 0.30 are resolved (problem- and dataset-cluster intervals both exclude zero; four datasets, no multiplicity correction). With exact coordinate solves in the sweep, the charts still gain 1.18 $[0.56, 1.72]$ at $c = 0.003$ and $1000d$ (0.50 pooled, interval reaching zero). The closed-form balance of Hoff (2017), BOBYQA, and PRAXIS are stronger on this model.

6.4 Standard COCO/BBOB evaluation

The frozen black-box instance (straight, affine, and curved geometries with shared Brent, radius, and poll policies) and the earlier controller v3, which also admits learned response slopes as directions (Supplement S9), follow the COCO procedure (Hansen et al., 2021): all 24 `bbob` functions in

Table 3: Diagonal linear network, Brent coordinate sweep, eight problems, 15 starts: mean $\log_{10} \max(F - F^*, 10^{-8})$, and for the flattest valleys this gap and the mean $\log_{10}$ of its two terms (floored at 10^{-12}). The anchors-only control skips the path search (follow-up cohort).

	All problems			$c = 0.003$, $1000d$		
Instance	$100d$	$300d$	$1000d$	Gap	Lasso	Imbalance
Straight path	-3.24	-5.06	-6.20	-4.51	-6.16	-4.55
Affine response	-3.00	-5.37	-7.20	-6.52	-6.56	-9.43
Curved response	-2.98	-5.37	-7.20	-6.60	-6.62	-9.42
Anchors only	-3.17	-5.08	-6.97	-6.04	-7.15	-6.18

Table 4: Fraction of the 51 COCO targets reached at the end of independent runs (24 functions, 15 default instances, 2–40D; last row 40D). Best in bold; details in Supplement S6.

Budget	Straight	Affine	Curve	v3	Powell	PRAXIS	BOBYQA	BIPOP-CMA
$10d$	.054	.054	.054	.100	.120	.095	**.185**	.079
$50d$	.157	.152	.151	.232	**.263**	.245	.257	.181
$200d$	.309	.328	.325	.334	.363	.369	**.371**	.366
$1000d$	.372	.406	.406	.454	.504	.497	.510	**.638**
$200d$, 40D	.230	.231	.231	.214	.224	.201	**.244**	.209

2–40D, default instances (1–5, 71–80), the official observer, and COCO restarts, at independent budgets from $10d$ to $1000d$ ($300d$ and $1000d$ as a follow-up cohort). Baselines are coordinate Brent, Powell, PRAXIS, BOBYQA, and IPOP/BIPOP-CMA-ES (166,320 runs, plus 90,720 in two follow-up studies; Supplement S6).

The hierarchy on a standard suite. Within the matched controller, affine responses (estimated conjugate lines) improve over straight history search at 200, 300, and $1000d$ (mean $\log_{10}$-gap differences -0.20 to -0.36, function-cluster intervals excluding zero; at $50d$ they trail by 0.09). We resolve no further aggregate gain from curved responses (differences to affine at most 0.04, all intervals including zero); this is not an equivalence test. On Rosenbrock (f8–f9, post hoc), where the bound of Theorem A.16 is below one step, curves trail affine responses at 100–$300d$.

Against native solvers. Native solvers remain ahead: Powell is significantly better at all budgets but $200d$, BOBYQA and PRAXIS at 10–$50d$ (BOBYQA also at $1000d$, lower bound 0.005), and restart CMA-ES at 10–20 and 300–$1000d$. At $1000d$ the frozen controller, which stops only at radius 10^{-12}, restarts 0.5 to 0.6 times per run against 5.1 to 12.0 for Powell, PRAXIS, and BOBYQA; a follow-up poll-radius floor of 10^{-4} (Powell's `xtol`) gains 0.11 to 0.12, leaving BOBYQA ahead by 0.96 (interval including zero) and Powell and CMA-ES significantly ahead (Supplement S6). At $200d$ in 20D and 40D, the curved controller uses 30 and 34 μs of CPU per evaluation, BOBYQA 105 and 361 μs (not a time to target).

7 Limitations and Conclusion

Powell's conjugate lines are the quadratic case of conditional-conjugate search, conjugate curves its general case, and the cost of straightening states when bending pays. Acquiring a response, searching along it, and bending it must each repay their cost: bending pays on structured curved valleys, while in the network and on BBOB affine responses already achieve the aggregate gains, with no further benefit from curvature resolved; native solvers lead on BBOB and the closed-form balance, BOBYQA, and PRAXIS on the network. Certificates rely on supplied charts and bounds, and the black-box instance on its poll safeguard. The chart certificate of Proposition 3 is modular in chart, response model, and scalar kernel; cheaper responses, learned charts, and amortized reuse are natural next steps.

AI USE STATEMENT

In this work, we used generative AI tools, under the authors' direction, to help conceive and write parts of the proofs, propose and refine hypotheses and experimental protocols, design experiments, implement methods, baselines, and synthetic objective generators, and interpret results. We have not used generative AI tools for translation, and cleaning or reformatting datasets and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to search, identify, and summarize literature, format references, create figures, and draft, structure, and edit text. We have reviewed all AI-assisted work: the authors checked the proofs, verified code by unit tests, bit-identical replays, and independent recomputation of the reported statistics from raw records, and checked every reference against its source. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Supplement S3 specifies the objective generator, seeds, budgets, supplied information, scoring, and accounting of the structured studies. The anonymous supplementary archive contains the frozen source, the protocol and file hashes recorded before each confirmation run, per-run records, and analysis and verification scripts for the structured studies, with their complete query records and the secant control (Supplement S3); the standard COCO/BBOB evaluation (Supplement S6), whose observer output regenerates from the frozen source; the COCP study (Supplement S7); and the diagonal-linear-network studies (Supplement S8). For the earlier controller studies (S5, S9–S12) it contains the archives that fit the upload limit and SHA-256 hashes of the others, and a script checks every file against the recorded hashes. Appendix A and Supplement S1 give the proofs and distinguish exact-real guarantees from float64 implementations. Protocols were frozen locally by file hash before each confirmation run.

The supplementary material (Sections S1–S12; `supplement.pdf` at the top level of the supplementary archive) contains the remaining appendices, including earlier controller studies (S5, S9–S12); numbers with prefix S refer to it.

REFERENCES

I. P. Androulakis, C. D. Maranas, and C. A. Floudas. αBB: A global optimization method for general constrained nonconvex problems. *Journal of Global Optimization*, 7(4):337–363, 1995. doi: 10.1007/BF01099647.

Douglas M. Bates and Donald G. Watts. *Nonlinear Regression Analysis and Its Applications*. Wiley, 1988. doi: 10.1002/9780470316757.

Leo Breiman and Adele Cutler. A deterministic algorithm for global optimization. *Mathematical Programming*, 58(1–3):179–199, 1993. doi: 10.1007/BF01581266.

Richard P. Brent. *Algorithms for Minimization without Derivatives*. Prentice-Hall, 1973. URL https://maths-people.anu.edu.au/~brent/pub/pub011.html.

José A. Carrillo, Young-Pil Choi, Claudia Totzeck, and Oliver Tse. An analytical framework for consensus-based global optimization method. *Mathematical Models and Methods in Applied Sciences*, 28(6):1037–1066, 2018. doi: 10.1142/S0218202518500276. URL https://arxiv.org/abs/1602.00220.

Coralia Cartis, Jan Fiala, Benjamin Marteau, and Lindon Roberts. Improving the flexibility and robustness of model-based derivative-free optimization solvers. *ACM Transactions on Mathematical Software*, 45(3):1–41, 2019. doi: 10.1145/3338517.

Jian-Shen Chen and Robert I. Jennrich. Simple accurate approximation of likelihood profiles. *Journal of Computational and Graphical Statistics*, 11(3):714–732, 2002. doi: 10.1198/106186002493.

