INITIALIZATION SHAPES THE COMPUTATIONAL REGIME OF ADAPTIVE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimization is central to modern machine learning, and its efficiency directly shapes the cost of model training. However, initialization is usually viewed as affecting only the starting point and transient trajectory, rather than the asymptotic convergence rate. We show that this separation can fail for adaptive quasi-Newton optimization. Parameter initialization alone can select distinct asymptotic computational regimes with the learning objective, optimizer family, and line-search hyperparameters held fixed. On a multidimensional regularized pinball-loss family, we identify an infinite hierarchy of such regimes for the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method and its limited-memory variant (L-BFGS), and develop a trace-to-complexity analysis that links recurring line-search decisions to asymptotic progress, evaluation cost, and stability. The analysis reveals an exact tradeoff between contraction and sensitivity to initialization, and we further show that each exact regime persists over any fixed finite horizon within an open full-dimensional neighborhood. Extensive experiments, including multidimensional and real-world learning settings, show that the predicted structure is observable in finite precision and persists beyond the exact analytic construction.

1 INTRODUCTION

Optimization is a fundamental component of modern machine learning, determining the parameters ultimately learned by a model and strongly influencing the computational cost of training. A typical optimization process begins from an initialization and iteratively updates model parameters using information from the current state, including gradients, curvature estimates, and adaptive step-selection rules. Recent work has examined how initialization and parameterization shape convergence and representation learning (Li et al., 2025a; Dominé et al., 2025; Tu et al., 2024), how adaptive learning-rate and line-search rules affect optimization dynamics and convergence (Collins-Woodfin et al., 2024; Cavalcanti et al., 2025), and how the geometry encountered along the optimization trajectory, such as sharpness and edge-of-stability behavior, interacts with training dynamics (Wang et al., 2022; Lyu et al., 2022).

However, initialization can leave a persistent computational imprint on optimization that goes beyond selecting a solution or shaping the early optimization trajectory. We uncover a form of path dependence in adaptive optimization in which different parameter initializations, with the objective, optimizer, and hyperparameters held fixed, select distinct recurring internal decision patterns. These differences can persist asymptotically, placing the same optimizer in different computational regimes and producing different evaluation-normalized convergence rates. Moreover, we establish the same phenomenon for a convex objective with a unique optimum, separating persistent computational path dependence from nonconvex basin-selection effects.

The effect of initialization is examined in a multidimensional convex pinball-loss family using BFGS and its limited-memory variant. This setting exposes the optimizer's internal control state while retaining a standard nonsmooth learning loss and matrix-valued parameter geometry. A trace-to-complexity analysis connects recurring line-search decisions with asymptotic progress, evaluation cost, and stability, and a finite-horizon persistence result shows that the resulting control patterns extend to full-dimensional neighborhoods of the exact regimes. The experiments then test the same mechanism across ambient dimension and limited-memory optimization, reproduce the hierarchy in

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.

forecasting, measure its cost on real and frozen-representation tasks, and finally transfer the effect to native strong-Wolfe L-BFGS.

Our contributions are:

- **Initialization-selected regime hierarchy.** We prove that initialization alone can select infinitely many recurrent computational regimes with distinct evaluation-normalized convergence rates, with the learning objective, optimizer configuration, and line search fixed. This establishes persistent path dependence in adaptive optimization. We further show that this hierarchy is dimension-free on the active control state and is shared exactly by BFGS and L-BFGS for any memory $M \geq 1$.

- **Trace-to-complexity calculus.** We develop an exact analysis that maps a complete line-search decision trace to its state contraction, evaluation cost, trial-normalized rate, and initialization sensitivity. This yields an exact reciprocity between optimization contraction and initialization sensitivity, together with a quantitative tradeoff between computational progress and robustness to initialization.

- **Full-dimensional persistence and empirical validation.** We prove that each exact recurrent regime is the center of a positive-volume family of genuinely multidimensional trajectories that preserves the same control trace over any fixed finite horizon. We then validate the predicted structure across ambient dimension and limited-memory optimization, reproduce the hierarchy in forecasting, observe initialization-dependent cost on a real-world electricity task and a frozen learned representation, and transfer the effect to native strong-Wolfe L-BFGS.

2 Related Work

Initialization and optimization regimes. Recent work has made the long-term role of initialization increasingly explicit. Initialization can change convergence rates in matrix factorization (Li et al., 2025a), select lazy, active, or feature-learning regimes (Tu et al., 2024; Dominé et al., 2025), and alter the dynamics and learning-rate robustness of low-rank adaptation (Hayou et al., 2024; Li et al., 2025b). These results primarily characterize how initialization changes parameter or representation dynamics. We study a complementary computational effect: initialization changes the recurring internal control state of a fixed adaptive optimizer and thereby changes its evaluation-normalized asymptotic behavior.

Adaptive step selection and quasi-Newton optimization. Adaptive step-size mechanisms are themselves rich dynamical systems. Exact analyses have characterized long-run behavior of adaptive learning rates and line search (Collins-Woodfin et al., 2024), while stochastic and adaptive line-search methods target robustness and efficiency in modern training (Galli et al., 2023; Jiang & Stich, 2023; Cavalcanti et al., 2025). For quasi-Newton methods, recent work gives global complexity guarantees for BFGS with Armijo–Wolfe search (Jin et al., 2024), uses BFGS and related updates in bilevel learning (Fang et al., 2025), and studies L-BFGS as a practical component of scientific-machine-learning pipelines (Rathore et al., 2024). Our analysis keeps the quasi-Newton rule and line-search policy fixed and asks which computational regime is selected by the initialization.

Nonsmooth quasi-Newton dynamics. Classical nonsmooth BFGS analyses (Lewis & Overton, 2008; 2013) provide a particularly clean setting in which line-search control and curvature updates interact. Subsequent work established convergence results for nonsmooth BFGS (Guo & Lewis, 2018), analyzed limited-memory BFGS on structured nonsmooth convex objectives (Asl & Overton, 2021), and continued the study of nonsmooth quasi-Newton convergence mechanisms (Gebken, 2026). Our focus is different: we characterize initialization-selected recurrent decision regimes, derive their evaluation complexity and sensitivity, and track how this structure persists under multidimensional and limited-memory extensions.

3 THEORY: INITIALIZATION-SELECTED REGIMES IN MULTIDIMENSIONAL QUASI-NEWTON OPTIMIZATION

We investigate how initialization can select different long-run computational regimes after the learning problem, quasi-Newton method, and line-search rule are fixed. Section 3.1 introduces the notation and optimization setup, Section 3.2 isolates the dimension-free active dynamics, Section 3.3 constructs the infinite regime hierarchy, and Section 3.4 extends its finite-horizon signatures to full-dimensional neighborhoods. All proofs and branch certificates are provided in Appendices A–J.

3.1 PRELIMINARIES AND NOTATION

Let $d \geq 1$, let $a \in \mathbb{R}^d$ satisfy $\|a\|_2 = 1$, and define the active and orthogonal projectors

$$P = aa^\top, \qquad P_\perp = I - P.$$

For $u > 0$, set $\tau = u/(1+u)$ and consider

$$\mathcal{F}_{u,\mu}(\theta) = \ell_\tau\big(-(1+u)a^\top\theta\big) + \frac{\mu}{2}\|P_\perp\theta\|_2^2, \qquad \mu > 0, \tag{1}$$

where $\ell_\tau(r) = \max\{\tau r, (\tau-1)r\}$ (Koenker & Bassett, 1978). Its unique minimizer is $\theta^\star = 0$. We write

$$x = a^\top\theta$$

for the active coordinate. Along the active subspace,

$$f_u(x) := \max\{x, -ux\} = \ell_\tau(-(1+u)x)\,. \tag{2}$$

We consider full inverse BFGS with $H_0 = I$ and standard L-BFGS with empty history, identity scaling, and memory $M \geq 1$. Both use the same deterministic Armijo–weak-Wolfe doubling–bisection line search (Lewis & Overton, 2008): trial 1 is tested first, trials are doubled before an upper bracket exists and bisected afterward, and every rejected or accepted trial contributes to the evaluation clock. We denote Wolfe failure, Armijo failure, and acceptance by W, A, and S, respectively. The Armijo and weak-Wolfe parameters are c_1 and c_2, with $0 \leq c_1 < c_2 < 1$. We use strict Armijo acceptance; Wolfe is tested after Armijo passes, and a trial exactly at the active kink is assigned W. Appendix C specifies the bracket updates and boundary conventions.