Andrew R. Conn, Katya Scheinberg, and Luis N. Vicente. *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics, 2009. doi: 10.1137/1.9780898718768.

Krishnamurthy Dvijotham, Maryam Fazel, and Emanuel Todorov. Universal convexification via risk-aversion. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 162–171. AUAI Press, 2014. arXiv:1406.0554.

Nadav Dym. Quasi branch and bound for smooth global optimization. *arXiv preprint arXiv:2005.13728*, 2020. URL https://arxiv.org/abs/2005.13728.

Massimo Fornasier, Timo Klock, and Konstantin Riedl. Consensus-based optimization methods converge globally. *SIAM Journal on Optimization*, 34(3):2973–3004, 2024. doi: 10.1137/22M1527805. URL https://arxiv.org/abs/2103.15130.

Donald Goldfarb. Curvilinear path steplength algorithms for minimization which use directions of negative curvature. *Mathematical Programming*, 18(1):31–40, 1980. doi: 10.1007/BF01588294. URL https://link.springer.com/article/10.1007/BF01588294.

Gene Golub and Victor Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems*, 19(2):R1–R26, 2003. doi: 10.1088/0266-5611/19/2/201. URL https://doi.org/10.1088/0266-5611/19/2/201.

Gene H. Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432, 1973. doi: 10.1137/0710036. URL https://epubs.siam.org/doi/10.1137/0710036.

Nikolaus Hansen, Anne Auger, Raymond Ros, Olaf Mersmann, Tea Tušar, and Dimo Brockhoff. COCO: A platform for comparing continuous optimizers in a black-box setting. *Optimization Methods and Software*, 36(1):114–144, 2021. doi: 10.1080/10556788.2020.1808977. URL https://doi.org/10.1080/10556788.2020.1808977.

Michael Hirsch and Wolfgang Quapp. The reaction pathway of a potential energy surface as curve with induced tangent. *Chemical Physics Letters*, 395(1–3):150–156, 2004. doi: 10.1016/j.cplett.2004.07.079.

Peter D. Hoff. Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017. doi: 10.1016/j.csda.2017.06.007.

Florian Jarre. On Nesterov's smooth Chebyshev–Rosenbrock function. *Optimization Methods and Software*, 28(3):478–484, 2013. doi: 10.1080/10556788.2011.638924.

Xiaoxi Jia, Matteo Lapucci, and Pierluigi Mansueto. Projection-based curve pattern search for black-box optimization over smooth convex sets. *Optimization Methods and Software*, 2026. doi: 10.1080/10556788.2026.2647280. URL https://arxiv.org/abs/2503.20616.

D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, 1993. doi: 10.1007/BF00941892.

Tomoya Kamijima, Naoki Marumo, and Akiko Takeda. A simple yet highly accurate prediction-correction algorithm for time-varying optimization. *arXiv preprint arXiv:2504.05798*, 2025. URL https://arxiv.org/abs/2504.05798.

Tamara G. Kolda, Robert Michael Lewis, and Virginia Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, 45(3):385–482, 2003. doi: 10.1137/S003614450242889.

Dmitri E. Kvasov and Yaroslav D. Sergeyev. A univariate global search working with a set of Lipschitz constants for the first derivative. *Optimization Letters*, 3(2):303–318, 2009. doi: 10.1007/s11590-008-0110-9.

Jeffrey Larson, Matt Menickelly, and Stefan M. Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019. doi: 10.1017/S0962492919000060.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1540–1552, 2020. URL https://proceedings.mlr.press/v108/lorraine20a.html.

Zhipeng Lu. Convolutional optimization with convex kernel and power lift. *arXiv preprint arXiv:2503.22135*, 2025. URL https://arxiv.org/abs/2503.22135.

Matthew MacKay, Paul Vicol, Jon Lorraine, David Duvenaud, and Roger Grosse. Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1903.03088.

Hossein Mobahi and John W. Fisher, III. On the link between Gaussian homotopy continuation and convex envelopes. In Xue-Cheng Tai, Egil Bae, Tony F. Chan, and Marius Lysaker (eds.), *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 8932 of *Lecture Notes in Computer Science*, pp. 43–56, Cham, 2015. Springer. doi: 10.1007/978-3-319-14612-6_4. URL https://people.csail.mit.edu/hmobahi/pubs/gaussian_convenv_2015.pdf.

Jorge J. Moré and Danny C. Sorensen. On the use of directions of negative curvature in a modified Newton method. *Mathematical Programming*, 16(1):1–20, 1979. doi: 10.1007/BF01582091. URL https://link.springer.com/article/10.1007/BF01582091.

Jorge J. Moré and Stefan M. Wild. Benchmarking derivative-free optimization algorithms. *SIAM Journal on Optimization*, 20(1):172–191, 2009. doi: 10.1137/080724083.

René Pinnau, Claudia Totzeck, Oliver Tse, and Stephan Martin. A consensus-based model for global optimization and its mean-field limit. *Mathematical Models and Methods in Applied Sciences*, 27 (1):183–204, 2017. doi: 10.1142/S0218202517400061. URL https://arxiv.org/abs/1604.05648.

S. A. Piyavskii. An algorithm for finding the absolute extremum of a function. *USSR Computational Mathematics and Mathematical Physics*, 12(4):57–67, 1972. doi: 10.1016/0041-5553(72)90115-2.

M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, 1964. doi: 10.1093/comjnl/7.2.155.

M. J. D. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical Report DAMTP 2009/NA06, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, August 2009. URL https://www.damtp.cam.ac.uk/user/na/NA_papers/NA2009_06.pdf.

Wolfgang Quapp. A valley following method. *Optimization*, 52(3):317–331, 2003. doi: 10.1080/0233193031000149902.

Wolfgang Quapp, Michael Hirsch, Olaf Imig, and Dietmar Heidrich. Searching for saddle points of potential energy surfaces by following a reduced gradient. *Journal of Computational Chemistry*, 19(9):1087–1100, 1998. doi: 10.1002/(SICI)1096-987X(19980715)19:9<1087::AID-JCC9>3.0.CO;2-M.

H. H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *The Computer Journal*, 3(3):175–184, 1960. doi: 10.1093/comjnl/3.3.175.

Axel Ruhe and Per Åke Wedin. Algorithms for separable nonlinear least squares problems. *SIAM Review*, 22(3):318–337, 1980. doi: 10.1137/1022057. URL https://epubs.siam.org/doi/10.1137/1022057.

Yaroslav D. Sergeyev. Global one-dimensional optimization using smooth auxiliary functions. *Mathematical Programming*, 81(1):127–146, 1998. doi: 10.1007/BF01584848.

Bruno O. Shubert. A sequential method seeking the global maximum of a function. *SIAM Journal on Numerical Analysis*, 9(3):379–388, 1972. doi: 10.1137/0709036.

Andrea Simonetto, Emiliano Dall'Anese, Santiago Paternain, Geert Leus, and Georgios B. Giannakis. Time-varying convex optimization: Time-structured algorithms and applications. *Proceedings of the IEEE*, 108(11):2032–2048, 2020. doi: 10.1109/JPROC.2020.3003156. URL https://doi.org/10.1109/JPROC.2020.3003156.

Sebastian U. Stich, Christian L. Müller, and Bernd Gärtner. Optimization of convex functions with random pursuit. *SIAM Journal on Optimization*, 23(2):1284–1309, 2013. doi: 10.1137/110853613.

Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Physical Review E*, 83(3):036701, 2011. doi: 10.1103/PhysRevE.83.036701.

Tristan van Leeuwen and Aleksandr Y. Aravkin. Variable projection for NonSmooth problems. *SIAM Journal on Scientific Computing*, 43(5):S249–S268, 2021. doi: 10.1137/20M1348650. URL https://epubs.siam.org/doi/10.1137/20M1348650.

Luis N. Vicente. Worst case complexity of direct search. *EURO Journal on Computational Optimization*, 1:143–153, 2013. doi: 10.1007/s13675-012-0003-7. URL https://www.mat.uc.pt/~lnv/papers/complex.pdf.

Peng Wang and Detong Zhu. A derivative-free modified tensor method with curvilinear linesearch for unconstrained nonlinear programming. *Optimization*, 73(3):597–621, 2024. doi: 10.1080/02331934.2022.2124112.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 2020.

Stephen J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015. doi: 10.1007/s10107-015-0892-3. URL https://link.springer.com/article/10.1007/s10107-015-0892-3.

A PROOFS, ASSUMPTIONS, AND THE SCOPE OF THE GUARANTEES

All statements in this appendix concern exact real arithmetic. The implementation evaluates these inequalities in floating point without outward rounding. Curvature, response regularity, and drift bounds are supplied assumptions. Every value used as a current upper bound is a value of the current original objective at an evaluated point. Requested accuracy and achieved accuracy are distinct whenever a local or global budget is exhausted.

A.1 THREE DISTINCT RESPONSE OBJECTS

For a fixed affine chart, write

$$q(t,s) = F(x + ta + Cs), \qquad t \in I, \quad s \in J \subset \mathbb{R}^m, \qquad b(t) = \min_{s \in J} q(t,s). \tag{A.1}$$

There are three different objects. An ideal response $s^*(t)$ minimizes $q(t,\cdot)$. A finite-power COCP response minimizes a convex convolution of a weighted scalar profile and in general differs from $s^*(t)$. The learned response $P(t)$ is fitted to paid, evaluated conditional incumbents and their objective-gap bounds. Objective-gap bounds come from the envelope, while root enclosures serve the surrogate. The prototype computes scalar mass proposals and polynomial fits.

Proposition A.1 (Ideal response and affine-chart conjugacy)**.** *Let F be C^2 on a neighborhood of the chart. Suppose J is compact and, for each $t \in I$, $q(t,\cdot)$ has a unique minimizer $s^*(t)$ in the interior of J, with $q_{ss}(t, s^*(t))$ positive definite. Then s^* is C^1 on the interior of I and*

$$(s^*)'(t) = -q_{ss}^{-1} q_{st} = -(C^\top H C)^{-1} C^\top H a, \qquad C^\top H \gamma'(t) = 0, \tag{A.2}$$

where $\gamma(t) = x + ta + Cs^*(t)$ and $H = \nabla^2 F(\gamma(t))$.

Proof. Uniqueness and compactness imply continuity of the minimizer: every convergent subsequence of minimizers for $t_k \to t$ has a limit minimizing $q(t,\cdot)$, and that limit must be $s^*(t)$. Interior optimality gives $q_s(t, s^*(t)) = 0$. The implicit function theorem applies because q_{ss} is invertible. Its local branch agrees with the continuous minimizing branch, and differentiating this identity gives the first equality. For the affine chart, $q_{ss} = C^\top H C$ and $q_{st} = C^\top H a$. Substituting into $\gamma' = a + C(s^*)'$ proves the last identity. □

Remark A.2 (Regularity and parameterization). Positive definiteness of q_{ss} supplies the nondegeneracy used by the implicit-function argument; uniqueness alone does not make the response differentiable: $q(t,s) = s^4 - ts$ has the unique response $s^*(t) = (t/4)^{1/3}$, which has a vertical tangent at zero. For a nonlinear chart $\Phi(t,s)$, the identity $q_{ss}(s^*)' = -q_{st}$ still holds under the stated nondegeneracy conditions, but q_{ss} and q_{st} include second derivatives of Φ contracted with ∇F. The ambient Hessian-conjugacy identity in (A.2) is specific to an affine chart. For quadratic F its ideal response is affine; for nonquadratic F the path can be nonlinear while remaining in the same fixed affine subspace.

Proposition A.3 (Finite-power convolution responses)**.** *Let $h(t,u) > 0$ be continuous on a compact interval J and differentiable in t, with differentiation under the integral justified. Fix $\beta > 0$ and define $p_{\beta,t}(u) = h(t,u)^\beta / \int_J h(t,v)^\beta \, dv$ with respect to the stated coordinate measure du. Let $\chi_\delta(r) = \mathrm{clip}(r/\delta, -1, 1)$ for $\delta > 0$. The root $m = m(t,\beta)$ of $\mathbb{E}_{p_{\beta,t}} \chi_\delta(m - U) = 0$ is unique, and*

$$\begin{aligned} J_m &= \frac{\mathbb{P}_{p_{\beta,t}}(|U - m| < \delta)}{\delta} > 0, \\ \partial_t m &= -\frac{\beta \, \mathrm{Cov}_{p_{\beta,t}}(\chi_\delta(m - U), \partial_t \log h(t,U))}{J_m}, \\ \partial_\beta m &= -\frac{\mathrm{Cov}_{p_{\beta,t}}(\chi_\delta(m - U), \log h(t,U))}{J_m}. \end{aligned} \tag{A.3}$$

For the absolute-value kernel used in the prototype, the convolution minimizer is instead the weighted median. For a positive continuous density it is unique and its derivative follows the same implicit calculation with denominator $2p_{\beta,t}(m)$ and χ_δ replaced by sign.

Proof. The Huber convolution is convex. Its derivative, divided by total mass, is the displayed expectation, which is nonpositive at the left endpoint and nonnegative at the right endpoint. At its root, strict positivity of the density makes $J_m > 0$, and the derivative is strictly increasing there, giving uniqueness. For a parameter v, differentiating a normalized expectation of a function held fixed in v gives its covariance with $\partial_v \log h^\beta$. Differentiate the root identity, keeping m fixed for that covariance term, and divide by J_m. For the absolute-value kernel, the normalized derivative is $2\mathbb{P}(U \le m) - 1$, whose derivative in m is $2p_{\beta,t}(m)$. □

The fixed-power identities hold between changes of the power or of the improvement mask, and the conjugacy identity (A.2) concerns the ideal response. A change of scalar parameterization multiplies the weighted measure by its Jacobian.

A.2 Curvature of the actual frozen path

Proposition A.4 (Path curvature). *Let $\widehat{\gamma} : I \to \mathbb{R}^d$ and F be C^2. If, on the path, $\|\nabla^2 F\| \le L$, $\|\nabla F\| \le G$, $\|\widehat{\gamma}'\| \le V$, and $\|\widehat{\gamma}''\| \le A$, then*

$$\left| \frac{d^2}{dt^2} F(\widehat{\gamma}(t)) \right| \le LV^2 + GA. \tag{A.4}$$

For a scalar navigator with $f_P(t) = q(t, P(t))$, the sufficient bound is

$$|f_P''| \le B_{tt} + 2B_{ts}V_P + B_{ss}V_P^2 + G_s A_P, \tag{A.5}$$

provided $|q_{tt}| \le B_{tt}$, $|q_{ts}| \le B_{ts}$, $|q_{ss}| \le B_{ss}$, $|q_s| \le G_s$, $|P'| \le V_P$, and $|P''| \le A_P$ on the relevant domain.

Proof. The chain rule gives $(F \circ \widehat{\gamma})'' = \widehat{\gamma}'^\top \nabla^2 F \widehat{\gamma}' + \nabla F^\top \widehat{\gamma}''$. Apply the norm bounds. Similarly, $f_P'' = q_{tt} + 2q_{ts}P' + q_{ss}(P')^2 + q_s P''$ gives (A.5). □

Every bound is for the actual frozen path, including any domain-preserving model adjustment. Hard clipping of a polynomial generally destroys C^2 regularity, so segments are certified separately and smoothness is needed only within each segment. A changed path is a new path, and its values at parameter t are new evaluations.