For an active state $\theta = xa$ and search direction p, define the kink-reaching trial length σ and the thresholds Θ_+ and Θ_- by

$$\sigma = -\frac{x}{a^\top p} > 0, \qquad \Theta_+ = \frac{u+1}{u+c_1}, \qquad \Theta_- = \frac{u+1}{1+c_1 u}. \tag{3}$$

We write $\Theta_\epsilon = \Theta_+$ on the positive sheet and $\Theta_\epsilon = \Theta_-$ on the negative sheet.

A complete line-search trace ω records the starting sign sheet, every branch decision, and every trial length. For a recurrent trace, N_ω is the number of line-search trials in one return, $q_\omega \in (0,1)$ is its state-return contraction,

$$\rho_\omega = q_\omega^{1/N_\omega}$$

is its trial-normalized rate, and λ_ω denotes the local sensitivity of the phase return at the recurrent phase. The indexed notation $N_m, q_m, \rho_m, \lambda_m$ is used for the explicit hierarchy constructed below.

3.2 DIMENSION-FREE ACTIVE DYNAMICS

Proposition 3.1 (Active line-search control law)**.** *For every active state, the line-search decision at trial t depends only on the current phase, sign sheet, and trial value:*

$$\mathsf{W}:\ 0 < t \leq \sigma, \qquad \mathsf{S}:\ \sigma < t < \Theta_\epsilon\sigma, \qquad \mathsf{A}:\ t \geq \Theta_\epsilon\sigma. \tag{4}$$

An accepted crossing at trial t updates the phase by

$$F_{+,t}(\sigma) = \frac{u+1}{u}\left(1 - \frac{\sigma}{t}\right), \qquad F_{-,t}(\sigma) = (u+1)\left(1 - \frac{\sigma}{t}\right). \tag{5}$$

Thus the adaptive decisions on the active subsystem are completely described by a one-dimensional control state and its literal branch word.

Remark 3.2 (Control-state interpretation). For a fixed sign sheet, the phase σ determines the decision at each trial, while preceding decisions determine the bracket and next trial. Thus one continuous phase together with the discrete search state describes the active line search. The scalar specialization and all branch-edge conventions are derived in Appendix C.

Proposition 3.3 (Dimension-free BFGS dynamics)**.** *Start from an active initialization $\theta_0 = x_0 a$. Full BFGS remains in the active subspace and preserves the block form*

$$H_k = P_\perp + h_k P.$$

If an accepted step has $s = \delta a$ and $y = \gamma a$ with $\delta\gamma > 0$, then $h_{k+1} = \delta/\gamma$. L-BFGS with any memory $M \geq 1$ produces the same active search direction after every accepted step. Consequently, both methods generate identical active iterates, accepted steps, and line-search traces, independently of the ambient dimension and the limited-memory history size.

Remark 3.4 (Why dimension and memory disappear). The multidimensional matrix dynamics contain an exact invariant control subsystem: the orthogonal curvature block remains decoupled, while the newest curvature pair determines the active secant action. Section 5.1 tests this dimension- and memory-invariance numerically under rotated high-dimensional embeddings and float64 arithmetic. The detailed matrix update and L-BFGS two-loop proof are given in Appendix A.

3.3 Infinite initialization-selected regime hierarchy

For $m \geq 1$, let $h = 2^{-m}$ and

$$D = (u+1)^2 - 2hu,$$

and define

$$\sigma_m^+ = \frac{2(u+1)(u+1-hu)}{D}, \qquad \sigma_m^- = \frac{h(u^2-1)}{D}. \tag{6}$$

Let $u_* > 2$ be the unique real root of

$$u^3 - u^2 - 3u - 3 = 0. \tag{7}$$

We call (u, c_1, c_2) admissible when either $c_1 = 0$, $2 \leq u \leq u_*$, and $0 < c_2 < 1$, or $21/10 \leq u \leq 13/6$, $0 \leq c_1 \leq (u-2)/u$, and $c_1 < c_2 < 1$.

Theorem 3.5 (Initialization-selected regime hierarchy)**.** *Fix any admissible parameter tuple. For every $m \geq 1$, the initialization $\theta_0 = \sigma_m^+ a$ generates under both methods the recurring line-search trace $\mathsf{WS}/\mathsf{A}^m\mathsf{S}$. One two-step return uses $N_m = m + 3$ trials and scales the active pair (x, h_k) by*

$$q_m(u) = \frac{2^{1-m}u}{(u+1)^2},$$

and has exact trial-normalized rate

$$\rho_m(u) = q_m(u)^{1/(m+3)}. \tag{8}$$

For fixed admissible parameters, the rates are pairwise distinct and satisfy $\rho_1(u) > \rho_2(u) > \cdots > 1/2$ with $\rho_m(u) \to 1/2$.

Remark 3.6 (What changes with the regime index). The first members are WS/AS, WS/AAS, and $\mathsf{WS}/\mathsf{AAAS}$. Increasing m adds one rejected Armijo trial before the accepted trial 2^{-m}, while the learning problem, optimizer, and line-search hyperparameters remain fixed. Thus initialization changes both the recurring internal decision pattern and the evaluation-normalized rate. The complete construction, branch certificates, state scaling, and rate ordering are given in Appendices D–G.

Remark 3.7 (Parameter robustness). The hierarchy holds on nondegenerate line-search parameter regions. At $c_1 = 0$, the common all-m interval $[2, u_*]$ is maximal. On $21/10 \leq u \leq 13/6$, the pointwise all-m Armijo ceiling is $c_1 \leq (u-2)/u$, yielding the maximal uniform closed height $c_1 \leq 1/21$. The sharp boundary arguments are given in Appendices E and F.

3.4 Full-dimensional finite-horizon persistence

The exact recurrent trajectories define the asymptotic regimes. Their finite-horizon control signatures extend beyond the invariant active subsystem.

Theorem 3.8 (Full-dimensional finite-horizon persistence)**.** *Fix admissible parameters and a regime index m whose reference trace has strict branch margins, and let the return horizon $R \geq 1$ be finite. For each optimizer there exists $\varepsilon_R > 0$ such that every parameter initialization in $B(\sigma_m^+ a, \varepsilon_R) \subset \mathbb{R}^d$ executes the same line-search trace as the exact regime for its first R returns. For full BFGS, the conclusion also holds jointly for all SPD initial inverse-Hessian matrices in a sufficiently small neighborhood of I.*

Remark 3.9 (From exact regimes to multidimensional training trajectories). Strict margins hold throughout the stated positive-Armijo region and at $c_1 = 0$ for $2 \leq u < u_*$. For $d > 1$, the neighborhood includes initializations outside the active subspace. The exact recurrent initialization is the center of a positive-volume family of trajectories over every fixed training horizon; its radius may shrink with the horizon. This separates exact asymptotic recurrence from finite-horizon observability. Section 5.2 directly tests this prediction by perturbing both the parameter initialization and the BFGS inverse-Hessian state. The detailed continuity and strict-margin proof is given in Appendix B.

4 Trace-to-Complexity Calculus

A complete line-search trace carries two kinds of information. Accepted steps determine how far the optimizer progresses when the state returns, while the full accepted-and-rejected trial pattern determines whether that return is realizable and how much adaptive computation it costs. This section turns a literal trace into contraction, evaluation complexity, and initialization sensitivity.

4.1 Complete traces determine recurrent complexity

Let ω be a fully resolved sign-return trace with accepted trials $t_1, \ldots, t_p$. These p accepted steps strictly alternate across the kink, so p is even, with $p/2$ steps originating on each sign sheet.

Proposition 4.1 (Trace-to-complexity calculus)**.** *For any feasible trace ω, the accepted-step maps compose to an affine phase return $R_\omega(\sigma) = A_\omega \sigma + B_\omega$. Its active inverse-curvature multiplier and phase derivative are*

$$q_\omega = \frac{u^{p/2} \prod_{j=1}^{p} t_j}{(u+1)^p}, \qquad \lambda_\omega = A_\omega = q_\omega^{-1}. \tag{9}$$

If $0 < q_\omega < 1$ and the fixed phase $\sigma_\omega^* = B_\omega/(1 - A_\omega)$ satisfies the complete trace conditions, the active pair (x, h_k) scales by q_ω at each return. The trace then defines a recurrent computational regime with trial-normalized rate $\rho_\omega = q_\omega^{1/N_\omega}$.

Remark 4.2 (Trace decomposition). The accepted trials determine q_ω, whereas rejected trials enter through the literal feasibility conditions and the clock N_ω. The complete trace must also generate the accepted trial lengths through the prescribed bracket updates; accepted steps alone are not a feasibility certificate. The complete feasibility calculus is given in Appendix H.