A.3 Paid anchor gaps imply a uniform response-error bound

Write n for the block count denoted by r in the main text. Consider a conditional chart with separable navigator penalties,

$$q(t, s) = b(t) + \frac{\kappa}{n} \sum_{j=1}^{m} \psi(s_j - r_j(t)), \qquad \kappa > 0, \tag{A.6}$$

on $I = [c-T, c+T]$, $T > 0$. Assume that the ideal responses are feasible, $r_j \in C^3(I)$, $|r_j'''| \le R_{3,j}$, and

$$\psi(0) = \psi'(0) = 0, \qquad \psi(e) \ge e^2, \qquad |\psi''(e)| \le M_\psi. \tag{A.7}$$

The prototype uses $\psi(e) = e^2 + A_\psi \sin^2(\omega e)$, with $A_\psi \ge 0$ and $M_\psi = 2 + 2A_\psi\omega^2$. The growth inequality in (A.7) allows nonconvex ψ.

Lemma A.5 (From conditional value gaps to position errors). *At each $t_k \in \{c - T, c, c + T\}$, suppose a paid scalar solve for navigator j returns $s_{j,k}$ and an achieved gap $g_{j,k}$ over its entire navigator domain. The other navigator coordinates are held fixed in the same chart context. Then*

$$|s_{j,k} - r_j(t_k)| \le \sqrt{n g_{j,k}/\kappa}. \tag{A.8}$$

Proof. By separability, all terms except navigator j contribute a constant to this scalar minimization. Its feasible minimizer is $r_j(t_k)$ and its objective gap is $\kappa\psi(s_{j,k} - r_j(t_k))/n$. Apply $\psi(e) \ge e^2$ and the achieved value-gap bound. □

Lemma A.6 (Three-anchor interpolation with imperfect anchors)**.** *Let $r \in C^3(I)$, $|r'''| \le R_3$, and let Q interpolate three values s_k at $c-T, c, c+T$, with $|s_k - r(t_k)| \le e_a$. For any actual C^2 model P, set $D_\ell = \sup_I |(P-Q)^{(\ell)}|$ for $\ell = 0, 1, 2$. Then $\sup_I |(P-r)^{(\ell)}| \le E_\ell$, where*

$$\begin{aligned} E_0 &= D_0 + \tfrac{5}{4} e_a + \frac{R_3 T^3}{9\sqrt{3}}, \\ E_1 &= D_1 + \frac{4e_a}{T} + R_3 T^2, \\ E_2 &= D_2 + \frac{4e_a}{T^2} + R_3 T. \end{aligned} \tag{A.9}$$

Proof. Translate c to zero. Let Q_* interpolate the exact response values. The Lagrange remainder gives $|r(t) - Q_*(t)| \le (R_3/6)|t(t^2 - T^2)|$; the maximum of the cubic factor is $2T^3/(3\sqrt{3})$. Write $Q_*(t) = r(0) + at + bt^2$. Taylor expansion at the two endpoints yields $|a - r'(0)| \le R_3 T^2/6$ and $|2b - r''(0)| \le R_3 T/3$. Since $|r'(t) - r'(0) - r''(0)t| \le R_3 t^2/2$, their sum is at most $R_3 T^2$ on I.

For the second derivative, the exact identity

$$Q_*'' = \int_{-T}^{T} \frac{T - |u|}{T^2}\, r''(u)\, du$$

expresses Q_*'' as an average under a symmetric triangular distribution. Because r'' is R_3-Lipschitz, $|r''(t) - Q_*''| \le R_3 \mathbb{E}|t - U|$. This expectation is convex in t; its maximum on $[-T, T]$ is attained at an endpoint and equals T.

In normalized coordinates $u = t/T$, the three Lagrange basis functions are $(u^2 - u)/2$, $1 - u^2$, and $(u^2 + u)/2$. The suprema of the sums of their absolute values, first derivatives in t, and second derivatives in t are respectively $5/4$, $4/T$, and $4/T^2$. These bound the propagated anchor error $Q - Q_*$. Add the three bounds for $P - Q$ to finish. □

Proposition A.7 (Uniform value and curvature certificates)**.** *Under (A.6)–(A.7), suppose every $P_j(I)$ lies in its navigator domain. Obtain $E_{\ell,j}$ from Lemmas A.5 and A.6, using $e_{a,j} = \max_k \sqrt{n g_{j,k}/\kappa}$. If $|b''| \le M_b$, then the actual frozen profile $f_P(t) = q(t, P(t))$ satisfies*

$$\begin{aligned} 0 \le f_P(t) - b(t) \le \tau, \qquad & \tau = \frac{\kappa M_\psi}{2n} \sum_j E_{0,j}^2, \\ |f_P''(t)| \le M_P, \qquad & M_P = M_b + \frac{\kappa M_\psi}{n} \sum_j (E_{1,j}^2 + E_{0,j} E_{2,j}). \end{aligned} \tag{A.10}$$

Proof. Integrating $|\psi''| \le M_\psi$ from zero gives $|\psi'(e)| \le M_\psi |e|$ and $\psi(e) \le M_\psi e^2/2$. Write $e_j = P_j - r_j$. The value bound follows by substitution, while

$$f_P'' = b'' + \frac{\kappa}{n} \sum_j \{\psi''(e_j)(e_j')^2 + \psi'(e_j) e_j''\}$$

proves the curvature bound. □

Corollary A.8 (Allocation of conditional accuracy)**.** *Put $a_0 = 5/4$ and $B_j = D_{0,j} + R_{3,j} T^3/(9\sqrt{3})$. The same value bound obeys*

$$\tau \le M_\psi a_0^2 \sum_j \max_k g_{j,k} + \frac{\kappa M_\psi}{n} \sum_j B_j^2. \tag{A.11}$$

Thus $g_{j,k} \le \tau_{\text{target}}/(2m a_0^2 M_\psi)$ allocates at most half the target to anchor uncertainty. When $P_j = Q_j$ and $R_{3,j} = 0$, navigator j contributes at most $21 M_\psi \max_k g_{j,k}/T^2$ to $M_P - M_b$.

Proof. Use $(a_0 e_{a,j} + B_j)^2 \le 2a_0^2 e_{a,j}^2 + 2B_j^2$ and $e_{a,j}^2 = n \max_k g_{j,k}/\kappa$. In the quadratic-response case, $E_{1,j}^2 + E_{0,j} E_{2,j} \le (16 + 5) e_{a,j}^2/T^2$. □

This cancellation concerns the error interface; the work needed to obtain $g_{j,k}$ still scales with the conditional scalar curvature, which is proportional to κ. While response uncertainty is uncontrolled, decreasing T can increase the terms $e_{a,j}/T$ and $e_{a,j}/T^2$; once anchor uncertainty is controlled, halving T reduces the cubic position remainder by a factor of eight. These facts motivate recalibration before geometric subdivision.

A.4 Scalar objective bounds, mass bounds, and safeguarded cost

Lemma A.9 (Objective envelope with bounded observation errors). *Suppose f' is M-Lipschitz on I, and the endpoints $u < v$ have observations v_u, v_v with errors at most η_u, η_v. Let C be their affine interpolant and $\eta = \max(\eta_u, \eta_v)$. Then, for $t \in [u, v]$,*

$$|f(t) - C(t)| \le \frac{M}{2}(t-u)(v-t) + \eta. \tag{A.12}$$

Consequently, the minimum of the lower quadratic envelopes over all covering panels is a lower bound $L \le \min_I f$. A current evaluated incumbent U therefore has original-objective gap at most $U - L$.