Remark 4.3 (Regime cells). For fixed parameters and starting sheet, the strict conditions of a fully resolved finite trace define an open interval of feasible phases when nonempty. Allowing parameters to vary gives a semialgebraic feasibility system. Points in a cell share a finite trace; a feasible fixed point of its return map generates exact recurrence. Line-search equalities form the boundaries. The full guard system and boundary conventions are given in Appendix H.

4.2 Complexity, stability, and finite observations

Corollary 4.4 (Complexity–stability reciprocity)**.** *Within the feasibility cell, every certified recurrent regime satisfies*

$$R_\omega(\sigma) - \sigma_\omega^* = \lambda_\omega(\sigma - \sigma_\omega^*), \qquad \lambda_\omega = q_\omega^{-1} = \rho_\omega^{-N_\omega}. \tag{10}$$

Corollary 4.5 (Explicit scaling along the regime hierarchy)**.** *Let*

$$A(u) = \frac{(u+1)^2}{u}.$$

Along the hierarchy of Theorem 3.5,

$$q_m = \frac{2^{1-m}}{A(u)}, \qquad \log_2 \rho_m = -1 + \frac{4 - \log_2 A(u)}{m+3}, \qquad \lambda_m = A(u) 2^{m-1}. \tag{11}$$

In particular, $\rho_m \downarrow 1/2$ while $\lambda_{m+1} = 2\lambda_m$.

Remark 4.6 (Rate–sensitivity interpretation). Each increment in m adds one rejected Armijo trial, halves the return contraction q_m, and doubles the local phase sensitivity λ_m. Stronger evaluation-normalized contraction is therefore coupled to a progressively narrower tolerance to initialization perturbations. The closed-form ordering and asymptotic calculation are given in Appendix G.

Corollary 4.7 (Finite-trace ambiguity)**.** *Let J_ω be a strict one-period feasibility cell containing a fixed phase with $\lambda_\omega > 1$. The active phases that execute r complete copies of the trace have width $|J_\omega|\lambda_\omega^{-(r-1)}$; requiring the returned phase to remain in the cell changes the width to $|J_\omega|\lambda_\omega^{-r}$.*

Corollary 4.7 quantifies the shrinking active-phase set, while Theorem 3.8 guarantees a nonzero-volume full-dimensional neighborhood for every fixed finite horizon. Together they separate asymptotic regime sensitivity from finite-horizon observability and motivate the shadowing experiments in Section 5. The exact return-cylinder calculation is given in Appendix I.

Computational cross-check. The calculus can be enumerated over a finite trace budget. For a total trial budget $N \leq 12$, exact rational enumeration yields 265,708 raw occurrences, 68,388 cyclic classes, and 152 classes feasible at the declared probes, recovering the analytic hierarchy within the budget. The complete grammar, probe set, and certificates are reported in Appendix J; the infinite hierarchy itself is established by Theorem 3.5.

5 Experiments

We organize the experiments as a progression from validating the exact theoretical structure to testing its persistence beyond the analytic construction:

- **Q1.** Does the exact regime structure persist across multidimensional and limited-memory settings?
- **Q2.** Does this structure remain observable under full-dimensional perturbations over finite horizons?
- **Q3.** Can the exact hierarchy be realized and observed in a forecasting objective under finite-precision computation?
- **Q4.** Does initialization-dependent computation remain observable beyond the exact construction in real learning settings?
- **Q5.** Does the same effect persist under a native strong-Wolfe L-BFGS implementation?

Complete grids, seeds, stopping rules, accounting conventions, and audit details are given in Appendix K.1.

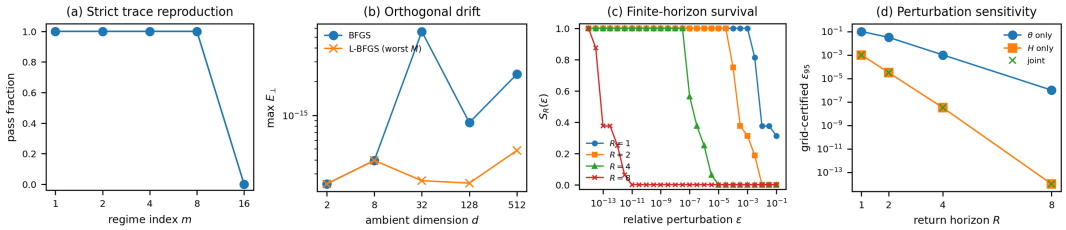


Figure 1: **Dimension invariance and finite-horizon shadowing. (a–b)** Strict trace reproduction and orthogonal drift under dimension, rotation, and limited-memory changes. **(c–d)** Same-trace survival and the corresponding 95% shadowing radius under parameter- and optimizer-state perturbations.

5.1 Dimension and optimizer invariance

To answer Q1, panels (a–b) of Fig. 1 test the predictions of Proposition 3.3 and Theorem 3.5 under rotated high-dimensional embeddings and different L-BFGS memories. Panel (a) shows exact reproduction of the complete scalar branch word and accepted trials for every resolved regime with $m \leq 8$, independently of ambient dimension, rotation, optimizer, or memory. At $m = 16$, repeated

contraction drives the active state to the scale of float64 machine precision, and the same breakdown is also observed in the aligned $d = 1$ runs. This locates the observed boundary at finite precision.

Panel (b) tests the second implication of the theory: the active dynamics remain decoupled from orthogonal coordinates. Across random rotations up to $d = 512$, the normalized orthogonal drift stays at approximately 10^{-16}–10^{-15} for both full BFGS and L-BFGS. Together, the two panels verify the predicted dimension- and memory-invariant control dynamics throughout the numerically resolved regimes. Full protocol and reconciliation statistics are given in Appendix K.1.

5.2 Full-dimensional finite-horizon shadowing

To answer Q2, Fig. 1c–d moves away from the exact invariant subsystem and tests the finite-horizon prediction of Theorem 3.8. Panel (c) plots the probability of retaining the complete trial-level control trace under increasing perturbation scale. The transition shifts left by several orders of magnitude as the requested return horizon grows from $R = 1$ to $R = 8$. This turns the theorem's existence statement into a quantitative stability picture: a nonzero same-trace neighborhood is observable at every resolved horizon, but longer shadowing requires progressively tighter initialization.

Panel (d) reveals that this neighborhood is strongly anisotropic in optimizer state. Parameter-only perturbations tolerate much larger radii than perturbations to the inverse-Hessian state, while the H-only and joint curves nearly coincide on the frozen grid. For the representative configuration, the curvature state is therefore the limiting component of finite-horizon regime stability. The experiment quantifies how the same-trace neighborhood shrinks with the return horizon and varies across perturbation types. Full perturbation grids and resolution-aware statistics are reported in Appendix K.1.

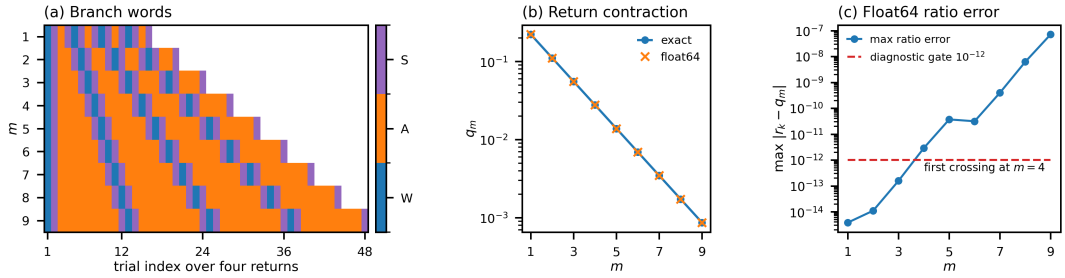


Figure 2: **Forecasting realization of the exact regime hierarchy. (a)** Line-search branch words. **(b)** Exact and float64 return contractions. **(c)** Float64 state-ratio error relative to the frozen numerical diagnostic tolerance.

5.3 Exact forecasting realization and numerical observability

To answer Q3, Fig. 2 embeds the analytic hierarchy in a periodic pinball-loss forecasting problem whose objective preserves the exact control dynamics; construction details are deferred to Appendix K.1. Panel (a) makes the computational regimes visible as line-search decision words. As m increases, the growing $\mathsf{A}^m\mathsf{S}$ blocks require progressively more rejected trials before acceptance. The hierarchy is therefore visible directly as a hierarchy of repeated adaptive decisions.