Proof. First use exact endpoint values. Subtract their chord and call the result h. The functions $h(t) + (M/2)(t-u)(v-t)$ and $h(t) - (M/2)(t-u)(v-t)$ have respectively nonpositive and nonnegative second derivative almost everywhere; f' being Lipschitz suffices. Concavity and convexity, together with their zero endpoint values, imply the two desired inequalities. The error in an affine interpolation of the endpoint observations is at most the corresponding convex combination of their error bounds, and hence at most η. Taking panel minima gives the conclusion. □

Proposition A.10 (Mass uncertainty encloses a finite-power root). *On each panel let $e_j = M(v-u)^2/8 + \max(\eta_u, \eta_v)$. Fix $\beta > 0$ and a union of active panels A of positive length. Define*

$$w(t) = \mathbf{1}_A(t)e^{-\beta f(t)}, \qquad w_-(t) = \mathbf{1}_A(t)e^{-\beta(C(t)+e_j)}, \qquad w_+(t) = \mathbf{1}_A(t)e^{-\beta(C(t)-e_j)}.$$

Let $Z_- = \int_I w_-$, $Z_+ = \int_I w_+$, and $H(s) = \int_{I\cap(-\infty,s]}(w_- + w_+)$. The derivative $D(s) = \int_I \operatorname{sign}(s-t)w(t)\,dt$ of the absolute-value convolution satisfies

$$H(s) - Z_+ \le D(s) \le H(s) - Z_-. \tag{A.13}$$

Thus every median lies in $\{s : Z_- \le H(s) \le Z_+\}$. Endpoints can be obtained as outward generalized quantiles of the density $(w_- + w_+)/(Z_- + Z_+)$ at probabilities $Z_-/(Z_- + Z_+)$ and $Z_+/(Z_- + Z_+)$.

Proof. Lemma A.9 gives $w_- \le w \le w_+$. Split the derivative into its positive left mass and negative right mass. Lower-bound the former and upper-bound the latter for the lower inequality; reverse these choices for the upper inequality. At a median $D(s) = 0$, yielding the stated enclosure. Generalized quantiles include any zero-density plateau; a positive density on the entire interval has an ordinary unique median. □

An active mask may omit a panel whose valid objective lower bound is at least U, since every point of that panel is at least as bad as the incumbent. The resulting convolution is the convolution of the masked density. Exponential-affine panel masses and their quantiles are computed from existing observations alone, so they cost computation only. The root enclosure controls a convex surrogate, whereas Lemma A.9 controls the original objective; the two bounds serve different tasks.

Theorem A.11 (Static safeguarded scalar query bound). *Fix a scalar objective with M-Lipschitz derivative and a target $\varepsilon > 0$. Assume exact, noiseless current evaluations, a covering initialization of n_0 retained nodes including both endpoints, no eviction, no drift, and no budget or arithmetic termination. Whenever $U - L > \varepsilon$, a safeguard chooses a panel attaining the smallest objective lower bound and evaluates a point in its middle half. Suppose at most p other queries are allowed before each safeguard, including at most p queries in a final unfinished block. For $M > 0$, set $h = \sqrt{8\varepsilon/M}$ and $W = |I|$. The number of safeguards is at most $1 + 4W/h$, and*

$$N_f \le n_0 + (p+1)(1 + 4W/h) + p = O\Big(n_0 + (p+1)\left[1 + W\sqrt{M/\varepsilon}\right]\Big). \tag{A.14}$$

For $M = 0$, endpoint values already certify the affine objective.

Proof. At a safeguard, let its panel have width w. All exact endpoint values are at least the current incumbent U, so its chord is at least U. Therefore $U - L \le Mw^2/8$ for the selected panel. A still-unsatisfied gap implies $w > h$. A point in the middle half is at distance at least $w/4 > h/4$ from every previous node, and thus from every previous safeguard point. At most $1 + 4W/h$ such mutually separated points fit in I. Counting at most p other queries per safeguard plus initialization and a final block proves the claim and rules out indefinite continuation. □

The prototype alternates a mass proposal/refinement opportunity with the same lower-envelope safeguard; in this ideal static regime it has $p = 1$, and the branch-and-bound control has $p = 0$. The theorem gives both the same worst-case order. Uncertain reused nodes, endpoint refreshes, cache eviction, and finite local caps require additional accounting, as do multi-chart and dynamic amortized bounds.

A.5 Transport of cached information across profile drift

Proposition A.12 (Centered drift transport)**.** *Let $q_\theta(s)$ be a family on a fixed scalar interval. Suppose, along the segment joining θ_0 and θ_1, $\|\partial_\theta \partial_s q_\theta(s)\|_* \le K$, where the dual norm corresponds to the norm on θ. Then for a fixed anchor a,*

$$|q_{\theta_1}(s) - q_{\theta_0}(s) - q_{\theta_1}(a) + q_{\theta_0}(a)| \le K\|\theta_1 - \theta_0\|\,|s - a|. \tag{A.15}$$

If an additional bound $|q_{\theta_1}(s) - q_{\theta_0}(s)| \le A$ holds and the old value is in $[l_s, u_s]$, the new value lies in the intersection

$$\begin{aligned} &[l_s - A, u_s + A] \cap [l_s + \delta_a - \rho_s, u_s + \delta_a + \rho_s], \\ &\delta_a = q_{\theta_1}(a) - q_{\theta_0}(a), \qquad \rho_s = K\|\theta_1 - \theta_0\|\,|s - a|. \end{aligned} \tag{A.16}$$

provided both anchor values are exact. Uncertain anchor values must contribute their own interval uncertainty.

Proof. Subtract anchor values and integrate first in s from a to s, then in θ along its line segment. The dual-norm inequality bounds the double integral by the right-hand side of (A.15). Each interval in (A.16) separately contains the new value, so their intersection does too. The old interval width remains present in both. □

For $q_\theta(s) = B_\theta + (\kappa/n)\psi(s - r_\theta)$, one may use $\rho_s = (\kappa M_\psi/n)|r_{\theta_1} - r_{\theta_0}|\,|s - a|$; the vertical term B_θ cancels exactly. Along successive stages, valid drift bounds can be accumulated, at the cost of increased uncertainty. The anchor refresh and every other reevaluation count against the objective budget. An old incumbent value becomes a current upper bound through its transported error or a fresh evaluation. This proposition proves that reuse is valid; its efficiency relative to retaining a previous solution and a scalar certificate is measured in Section 6.

A.6 From frozen paths to a joint-chart certificate

Proposition A.13 (Covering path bounds and relative descent)**.** *Fix the chart context in (A.1). Suppose intervals I_ℓ cover I. On each interval, a frozen feasible path has $0 \le f_\ell(t) - b(t) \le \tau_\ell$ and a paid scalar search returns U_ℓ with achieved path gap η_ℓ, so $U_\ell - \eta_\ell \le \min_{I_\ell} f_\ell$. Then*

$$L_\ell = U_\ell - \eta_\ell - \tau_\ell \le \min_{I_\ell} b, \qquad L_{\mathrm{chart}} = \min_\ell L_\ell \le \min_{t,s} q(t,s). \tag{A.17}$$

Let F_0 be the current objective at chart entry, and let $U \le F_0$ be the best value among it and all evaluated candidates within the same joint chart. If $U - L_{\mathrm{chart}} \le \varepsilon + \xi(F_0 - U)$, for $\xi \ge 0$, then

$$F_0 - U \;\ge\; \frac{F_0 - \min_{t,s} q(t,s) - \varepsilon}{1 + \xi}. \tag{A.18}$$

Proof. Since $f_\ell \le b + \tau_\ell$ pointwise, $\min f_\ell - \tau_\ell \le \min b$, proving the first inequality. The minimum over a covering collection gives the second. Write $D = F_0 - \min q$ and $d = F_0 - U$. Validity of the lower bound implies $D \le F_0 - L_{\mathrm{chart}} = d + (U - L_{\mathrm{chart}}) \le (1 + \xi)d + \varepsilon$, which rearranges to the result. □

The current point may lie off the learned path; it is retained independently as a candidate, and $P(0)$ is kept as fitted. Calibration candidates may also be accepted if they lie in the same fixed-context joint chart. Old and new paths on the same interval may give different valid lower bounds on the same b; their maximum is valid. If a subdivision is interrupted, retaining its covering parent preserves the bound. The chart certificate requires complete interval coverage.

A.7 A STRUCTURAL COMPLETE-SWEEP GUARANTEE

Consider the supplied family

$$F(y,z) = \frac{1}{n}\sum_{j=1}^{n}\left[\phi_j(y_j) + \kappa\psi\big(z_j - p_j(y_j) - \lambda y_{j-1}\big)\right] \tag{A.19}$$

on a compact product box, with cyclic indices for $n \geq 2$ and $\lambda = 0$ for $n = 1$. Assume each ϕ_j is continuous, $\psi \geq 0$, $\psi(0) = 0$, and that $p_j(y_j) + \lambda y_{j-1}$ is feasible in the navigator domain for every allowed y. The baseline chart i changes only y_i and its navigator set S_i, which includes its own navigator and every residual affected by changing y_i. For this family, $S_i = \{i, i+1\}$ when $\lambda \neq 0$ and $n \geq 2$, and $S_i = \{i\}$ otherwise. Each chart searches the entire baseline interval and navigator box. The exact conditional minimum therefore removes all penalties in S_i simultaneously.

Theorem A.14 (Complete sweep from achieved chart gaps)**.** *Visit every baseline once in any fixed order $i_1, \ldots, i_n$. Let x^k denote the accepted point after chart i_k, which changes only its chart coordinates. Suppose the accepted objective has achieved joint-chart gap at most g_k relative to the fixed context at that chart's entry. Then*

$$F(x^n) - F^* \leq \sum_{k=1}^{n} g_k, \qquad F^* = \frac{1}{n}\sum_{j=1}^{n}\min\phi_j. \tag{A.20}$$

The bound uses only the achieved chart gaps g_k.

Proof. The idea is to charge each final residual to the last chart that can change it; dependence closure ensures that no later chart changes either that residual's navigator or a baseline on which it depends. Feasibility of all zero residuals and nonnegativity of ψ establish the expression for F^*. Let $e_j = z_j - p_j(y_j) - \lambda y_{j-1}$ and $\phi_j^* = \min\phi_j$. Outside S_{i_k} every residual is constant during chart i_k. Minimizing that chart eliminates all penalties in S_{i_k} and minimizes ϕ_{i_k}, so its achieved gap bounds

$$H_k = \frac{1}{n}\left[\phi_{i_k}(y_{i_k}^k) - \phi_{i_k}^* + \kappa\sum_{j\in S_{i_k}}\psi(e_j(x^k))\right] \leq g_k.$$

Assign residual j to the last chart $\ell(j)$ whose navigator set contains it. Every residual has such a chart. Complete dependence closure implies that no later chart changes either its navigator or a baseline on which it depends. Hence $e_j(x^n) = e_j(x^{\ell(j)})$. Likewise, every baseline's ϕ value is unchanged after its unique visit. Let $J_k = \{j : \ell(j) = k\} \subseteq S_{i_k}$. The sets J_k partition all residuals, and therefore

$$F(x^n) - F^* = \frac{1}{n}\sum_{k}\left[\phi_{i_k}(y_{i_k}^k) - \phi_{i_k}^* + \kappa\sum_{j\in J_k}\psi(e_j(x^k))\right] \leq \sum_k H_k \leq \sum_k g_k.$$

The inequality discards only nonnegative penalties. □

Corollary A.15 (Conditional contraction for the implemented cyclic order)**.** *Under Theorem A.14, retain the incumbent at every chart and write $d_k = F(x^{k-1}) - F(x^k) \geq 0$. If every chart actually attains $g_k \leq \varepsilon_k + \xi d_k$ for the same $\xi \geq 0$, then*

$$\Delta_+ \leq \frac{\xi}{1+\xi}\Delta_0 + \frac{\sum_k \varepsilon_k}{1+\xi}, \qquad \Delta_0 = F(x^0) - F^*, \quad \Delta_+ = F(x^n) - F^*. \tag{A.21}$$

With $\varepsilon_k = \varepsilon$ and T successive sweeps satisfying these conditions, put $q_\xi = \xi/(1+\xi)$. Then

$$\Delta_T \leq q_\xi^T\Delta_0 + n\varepsilon(1 - q_\xi^T). \tag{A.22}$$

Proof. Theorem A.14 and telescoping descent give $\Delta_+ \le \sum_k \varepsilon_k + \xi \sum_k d_k = \sum_k \varepsilon_k + \xi(\Delta_0 - \Delta_+)$. Rearrange and iterate the affine recurrence. □

For $\xi > 0$, choosing $\varepsilon \le \delta/(2n)$ and $q_\xi^T \Delta_0 \le \delta/2$ suffices for accuracy δ, conditional on obtaining every required certificate. Query complexity additionally charges response learning, path reconstruction, failed attempts, refreshes, and all scalar searches, and fixed local caps must be large enough for the hypotheses to be attained. The theorem covers the stated structured family and applies equally to COCP and BB whenever either supplies the required chart gaps.

A.8 Flat curved valleys separate curves from every segment-descending line search

The next result strengthens the fixed-tangent example below to a lower bound for a whole class of line-search methods, with the response acquisition cost of the curve search counted. It is a fixed-dimension counterpart of Jarre (2013), who shows that on Nesterov's Chebyshev–Rosenbrock function in $\mathbb{R}^n$, started at $(-1, 1, \dots, 1)$, every continuous piecewise linear descent path has at least $0.72 \cdot 1.618^n$ segments before the value drops by 75%. Here the dimension is two, and the bound grows with the flatness A/μ.

Theorem A.16 (Curve–line separation on flat curved valleys)**.** *For $A, \mu > 0$ let $F_\mu(t, s) = A(s - t^2)^2 + \mu(1 - t)^2$, and start on the valley floor at $x_0 = (t_0, t_0^2)$ with $t_0 \le -1$, where $F_\mu(x_0) = \mu(1 - t_0)^2$.*

(a) Consider any method whose iterates satisfy $x_{k+1} = x_k + \tau_k d_k$ with the whole segment $[x_k, x_{k+1}]$ in $\{F_\mu \le F_\mu(x_k)\}$, for example exact local line searches that descend from $\tau = 0$ to the first local minimizer, along arbitrary, adaptively chosen directions. Before some iterate satisfies $F_\mu \le \mu$, it performs at least $|t_0|\,(A/\mu)^{1/4}/(\sqrt{8}\,(1 - t_0)^{1/2})$ iterations; for $t_0 = -1$ this is $(A/(4\mu))^{1/4}/\sqrt{8}$.

(b) For fixed t_0, some method in this class reaches $F_\mu \le \mu$ within $O((A/\mu)^{1/4})$ iterations as $A/\mu \to \infty$, so the order in (a) is sharp.

(c) With the chart $a = e_t$, $C = e_s$ on an interval containing t_0 and 1, three conditional parabolic fits (nine evaluations) recover the ideal response $s^*(t) = t^2$ exactly, one parabolic step on the reduced function $b(t) = \mu(1 - t)^2$ is exact, and at most 13 evaluations reach the global minimizer $(1, 1)$, for every μ.