Panel (b) connects those decisions to optimization progress. The float64 return contractions visually coincide with the exact theoretical values, showing that the longer decision traces correspond to the predicted changes in per-return contraction. Panels (a) and (b) together give a direct trace-to-progress interpretation of the theory: the branch word records the adaptive computation spent by the optimizer, while the return contraction records what that computation achieves.

Panel (c) isolates the finite-precision boundary. The discrete branch words remain unchanged through $m = 9$, even though the continuous state-ratio error first crosses the frozen 10^{-12} diagnostic tolerance at $m = 4$. Hence, under strong contraction, the symbolic trace is preserved even when continuous state ratios lose numerical accuracy. A compensated replay recovers the stringent ratio agreement, identifying cancellation as the source of the discrepancy. Together, the panels connect the exact hierarchy to its finite-precision realization.

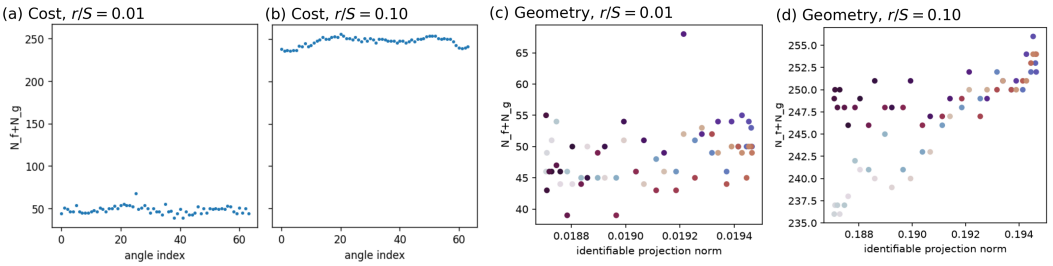


Figure 3: **Equal-distance initialization changes optimization cost on real forecasting. (a–b)** Oracle cost across initialization directions at two fixed radii. **(c–d)** Oracle cost against identifiable projection norm for the same configurations.

5.4 Real forecasting and learned-representation transfer

To answer Q4, Fig. 3 tests initialization-dependent computation in electricity forecasting with DLinear (Zeng et al., 2023) on ETTm1 (Zhou et al., 2021). With the data, convex forecasting objective, optimizer, stopping target, and audited reference fixed, only the direction of equal-distance parameter initialization is changed. Full protocol details are given in Appendix K.1.

Panels (a–b) show a $1.74\times$ spread in optimization work at the smaller radius: equal-distance initializations require 39–68 oracle calls to reach the same target, with narrower variation at the larger radius. To account for DLinear's structural rank deficiency, panels (c–d) examine distances after projection onto the identifiable subspace. At the smaller radius, these effective distances vary by only $1.0405\times$ and correlate only modestly with oracle cost. Thus the large computational spread is not matched by a comparable spread in identifiable distance, supporting initialization-dependent computation beyond a simple distance-to-reference explanation.

To test whether this effect depends on the nonsmooth kink of the pinball loss, we repeat the equal-distance initialization experiment with a C^∞ pseudo-Huber quantile loss in the identifiable DLinear quotient. Across all three smoothing levels, initialization-dependent variation in optimization cost remains clearly observable. Thus the empirical effect extends beyond the nonsmooth geometry used to expose the exact analytic hierarchy; the cross-setting comparison is summarized in Table 1, with complete results reported in Appendix K.1.

Transfer to a learned representation. The same question remains meaningful after replacing the direct forecasting representation with learned features. On the official frozen PatchTST representation (Nie et al., 2023), a certified 9,228-dimensional convex quantile head reaches the common target for all 64 small-radius starts, with oracle cost ranging from 3932 to 4179 and median 4053.5 ($1.063\times$ max/min). The effect persists with a learned representation, although it is smaller than for DLinear. This transfer is also geometrically cleaner: the frozen affine head is full-column-rank,

Table 1: **Initialization-dependent computation across transfer settings.** Equal-distance initialization directions are varied around a fixed reference while the learning objective and optimizer configuration within each setting remain fixed. In the *Opt./LS* column, WW denotes the deterministic Armijo–weak-Wolfe line search and SW denotes the native PyTorch strong-Wolfe line search. Spread is the max/min cost ratio among runs reaching the target.

Setting	Opt./LS	Dim.	Reached	Cost			Spread
				Min	Median	Max	
ETTm1 + DLinear	BFGS + WW	1,164	64/64	39	49	68	$1.74\times$
DLinear quotient, $\delta = 10^{-3}$	BFGS + WW	588	64/64	49	59	66	$1.347\times$
DLinear quotient, $\delta = 10^{-2}$	BFGS + WW	588	64/64	47	54.5	64	$1.362\times$
DLinear quotient, $\delta = 10^{-1}$	BFGS + WW	588	64/64	26	35	46	$1.769\times$
ETTm1 + frozen PatchTST head	BFGS + WW	9,228	64/64	3,932	4,053.5	4,179	$1.063\times$
ETTm1 + frozen PatchTST head	L-BFGS + SW	9,228	51/64	4,265	4,370	5,358	$1.256\times$

so the variation is observed without the exact parameter nullspace present in DLinear. Runs at the medium and large radii reach the prescribed budget; their final-gap diagnostics are reported in the appendix. The complete learned-representation protocol and a compact visualization are given in Appendix K.1 and Fig. 6. The cross-setting comparison is summarized in Table 1.

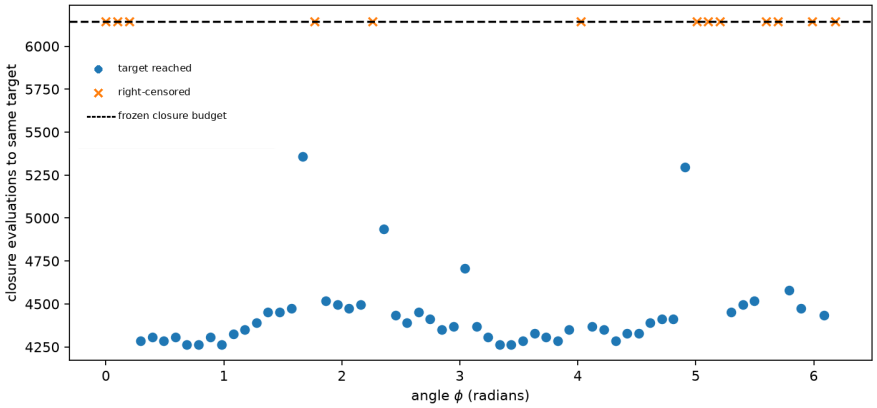


Figure 4: **Initialization-dependent computation under native strong-Wolfe L-BFGS.** Closure evaluations to the common target across 64 equal-distance initializations; unresolved runs are right-censored at 6,144 closures.

5.5 Transfer to native strong-Wolfe L-BFGS

To answer Q5, Fig. 4 asks whether the initialization–computation effect survives when the custom line-search rule is replaced by a standard optimizer implementation. We rerun the same small-radius starts on the frozen learned-representation objective using native PyTorch (Paszke et al., 2019) L-BFGS with strong-Wolfe line search (further details in Appendix K.1). Among runs reaching the target, closure costs have a $1.256\times$ spread across initializations and median 4,370. This extends the empirical initialization–computation effect beyond the deterministic weak-Wolfe construction to native strong-Wolfe L-BFGS on a learned representation: controlled initialization changes continue to produce measurable differences in optimization work. Table 1 summarizes the progression from direct forecasting to learned representations and native strong-Wolfe optimization.

6 Discussion and Limitations

These results suggest treating initialization and optimizer state as coupled design choices, since early parameter choices shape the curvature estimates and step-selection decisions that govern later computation. The same perspective raises an important question for larger nonconvex learning systems: whether such regimes arise, and whether they can be predicted or controlled to reduce training cost.

Further empirical validation across a broader range of learning architectures and optimizers would help characterize when initialization-dependent computation is most pronounced. Moreover, the present work focuses on the theoretical mechanisms, while how to exploit this dependence in practical training workflows to reduce overall computational cost remains to be explored.

7 Conclusion

Initialization can be a persistent computational variable in adaptive quasi-Newton optimization. We prove an infinite hierarchy of initialization-selected regimes, develop a trace-to-complexity calculus linking internal decisions to progress, evaluation cost, and sensitivity, and show finite-horizon persistence in full-dimensional neighborhoods. The experiments recover these structures across dimension and L-BFGS memory, and observe the same initialization–computation dependence in real forecasting, a frozen learned representation, and native strong-Wolfe L-BFGS.

AI USE STATEMENT

AI tools were used to polish the writing and check the manuscript. The human author reviewed the AI-assisted work and takes responsibility for the final content, including all text, claims, code, and artifacts.

REFERENCES

Azam Asl and Michael L. Overton. Analysis of limited-memory BFGS on a class of nonsmooth convex functions. *IMA Journal of Numerical Analysis*, 41(1):1–27, 2021. doi: 10.1093/imanum/drz052.

Joao V. Cavalcanti, Laurent Lessard, and Ashia C. Wilson. Adaptive backtracking line search. In *International Conference on Learning Representations*, pp. 78905–78951, 2025.

Elizabeth Collins-Woodfin, Inbar Seroussi, Begoña García Malaxechebarría, Andrew W. Mackenzie, Elliot Paquette, and Courtney Paquette. The high line: Exact risk and learning rate curves of stochastic adaptive learning rate algorithms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 6500–6548, 2024. doi: 10.52202/079017-0209.

Clémentine C. J. Dominé, Nicolas Anguita, Alexandra M. Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M. Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In *International Conference on Learning Representations*, pp. 102747–102798, 2025.

Sheng Fang, Yongjin Liu, Wei Yao, Chengming Yu, and Jin Zhang. qNBO: quasi-Newton meets bilevel optimization. In *International Conference on Learning Representations*, 2025.

Leonardo Galli, Holger Rauhut, and Mark Schmidt. Don't be so monotone: Relaxing stochastic line search in over-parameterized models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1509.

Bennet Gebken. Technical results on the convergence of quasi-Newton methods for nonsmooth optimization. *Journal of Optimization Theory and Applications*, 209(2):57, 2026. doi: 10.1007/s10957-026-03008-z.

Jiayi Guo and Adrian S. Lewis. Nonsmooth variants of Powell's BFGS convergence theorem. *SIAM Journal on Optimization*, 28(2):1301–1311, 2018. doi: 10.1137/17M1121883.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on LoRA finetuning dynamics. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3715.

Xiaowen Jiang and Sebastian U. Stich. Adaptive SGD with Polyak stepsize and line-search: Robust convergence and variance reduction. In *Advances in Neural Information Processing Systems*, volume 36, pp. 26396–26424, 2023. doi: 10.52202/075280-1148.

Qiujiang Jin, Ruichen Jiang, and Aryan Mokhtari. Non-asymptotic global convergence analysis of BFGS with the Armijo–Wolfe line search. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0536.

Roger Koenker and Gilbert Bassett, Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Adrian S. Lewis and Michael L. Overton. Nonsmooth optimization via BFGS, 2008. URL `https://optimization-online.org/2008/12/2172/`. Optimization Online, record 2172.

Adrian S. Lewis and Michael L. Overton. Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, 141:135–163, 2013. doi: 10.1007/s10107-012-0514-2.

Bingcong Li, Liang Zhang, Aryan Mokhtari, and Niao He. On the crucial role of initialization for matrix factorization. In *International Conference on Learning Representations*, pp. 4520–4555, 2025a.

Shiwei Li, Xiandi Luo, Xing Tang, Haozhao Wang, Hao Chen, Weihong Luo, Yuhua Li, Xiuqiang He, and Ruixuan Li. Beyond zero initialization: Investigating the impact of non-zero initialization on LoRA fine-tuning dynamics. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35519–35535. PMLR, 2025b.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34689–34708, 2022. doi: 10.52202/068431-2514.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Pratik Rathore, Weimu Lei, Zachary Frangella, Lu Lu, and Madeleine Udell. Challenges in training PINNs: A loss landscape perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42159–42191. PMLR, 2024.

Zhenfeng Tu, Santiago Aranguri, and Arthur Jacot. Mixed dynamics in linear networks: Unifying the lazy and active regimes. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106059–106104, 2024. doi: 10.52202/079017-3365.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9983–9994, 2022. doi: 10.52202/068431-0725.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

APPENDIX

A MULTIDIMENSIONAL ACTIVE-SUBSPACE REDUCTION

Let $P = aa^\top$ and $P_\perp = I - P$. If $\theta = xa$ then $\nabla \mathcal{F}_{u,\mu}(\theta) = g_u(x)a$. Suppose full inverse BFGS has $H = hP + G_\perp$ with $G_\perp = P_\perp G_\perp P_\perp$. Its search direction is $p = -hg_u(x)a$, so every trial remains in $\mathrm{span}(a)$. An accepted crossing has $s = \delta a$ and $y = \gamma a$ with $\delta\gamma > 0$. Since $\rho = (s^\top y)^{-1}$,

$$\rho s y^\top = \rho y s^\top = P,$$

and the inverse-BFGS update becomes

$$H^+ = (I - P)H(I - P) + \rho s s^\top = G_\perp + \frac{\delta}{\gamma} P.$$

Thus the active inverse curvature follows the scalar secant update while the orthogonal block remains decoupled.

For standard L-BFGS, let the newest stored pair be $(s, y) = (\delta a, \gamma a)$. In the backward two-loop recursion, for any current gradient $q = q_a a$,

$$\alpha = \frac{s^\top q}{s^\top y} = \frac{q_a}{\gamma}, \qquad q \leftarrow q - \alpha y = 0.$$

Every earlier pair therefore receives zero coefficient, and the initial L-BFGS scaling is applied to the zero vector. In the forward loop, earlier pairs again contribute nothing, while the newest pair gives

$$r = s\alpha = \frac{\delta}{\gamma} q_a a.$$

Hence the L-BFGS search direction equals the full-BFGS active direction and is independent of the memory size once the newest valid pair is present. The first direction also agrees when both methods start from the identity scaling. Induction over accepted steps proves Proposition 3.3.

B FULL-DIMENSIONAL FINITE-HORIZON PERSISTENCE

We prove Theorem 3.8 using the strict margins of the specified reference trace. Fix the parameters, m, and a finite horizon R. Up to its Rth two-step return, the reference execution contains finitely many trial points and curvature updates. Every tested point lies away from the kink, every recorded decision has a strict margin, and every accepted curvature product $s^\top y$ is positive. The certificate calculations in Appendices E and F establish these properties for the parameter ranges stated in the main text.

For each optimizer, follow this finite execution inductively. Near the first reference state, the objective and gradient are smooth on the relevant sign sheet. The search direction and first trial point depend continuously on the initialization. A sufficiently small perturbation therefore preserves the first decision. Once that decision is fixed, the deterministic bracket update produces the same next trial length. Repeating this argument preserves the entire first line search.

At acceptance, the positive curvature product remains bounded away from zero in a sufficiently small neighborhood of the reference state. The BFGS update is continuous there. The same holds for the L-BFGS two-loop recursion and its standard positive curvature scaling. This permits the induction to continue through the next line search. Intersecting the finitely many neighborhoods required through return R gives an open neighborhood of $\sigma_m^+ a$ containing a ball of radius $\varepsilon_R > 0$.

For full BFGS, all these operations are jointly continuous in θ_0 and H_0. The cone of symmetric positive-definite matrices is open and contains I, so the same argument gives a product neighborhood of $(\sigma_m^+ a, I)$. Shrinking its factors gives the stated joint conclusion. The minimum margin can decrease with R; the argument requires positivity only for the finite execution under consideration.

C Scalar Control-State Specialization and the Pinball Identity

Line-search convention. At a differentiable point θ with descent direction p, the search accepts a differentiable trial point if

$$\mathcal{F}(\theta + tp) < \mathcal{F}(\theta) + c_1 t \nabla\mathcal{F}(\theta)^\top p, \qquad \nabla\mathcal{F}(\theta + tp)^\top p > c_2 \nabla\mathcal{F}(\theta)^\top p.$$

These are the strict tests in the Lewis–Overton search. Initialize the lower and upper brackets by $\alpha = 0$ and $\beta = +\infty$, and test $t = 1$. An Armijo failure (A) sets $\beta = t$. If Armijo passes but the curvature test or differentiability requirement fails (W), set $\alpha = t$. Otherwise accept (S). Following a rejection, set

$$t \leftarrow \begin{cases} (\alpha + \beta)/2, & \beta < +\infty, \\ 2\alpha, & \beta = +\infty. \end{cases}$$

Each newly tested t contributes one unit to the trial clock.

Branch windows. For the scalar objective, write $p = -Hg_u(x)$, where $H > 0$ is the active inverse curvature, and set $\sigma = -x/p$. The direction points toward zero. Before crossing, the directional derivative is unchanged, so the curvature test fails because $c_2 < 1$. Armijo holds there because $c_1 < 1$. At $t = \sigma$, the trial is nondifferentiable and is assigned W.