Proof. (a) On a step segment write $g = s - t^2$. A direction without t-component makes no progress in t. Otherwise the segment is affine in t, so g is quadratic in t with leading coefficient -1; if the step spans $t \in [m - L/2, m + L/2]$, then $g(m - L/2) + g(m + L/2) - 2g(m) = -L^2/2$. On the segment $Ag^2 \le F_\mu \le F_\mu(x_k) \le F_\mu(x_0)$, so $|g| \le w := (F_\mu(x_0)/A)^{1/2}$ and $L^2/2 \le 4w$, that is, $L \le \sqrt{8}\,(1 - t_0)^{1/2}(\mu/A)^{1/4}$ for every step. A point with $F_\mu \le \mu$ has $|1 - t| \le 1$, hence $t \ge 0$, which requires total progress at least $|t_0|$ in t.

(b) Let $F_0 = F_\mu(x_0)$ and $W(t) = ((F_0 - \mu(1 - t)^2)/A)^{1/2}$ for $t_0 \le t \le 1$, so that there $\{F_\mu \le F_0\} = \{|s - t^2| \le W(t)\}$. The function W is nondecreasing and concave, as the square root of a nonnegative concave function, and $W(t)^2 \ge c\,(t - t_0)$ with $c = \mu(1 - t_0)/A$. Place the vertices on the lower boundary $s = t^2 - W(t)$. On the chord from the boundary point at t_k to the one at $t_k + L$, parameterized by $\lambda \in [0, 1]$, $g = -(1 - \lambda)W(t_k) - \lambda W(t_k + L) + \lambda(1 - \lambda)L^2$. Concavity gives $g \ge -W$, and $L^2 \le 8W(t_k)$ gives $g \le W$, so the chord stays in $\{F_\mu \le F_0\}$, and every vertex has value F_0. The first chord leaves x_0 with $L = c^{1/3}$, for which $g \le \lambda L^2 \le (c\lambda L)^{1/2} \le W$. Later chords take $L = \sqrt{8W(t_k)} \ge \sqrt{8}\,(c\,(t_k - t_0))^{1/4}$ until $t = 0$, and a last chord of length $\sqrt{4W(0)}$, which keeps $g \le \frac{9}{16}W(0)$, ends on the valley floor at some $t \in (0, 1]$ once A/μ is large, where $F_\mu = \mu(1 - t)^2 \le \mu$. Counting chords while the progress $t_k - t_0$ lies in $[2^j c^{1/3}, 2^{j+1}c^{1/3})$ gives $O(\log(A/\mu) + |t_0|^{3/4}c^{-1/4}) = O((A/\mu)^{1/4})$ iterations. Numerically, for $t_0 = -1$ and A/μ from 10^6 to 10^{12}, the construction uses between $0.41(A/\mu)^{1/4}$ and $0.35(A/\mu)^{1/4}$ iterations, against the lower bound $0.25(A/\mu)^{1/4}$.

(c) F_μ is quadratic in s with minimizer $s = t^2$, so a parabola through three evaluations at fixed t is exact. The quadratic response through three exact nodes is t^2, and b is quadratic, so parabolic interpolation of its three values is exact. Three fits, three reduced values, and the final evaluation give 13 evaluations. □

The class in (a) restricts the accepted segments, not the information a method may collect, and it counts iterations. The method in (b) keeps every vertex at the initial value, so (b) shows that the order is sharp for this class; whether methods that must decrease the value at every iterate attain it is left open. The class contains coordinate, Powell, and Rosenbrock-type methods (Rosenbrock, 1960) with segment-descending line searches. Bracketing and global line searches that cross a barrier, trust-region steps that test only endpoints, and population samplers form separate classes. Numerically, with $A = 100$ and $t_0 = -1$, coordinate descent with exact local line searches needs 1,544, 15,448, and 154,514 steps for $\mu = 10^{-1}, 10^{-2}, 10^{-3}$, consistent with the bound; the curve search in (c) returns $F_\mu \approx 10^{-27}$ at $\mu = 10^{-6}$ with 13 evaluations. Chart selection is supplied here. Section 5 charges response acquisition and path search within a supplied chart; discovering a suitable chart is not covered by these bounds.

A.9 The cost of straightening

Proof of Proposition 2. For quadratic s^*, $s^*(h) - s^*(-h) = 2hs^{*\prime}(0)$, so the secant line equals the tangent at 0 and $P(t) - s^*(t) = -\frac{1}{2}t^2 s^{*\prime\prime}$. Hence $q(t, P(t)) - b(t) = \frac{1}{2}(P - s^*)^\top Q(P - s^*) = \frac{1}{8}t^4 s^{*\prime\prime\top} Q s^{*\prime\prime}$. Quadratic interpolation through three exact values of a quadratic is exact, so the curve has $P = s^*$ and $q(t, P(t)) = b(t)$.

For an affine L, put $e = L - s^*$ and $M = \max_{|t|\le h} \|e(t)\|_Q$. Because L is affine and s^* is quadratic, $e(-h) + e(h) - 2e(0) = -h^2 s^{*\prime\prime}$, so the triangle inequality gives $h^2\|s^{*\prime\prime}\|_Q \le 4M$, and $\max_{|t|\le h}[q(t, L(t)) - b(t)] = M^2/2 \ge h^4\|s^{*\prime\prime}\|_Q^2/32$. For $L = P + \frac{1}{4}h^2 s^{*\prime\prime}$, $e(t) = (\frac{1}{4}h^2 - \frac{1}{2}t^2)s^{*\prime\prime}$ and $M = \frac{1}{4}h^2\|s^{*\prime\prime}\|_Q$, so equality holds.

For $q \in C^4$ with q_{ss} invertible along the response, $s^* \in C^3$ and $q(t, s) = b(t) + \frac{1}{2}(s - s^*(t))^\top Q(t)(s - s^*(t)) + O(|s - s^*(t)|^3)$ with $Q(t) = q_{ss}(t, s^*(t)) = Q(0) + O(h)$. For $|t| \le h$ the tangent at zero and the centered-secant line each satisfy $P(t) - s^*(t) = -\frac{1}{2}t^2 s^{*\prime\prime}(0) + O(h^3)$, which gives the identity up to $O(h^5)$. (The shifted minimax line instead has residual $(\frac{1}{4}h^2 - \frac{1}{2}t^2)s^{*\prime\prime}(0) + O(h^3)$.) □

Paid anchors and when bending pays. Let navigator j have anchor errors at most δ_j and $|s_j^{*\prime\prime\prime}| \le R_{3,j}$. Lemma A.6 bounds the fitted curve coordinatewise, $|P_j - s_j^*| \le e_j := \frac{5}{4}\delta_j + R_{3,j}h^3/(9\sqrt{3})$ for $|t| \le h$, so $\|P - s^*\|_Q^2 \le \sum_j e_j^2$. In the model of Proposition 2 the curve's value excess is $\frac{1}{2}\|P - s^*\|_Q^2 \le \frac{1}{2}\lambda_{\max}(Q)\sum_j e_j^2$; for $q \in C^4$, a bound on the transverse Hessian near the response replaces Q, up to higher-order terms. The secant line pays $\frac{1}{8}h^4\|s^{*\prime\prime}\|_Q^2$ at the interval ends, and no line has a smaller worst case than $\frac{1}{32}h^4\|s^{*\prime\prime}\|_Q^2$. Since the curve's value excess is $\frac{1}{2}\|P - s^*\|_Q^2$, its worst case is smaller than that of every line when $\sup_{|t|\le h}\|P - s^*\|_Q < \frac{1}{4}h^2\|s^{*\prime\prime}\|_Q$. This comparison must use the Q-norm: small coordinate errors along a stiff direction can cost more than straightening. For $s^*(t) = (t^2, 0)$, $Q = \mathrm{diag}(1, 10^{10})$, and $h = 0.1$, a curve error of 10^{-4} in the second coordinate costs 50, whereas the secant loses at most 5×10^{-5}.