If $x > 0$, then $p = -H$ and $\sigma = x/H$. For $t > \sigma$,

$$f_u(x + tp) = u(tH - x).$$

The Armijo test reduces to

$$(u + c_1)tH < (u + 1)x,$$

equivalently $t < \Theta_+\sigma$. After crossing, the directional derivative is $uH > 0$, so the curvature test holds. If $x < 0$, then $p = uH$ and $\sigma = -x/(uH)$. For $t > \sigma$, $f_u(x + tp) = x + tuH$, and Armijo becomes

$$(1 + c_1 u)tuH < -(u + 1)x,$$

equivalently $t < \Theta_-\sigma$. The directional derivative after this crossing is also positive. These calculations prove (4); in particular, an Armijo equality is an A decision.

Phase update. At an accepted crossing, the scalar secant equation gives

$$H_+ = \frac{x_+ - x}{g_u(x_+) - g_u(x)} = \begin{cases} tH/(u + 1), & x > 0, \\ tuH/(u + 1), & x < 0. \end{cases}$$

Substituting $x_+ = x + tp$ into the next phase produces $F_{+,t}$ and $F_{-,t}$ in (5).

Pinball representation. For $\tau = u/(u + 1)$,

$$\ell_\tau(-(u + 1)x) = \max\{-ux, x\} = f_u(x).$$

The gradients also agree away from the kink. Thus this loss identity preserves the optimizer trajectory when the parameter and inverse-curvature initializations agree.

D Complete B_m Construction and State Scaling

Closing the phase. Let $h = 2^{-m}$. Since $2h \le 1$,

$$D = (u + 1)^2 - 2hu \ge (u + 1)^2 - u = u^2 + u + 1 > 0.$$

Substitution of (6) into the two accepted-step maps gives

$$\frac{u + 1}{u}\left(1 - \frac{\sigma_m^+}{2}\right) = \frac{h(u^2 - 1)}{D} = \sigma_m^-, \qquad (u + 1)\left(1 - \frac{\sigma_m^-}{h}\right) = \frac{2(u + 1)(u + 1 - hu)}{D} = \sigma_m^+.$$

Thus accepted trials 2 and h restore the phase and starting sheet. Appendices E and F establish that the line search reaches these trials through exactly $\mathsf{WS}/\mathsf{A}^m\mathsf{S}$.

Scaling the active state. For an arbitrary scalar initial inverse curvature $H_0 > 0$, start at $x_0 = H_0\sigma_m^+$. Acceptance of trial 2 gives

$$x_1 = H_0(\sigma_m^+ - 2) < 0, \qquad H_1 = \frac{2H_0}{u+1}.$$

Acceptance of trial h then gives

$$x_2 = x_1 + huH_1 > 0, \qquad H_2 = \frac{huH_1}{u+1} = \frac{2hu}{(u+1)^2}H_0.$$

The restored phase implies $x_2 = H_2\sigma_m^+$, and hence

$$(x_2, H_2) = q_m(x_0, H_0), \qquad q_m = \frac{2hu}{(u+1)^2}.$$

Here H_j denotes the scalar active inverse curvature. In the multidimensional embedding, the active block scales by q_m while the orthogonal block remains $P_\perp$.

The trial-normalized limit. Let z_n be the objective value at the nth line-search trial, including rejected trials. Every trial in B_m avoids the kink, so $z_n > 0$. Positive homogeneity implies that corresponding trial points and objective values in successive returns scale by q_m. For $r = 1, \dots, N_m$ and $k \geq 0$,

$$z_{r+kN_m} = q_m^k z_r, \qquad N_m = m + 3.$$

Taking $(r + kN_m)$th roots in each of the finitely many residue classes gives

$$\lim_{n\to\infty} z_n^{1/n} = q_m^{1/N_m}.$$

The same argument applies to accepted objective values indexed by the cumulative trial count. Any fixed initial evaluation cost leaves this limit unchanged. This proves the contraction and rate assertions of Theorem 3.5 once the branch certificates are established.

E Literal Inequalities and Zero-Armijo Maximality

At $c_1 = 0$, the six branch margins are

$$\sigma_m^+ - 1 = \frac{(u+1)^2 - 2hu^2}{D}, \tag{12}$$

$$2 - \sigma_m^+ = \frac{2hu(u-1)}{D}, \tag{13}$$

$$\frac{u+1}{u}\sigma_m^+ - 2 = \frac{2\{(u+1)^2 - hu(u^2+1)\}}{uD}, \tag{14}$$

$$h - \sigma_m^- = \frac{2h(u+1-hu)}{D}, \tag{15}$$

$$(u+1)\sigma_m^- - h = \frac{h\{(u+1)^2(u-2) + 2hu\}}{D}, \tag{16}$$

$$2h - (u+1)\sigma_m^- = \frac{h\{(u+1)^2(3-u) - 4hu\}}{D}. \tag{17}$$

The first numerator is at least $(u+1)^2 - u^2 = 2u + 1$. The second is positive for $u > 1$. For the third, $h \leq 1/2$ gives

$$(u+1)^2 - hu(u^2+1) \geq \frac{-u^3 + 2u^2 + 3u + 2}{2}.$$

The polynomial $-u^3+2u^2+3u+2$ decreases on $[2, u_*]$ and, using $u_*^3 = u_*^2+3u_*+3$, has endpoint value $u_*^2 - 1 > 0$. Thus the displayed lower bound is at least $(u_*^2 - 1)/2 > 0$. The fourth numerator is at least $2h(1 + u/2) > 0$. The fifth is positive for $u \geq 2$, including $u = 2$ because $h > 0$.

For the sixth margin,

$$(u+1)^2(3-u) - 4hu \geq (u+1)^2(3-u) - 2u = -u^3 + u^2 + 3u + 3.$$

The last expression is nonnegative exactly up to u_*. It vanishes only when $u = u_*$ and $h = 1/2$, corresponding to $m = 1$. At that point the preceding negative trial lies on an Armijo-equality boundary and remains an A trial because Armijo is strict. All six margins are strict on $[2, u_*)$.

The polynomial $u^3 - u^2 - 3u - 3$ has negative value at 2, positive value at 3, and strictly positive derivative for $u \geq 2$, so u_* is unique. If $u > u_*$, margin (17) is negative already for $m = 1$. If $u < 2$, the numerator in (16) tends to $(u+1)^2(u-2) < 0$ as $h \downarrow 0$, so the final negative acceptance fails for all sufficiently large m. Hence $[2, u_*]$ is the maximal common literal interval.

F Positive-Armijo Hierarchy and Maximal Height

Write $n = 2^m$, $c = c_1$, and

$$\Delta_m = n(u+1)^2 - 2u = nD.$$

After clearing positive denominators, the three nontrivial branch certificates are

$$P_W = n(1+u)^2(1-c) - u^3 - u + 2uc, \tag{18}$$

$$P_S = n(1+u)^2(u-2-cu) + 2u(1+cu), \tag{19}$$

$$P_A = n(1+u)^2(3-u+2cu) - 4u(1+cu). \tag{20}$$

They satisfy

$$\sigma_m^+ \Theta_+ - 2 = \frac{2P_W}{\Delta_m(u+c)}, \tag{21}$$

$$\sigma_m^- \Theta_- - \frac{1}{n} = \frac{P_S}{n\Delta_m(1+cu)}, \tag{22}$$

$$\frac{2}{n} - \sigma_m^- \Theta_- = \frac{P_A}{n\Delta_m(1+cu)}. \tag{23}$$

Thus $P_W > 0$ certifies positive-side acceptance at 2, $P_S > 0$ certifies negative-side acceptance at $1/n$, and $P_A > 0$ certifies Armijo failure at the preceding trial $2/n$. Once $2/n$ fails Armijo, every earlier larger negative trial also fails.

Solving $P_S > 0$ for c gives the finite-m ceiling

$$c < C_{S,m}(u) = \frac{n(1+u)^2(u-2) + 2u}{u\{n(1+u)^2 - 2u\}}.$$

Since

$$C_{S,m}(u) = \frac{u-2}{u} + \frac{2(u-1)}{\Delta_m},$$

it decreases to $(u-2)/u$. Therefore $c \leq (u-2)/u$ passes for every finite m, including equality, while $c > (u-2)/u$ fails for a sufficiently large finite m.

It remains to verify the other polynomials on $21/10 \leq u \leq 13/6$. The coefficient of c in P_W is negative, so its minimum over $0 \leq c \leq (u-2)/u$ occurs at the ceiling. At $n = 2$,

$$uP_W\left(2, u, \frac{u-2}{u}\right) = -u^4 + 5u^2 + 4u + 4.$$

This polynomial decreases on the stated interval and has endpoint value $18275/1296 > 0$ at $u = 13/6$. The coefficient of c in P_A is $2u\{n(u+1)^2 - 2u\} > 0$, so P_A is minimized at $c = 0$ and $n = 2$. Its minimum is

$$P_A(2, 13/6, 0) = \frac{869}{108} > 0.$$

The remaining crossing and lower-window guards are (12), (13), and (15), which remain strict on this interval.

For the closed Cartesian box $0 \leq c \leq 1/21$, the slopes in n are nonnegative, with a possible zero slope only for P_S at the limiting corner. Direct endpoint differentiation reduces the minima to

$$P_W(2, 13/6, 1/21) = \frac{10537}{1512} > 0,$$

$$P_S(2, 21/10, 1/21) = \frac{231}{50} > 0, \qquad P_A(2, 13/6, 0) = \frac{869}{108} > 0.$$

Since $(u-2)/u$ is increasing and

$$\min_{u\in[21/10,13/6]} \frac{u-2}{u} = \frac{1}{21},$$

no taller closed Cartesian box can work for all m. A convenient open box is

$$\frac{21}{10} < u < \frac{13}{6}, \qquad 0 < c_1 < \frac{1}{21}, \qquad \frac{1}{21} < c_2 < 1.$$

There is no $m = 1$ exception. At the limiting pointwise ceiling, every finite m passes strictly, although no positive branch margin is uniform over all m.

G Rate Ordering and Family Asymptotics

Let

$$A(u) = \frac{(u+1)^2}{u}.$$

Then

$$q_m(u) = \frac{2^{1-m}}{A(u)}$$

and

$$\log_2 \rho_m(u) = -1 + \frac{4 - \log_2 A(u)}{m+3}. \tag{24}$$

On every parameter interval in Theorem 3.5, $4 < A(u) < 16$. The correction term in (24) is therefore positive and strictly decreases with m. This proves

$$\rho_1 > \rho_2 > \cdots > \frac{1}{2}$$

and the limit $1/2$. Reciprocity gives

$$\lambda_m(u) = A(u)2^{m-1},$$

which grows exponentially.

The phases are distinct as well. Treating (6) as functions of h, differentiation shows that σ_m^+ decreases with h, whereas σ_m^- increases with h. Since $h = 2^{-m}$ decreases,

$$\sigma_m^+ \uparrow 2, \qquad \sigma_m^- \downarrow 0.$$

H General Feasible-Word Calculus

Let ω be a fully resolved sequence of outer line-search signatures. Suppose its accepted crossed trials are $t_1, \ldots, t_p$, with sign sheets $\epsilon_1, \ldots, \epsilon_p$. Define

$$F_j(\sigma) = -\frac{\kappa_{\epsilon_j}}{t_j}\sigma + \kappa_{\epsilon_j}, \qquad \kappa_+ = \frac{u+1}{u}, \quad \kappa_- = u+1.$$

Starting from $A_0 = 1$ and $B_0 = 0$, the affine coefficients obey

$$A_j = -\frac{\kappa_{\epsilon_j}}{t_j}A_{j-1}, \qquad B_j = \kappa_{\epsilon_j} - \frac{\kappa_{\epsilon_j}}{t_j}B_{j-1}.$$

Thus $R_\omega(\sigma) = A_p\sigma + B_p$ is affine in σ.

For an even sign-return word, p is even and half of the updates cross from each sign. Multiplying the secant-update factors shows that the inverse-Hessian scalar satisfies $H_{\text{out}} = q_\omega H_{\text{in}}$, where

$$q_\omega = \left(\prod_{+\to-} \frac{t_j}{u+1}\right)\left(\prod_{-\to+} \frac{t_j u}{u+1}\right) = \frac{u^{p/2}\prod_j t_j}{(u+1)^p}.$$

Multiplying the affine slopes gives the phase-return derivative

$$A_p = \frac{(u+1)^p}{u^{p/2}\prod_j t_j} = q_\omega^{-1} = \lambda_\omega.$$

When $A_p \neq 1$, the affine return has the candidate fixed phase

$$\sigma_\omega^* = \frac{B_p}{1 - A_p}.$$

When $A_p = 1$, the fixed-point equation instead requires $B_p = 0$; feasibility of the literal word does not force this equality, so a word with $q_\omega = A_p = 1$ may have no fixed phase.

These algebraic formulas are not by themselves existence certificates. A reported cycle must satisfy all of the following guards:

1. $A_p \neq 1$ and $\sigma_\omega^* > 0$;
2. every intermediate phase is positive;
3. every literal branch satisfies
$$\mathsf{S}:\ \sigma < t < \Theta_\epsilon \sigma, \qquad \mathsf{A}:\ t \geq \Theta_\epsilon \sigma, \qquad \mathsf{W}:\ 0 < t \leq \sigma;$$
4. the doubling–bisection recurrence generates the stated trials;
5. every accepted trial crosses the kink under the stated convention;
6. the evaluation clock includes every rejected and accepted trial;
7. $0 < q_\omega < 1$.

Only after these conditions hold does the word define a nonzero scaling cycle. The returned sign and phase then make the active pair (x, H) contract by q_ω, while the orthogonal BFGS block remains unchanged, and

$$\rho_\omega = q_\omega^{1/N_\omega}$$

is its accepted-value rate indexed by cumulative trial count. The last guard also follows from strict Armijo for a genuine nonzero scaling cycle: each accepted step decreases the positive objective, while a full return multiplies it by q_ω. Formal candidates with $q_\omega \geq 1$ are therefore infeasible.

After clearing denominators whose signs are fixed by the guards, a complete word defines a finite semialgebraic feasibility system. Equality cases form lower-dimensional boundary cells and are not merged with open cycle cells. A coarse alphabet that suppresses inner depth may represent countably many affine pieces; finiteness requires an explicit trial or depth budget.

A general formal word may contain a trial exactly at the kink. Its all-trial value sequence can then contain periodic zeros, so an ordinary positive root limit need not exist. Such a word requires an accepted-value rate or an explicit limsup or continuation convention. The B_m words contain no kink trials, so their ordinary all-trial limits are defined.

I Return Cylinders and Finite-Prefix Ambiguity

Fix the parameters and a strict one-period feasibility interval J_ω containing a fixed phase σ_ω^*, with $\lambda_\omega > 1$. For brevity write $J = J_\omega$, $R = R_\omega$, $\lambda = \lambda_\omega$, and $\sigma^* = \sigma_\omega^*$. The affine return satisfies

$$R(\sigma) - \sigma^* = \lambda(\sigma - \sigma^*).$$

Complete returns. Executing $r \geq 1$ complete copies requires the phase at the beginning of each copy to belong to J. Thus the exact set of initial phases is

$$C_r = \bigcap_{j=0}^{r-1} R^{-j}(J).$$

Each inverse image is a contraction of J about its interior point σ^*. These intervals are nested, so

$$C_r = R^{-(r-1)}(J) = \sigma^* + \lambda^{-(r-1)}(J - \sigma^*), \qquad |C_r| = |J|\lambda^{-(r-1)}.$$

Requiring the phase after the last copy also to lie in J adds the condition $R^r(\sigma) \in J$ and gives width $|J|\lambda^{-r}$. This proves Corollary 4.7.

Partial trial prefixes. Let J_s be the set of positive phases that execute the first s trials of the same strict trace, where $0 \le s < N_\omega$ and $J_0 = (0, \infty)$. In particular, $J \subseteq J_s$. For $r \ge 1$, the set that executes r complete copies followed by these s trials is

$$\begin{aligned} C_{r,s} &= R^{-(r-1)}(J) \cap R^{-r}(J_s) \\ &= R^{-r}\big(R(J) \cap J_s\big). \end{aligned}$$

Consequently,

$$|C_{r,s}| = \gamma_s \lambda^{-r}, \qquad \gamma_s = |R(J) \cap J_s|.$$

The constant is independent of r and includes the final partial-trace condition. For $s = 0$, $R(J) \subset (0, \infty)$ because every full copy ends at a positive phase; hence $\gamma_0 = \lambda|J|$, recovering the complete-return formula. With no completed return ($r = 0$), the initial prefix set is simply J_s. This makes explicit both the residue-dependent constant and the one-return indexing shift.

J Bounded Enumeration of Feasible Regimes

For each inner length $1 \le n \le 11$, the bounded enumeration considers every string in $\{\mathsf{A}, \mathsf{W}\}^{n-1}\mathsf{S}$ together with its deterministic dyadic trial sequence generated by the literal doubling–bisection rule from $t_{\text{init}} = 1$. It forms alternating-sign outer occurrence sequences for both starting signs, with

$$N \le 12, \qquad K \in \{2, 4, 6, 8, 10, 12\},$$

where N is the total trial count and K is the even number of accepted steps. The quotient identifies cyclic rotations of the full signed outer occurrence sequence only. It does not identify reversal, and it retains repetitions of shorter words as longer occurrences.