A smaller worst case does not by itself guarantee descent. If a frozen path, curved or straight, has value excess at most τ over b and the path search achieves gap η, the returned value satisfies $U \le \min b + \tau + \eta$ (Proposition 3), so $\tau + \eta < F_0 - \min b$ guarantees improvement over an incumbent with value F_0. The geometric comparison alone does not: for $q(t, s) = a^4(1 - t)^2 + \frac{1}{2}(s - t^2)^2$, incumbent $(0, 0)$, and $P(t) = t^2 + a$, the curve error a is small against the straightening scale, yet $\min_t q(t, P(t)) \ge a^2/2 > a^4 = q(0, 0)$ for small a. These are local comparisons of values at a fixed step length, not evaluation counts; Theorem 5 charges acquisition and search together.

A.10 A curve–tangent separation with explicit approximation tolerance

The following example isolates the value of curvature in a fixed chart.

Proposition A.17 (An exact tangent can forbid descent)**.** *Let*

$$\phi(t) = (t^2-1)^2 + \tfrac{1}{10}(t-1)^2, \qquad p(t) = t + t^2/4, \qquad q(t,s) = \phi(t) + \kappa(s-p(t))^2.$$

Set $t_0 = (-1-\sqrt{4/5})/2$*,* $A_0 = 6t_0^2 - 19/10$*, and* $B_0 = 4t_0$*. If* $\kappa > 16(B_0^2/(4A_0) - 1)$*, then* $(t_0, p(t_0))$ *is the unique global minimizer of* q *along the entire tangent line to* $(t, p(t))$ *at* t_0*. In contrast, the response curve reaches the unique global minimum* $(1, 5/4)$ *with value zero.*

Proof. The derivative is $\phi'(t) = (t-1)(4t^2+4t+1/5)$, so $\phi'(t_0) = 0$ and $A_0 > 0$. Along the tangent, with $u = t - t_0$, the quadratic response gives the exact residual $-u^2/4$. Taylor expansion of the quartic ϕ is exact and gives

$$q(t_0+u, p(t_0)+p'(t_0)u) - q(t_0, p(t_0)) = u^2\{A_0 + B_0 u + (1+\kappa/16)u^2\}.$$

The assumed inequality makes the discriminant of the bracket negative and its leading coefficient positive. Its value is strictly positive for all u, proving the tangent claim. Both summands in ϕ are nonnegative, and they vanish simultaneously only at $t = 1$. The nonnegative transverse penalty then gives the unique global minimizer stated above. □

Corollary A.18 (Robustness to paid response and path errors)**.** *Let a feasible learned path* $(t, P(t))$ *be defined on an interval containing* 1*, with* $\sup|P - p| \le e$*. If a paid path search returns an evaluated value within* η *of that path's minimum, it returns* $U \le \kappa e^2 + \eta$*. Consequently,* $\kappa e^2 + \eta < \phi(t_0)$ *guarantees strict improvement over the tangent incumbent in Proposition A.17.*

Proof. The path's feasible point at $t = 1$ has value $\kappa(P(1) - p(1))^2 \le \kappa e^2$. Apply the achieved path gap. □

All response queries needed to establish e are part of the method's cost. In the special known class $q(t,s) = \phi(t) + a(t)(s - p(t))^2$, $a(t) > 0$, one could instead recover a scalar response exactly from three admissible queries at $s = 0, \pm h$:

$$p(t) = -\frac{h\{q(t,h) - q(t,-h)\}}{2\{q(t,h) + q(t,-h) - 2q(t,0)\}}.$$

Thus a known quadratic p can be recovered from nine such queries, a structure-exploiting response model that complements the COCP calibration routine. The separation explains the geometric mechanism behind curved paths.

In Figure 1 ($\kappa = 12$), the incumbent is a strict local minimizer of q with value $\phi(t_0) \approx 0.390$. Along the response, ϕ rises to about 1.105 at $t = (-1+\sqrt{4/5})/2$ before it falls to zero at $t = 1$, so the curve reaches the global minimum through a global scalar search along the path. The tangent is a typical line: among all lines through the incumbent, about 4% of directions contain a better point, and the tangent (at $27.8°$) and both coordinate lines are outside this set.

A.11 Proof of Theorem 5

Theorem 5 restates Theorem S8 of the supplement; we give its proof here so that the main text is self-contained. Strict unimodality with minimizer $p(t)$ gives $\min_s q(t,s) = b(t)$ and $\psi(e) \geq 0$ on the feasible residuals.

Proof of Theorem 5. Set $\delta = \frac{1}{2}\sqrt{\epsilon/K}$ and divide I into $N = \max\{1, \lceil (\ell/2)(R/\delta)^{1/3} \rceil\}$ equal leaves of half-width $h = \ell/(2N)$, so that $Rh^3 \leq \delta$, including $R = 0$. At the left, middle, and right baseline coordinates of each leaf, golden-section bracketing on the unimodal profile $q(t,\cdot)$ reduces a bracket of $p(t)$ to width at most 2δ, and a paid evaluation at its midpoint gives an anchor within δ of $p(t)$. Each anchor costs $O(1 + \log_+(W/\delta))$ queries, including that evaluation, and no unqueried point is treated as an incumbent.

Let P interpolate the three paid anchors of a leaf, with residual $e = P - p$. Lemma A.6 with $D_\ell = 0$, $e_a = \delta$, $T = h$, and $Rh^3 \leq \delta$ gives

$$E_0 \leq \left(\tfrac{5}{4} + \tfrac{1}{9\sqrt{3}}\right)\delta < \tfrac{4}{3}\delta, \qquad E_1 \leq \frac{4\delta}{h} + Rh^2 \leq \frac{5\delta}{h}, \qquad E_2 \leq \frac{4\delta}{h^2} + Rh \leq \frac{5\delta}{h^2}.$$

Since $\epsilon \leq K\rho^2$ gives $\delta \leq \rho/2$, we have $E_0 < 2\rho/3$, so the interpolant stays in J without clipping. Integrating $|\psi''| \leq M_\psi$ from $\psi(0) = \psi'(0) = 0$ gives $0 \leq \kappa\psi(e) \leq KE_0^2/2 \leq 2\epsilon/9$ and $|\kappa\psi'(e)| \leq K|e|$. The frozen leaf profile $f(t) = q(t, P(t))$ therefore satisfies $f - b \leq 2\epsilon/9$ and, by direct differentiation,

$$|f''| \leq M_b + K(E_1^2 + E_0 E_2) \leq M_b + \frac{95\,\epsilon}{12h^2}.$$

A safeguarded COCP or BB search with this derivative-Lipschitz bound and target $\epsilon/2$ costs $O(1 + h\sqrt{M_b/\epsilon})$ evaluations on each leaf by Proposition 4, because $h\sqrt{95/(12h^2)}$ is a constant. On the leaf that contains a minimizer t^* of b, the returned evaluated value is at most $f(t^*) + \epsilon/2 \leq b(t^*) + 2\epsilon/9 + \epsilon/2$, so the best evaluated value U over all leaves has $U - \min q < \epsilon$. Path search costs $O(N + \ell\sqrt{M_b/\epsilon})$ in total, and the $3N$ anchors cost $O(N(1+\log_+(W/\delta)))$. Since $(R/\delta)^{1/3} = 2^{1/3}(KR^2/\epsilon)^{1/6}$ gives $N = O(1 + \ell(KR^2/\epsilon)^{1/6})$, and $W/\delta = 2W\sqrt{K/\epsilon}$, the total is (11). □