The raw count is

$$\sum_{\substack{2 \le K \le N \le 12 \\ K \text{ even}}} 2^{N-K+1} \binom{N-1}{K-1}.$$

Indeed, a composition of N into K positive inner lengths has $\binom{N-1}{K-1}$ choices, the rejected trials have 2^{N-K} possible letter assignments, and the starting sign has two choices.

The exact probe set is

$$\mathcal{P} = \left\{ \left(\frac{19}{10}, 0\right), (2, 0), \left(\frac{21}{10}, 0\right), \left(\frac{19}{10}, \frac{1}{20}\right), \left(2, \frac{1}{20}\right), \left(\frac{21}{10}, \frac{1}{20}\right), \left(\frac{21}{10}, \frac{1}{100}\right), \left(2, \frac{1}{100}\right) \right\}.$$

At each $(u, c_1) \in \mathcal{P}$, exact rational arithmetic reconstructs every dyadic trial by literal doubling and bisection and composes the accepted-step maps into

$$R_\omega(\sigma) = A_\omega \sigma + B_\omega.$$

When $A_\omega \ne 1$, it computes $\sigma^*_\omega = B_\omega/(1 - A_\omega)$ and propagates every intermediate phase. An occurrence passes only if all phases are positive, the literal conditions

$$\mathsf{W}: \ 0 < t \le \sigma, \qquad \mathsf{S}: \ \sigma < t < \Theta_\epsilon \sigma, \qquad \mathsf{A}: \ t \ge \Theta_\epsilon \sigma$$

hold with these exact endpoint conventions, every required kink crossing occurs, the reconstructed trial count equals N, and $0 < q_\omega < 1$. A cyclic class is probe-feasible precisely when at least one of its signed occurrences passes every predicate at at least one declared probe. The output retains occurrence labels, exact predicates, phases, trials, q_ω, and the list of passing probes.

The resulting enumeration gives $265{,}708$ raw occurrences, $68{,}388$ cyclic classes, and 152 probe-feasible classes. Exact probe feasibility does not certify an entire parameter interval.

Within the declared budget, the enumeration recovers $B_1, \ldots, B_9$, including their phases, accepted trials, contractions, and trial counts. The all-m statement is established analytically by Theorem 3.5; the bounded enumeration provides an independent finite cross-check.

K Experimental Protocols and Additional Results

For every finite-precision experiment, N_f counts objective evaluations and N_g counts gradient evaluations. Initial objective and gradient evaluations are included. Each line-search trial evaluates the objective and evaluates the gradient after Armijo passes at a differentiable trial point. The theoretical trial count N_m is separate from this experimental accounting.

Table 2: Counts from the bounded trace enumeration with $N \leq 12$.

Accepted steps K	Raw occurrences	Cyclic classes	Probe-feasible classes
2	40,962	20,481	21
4	131,074	32,833	76
6	80,386	13,409	54
8	12,802	1,615	1
10	482	49	0
12	2	1	0
Total	265,708	68,388	152

K.1 Protocols for the main-text experiments

Q1: multidimensional and limited-memory regime invariance. We use

$$F(\theta) = f_2(a^\top \theta) + \frac{1}{2}\|P_\perp \theta\|_2^2,$$

corresponding to the multidimensional objective in Section 3.1 with $\mu = 1$. The evaluation grid is

$$d \in \{1, 2, 8, 32, 128, 512\}, \qquad m \in \{1, 2, 4, 8, 16\},$$

with full inverse BFGS and L-BFGS memories $M \in \{1, 10, 100\}$. Every configuration includes the coordinate-aligned active direction, and each $d > 1$ configuration additionally uses eight fixed random rotations shared across optimizers and regime indices.

Each run covers four complete two-step returns. A run passes the strict trace gate only when every branch label, trial value, and accepted trial agrees with the rational reference trace. Continuous state error, rate error, and orthogonal drift are recorded separately. The suite contains 920 runs. All 736 runs with $m \leq 8$ reproduce the exact trace, while the 184 runs with $m = 16$ reach the binary64 precision limit. Across the rotated configurations, the maximum normalized orthogonal drift remains on the order of 10^{-15}.

Q2: full-dimensional finite-horizon shadowing. We use the same $\mu = 1$ objective with

$$d \in \{2, 8, 32, 128\}, \qquad m \in \{1, 2, 4, 8\}.$$

Three perturbation families are considered: parameter-only, inverse-Hessian-only, and joint perturbations. Each configuration uses 27 logarithmic perturbation scales from 10^{-14} to 10^{-1} and 16 fixed direction/optimizer-state seed pairs.

A return is counted as shadowed only when every dyadic trial value, branch decision, accepted trial, and return boundary agrees with the exact reference. Let T_{shadow} denote the number of complete returns before the first mismatch. We report

$$S_R(\varepsilon) = \Pr[T_{\mathrm{shadow}} \geq R],$$

together with $\varepsilon_{95}(R)$, the largest grid radius with at least 95% survival. With 16 directions, this requires $16/16$ survivors.

The complete suite contains 20,736 random-direction runs. For the representative $d = 32, m = 4$ joint configuration, the largest grid scales with full survival are

$$10^{-3}, \quad 3.162 \times 10^{-5}, \quad 3.162 \times 10^{-8}, \quad 10^{-14}$$

for $R = 1, 2, 4, 8$, respectively. At $R = 4$, $\varepsilon_{95} = 10^{-3}$ for parameter-only perturbations and 3.162×10^{-8} for both inverse-Hessian-only and joint perturbations.

Q3: finite-precision forecasting realization of the exact hierarchy. The controlled forecasting construction fixes

$$P = 4, \qquad T = 64, \qquad \tau = \frac{2}{3}, \qquad y_t = \sin(2\pi t/P),$$

with predictor

$$\widehat{y}_{t+1} = -y_{t-1} + 3\theta.$$

The resulting forecasting objective satisfies the exact identity

$$L(\theta) = 64 f_2(\theta).$$

For each $m = 1, \ldots, 9$, four complete returns are replayed in exact rational arithmetic and binary64. The exact check records the complete line-search word, trial count, state ratio, inverse-Hessian ratio, and kink avoidance. Binary64 reproduces every discrete branch word through $m = 9$. The continuous state-ratio tolerance is 10^{-12} and is first crossed at $m = 4$; a compensated replay is used to diagnose the resulting cancellation error.

Q4: initialization-dependent computation in real learning settings. *Real electricity forecasting.* The real-data experiment uses the ETTm1 split and train-only standardization with the univariate OT channel, context length 48, horizon 12, and pinball loss at $\tau = 2/3$. DLinear uses moving-average kernel 25 with separate seasonal and trend linear maps and a combined intercept, giving 1,164 parameters across the horizon heads.

Because this parameterization is rank deficient, we fix a canonical representative of the near-optimal solution face and construct the initialization directions around this reference. A deterministic two-dimensional plane defines 64 equal-distance directions on each radius. All runs use $H_0 = I$, line-search parameters

$$(c_1, c_2) = (10^{-4}, 0.9),$$

a common stopping target, and $N_f + N_g$ oracle accounting. The tested relative radii are

$$r/S \in \{0.01, 0.03, 0.10\}.$$

All 192 runs reach the common target. The main-text geometry analysis uses the smallest radius together with the identifiable-subspace projection. At this radius, the Pearson and Spearman correlations between projected initialization norm and oracle cost are 0.341 and 0.348, respectively.

Smooth-objective transfer. To remove the nonsmooth pinball kink, we replace the loss by the pseudo-Huber quantile objective

$$\ell_{\tau,\delta}(r) = \left(\tau - \frac{1}{2}\right) r + \frac{1}{2}\left(\sqrt{r^2 + \delta^2} - \delta\right), \qquad \delta > 0, \tag{25}$$

with $\tau = 2/3$ and

$$\delta \in \{10^{-3}, 10^{-2}, 10^{-1}\}.$$

For every $\delta > 0$, the loss is C^∞ and

$$\ell''_{\tau,\delta}(r) = \frac{\delta^2}{2(r^2 + \delta^2)^{3/2}}.$$

The data split, standardization, DLinear construction, and forecasting horizon are inherited from the preceding experiment. Optimization is performed in the rank-49 identifiable quotient of the 97-column DLinear design for each of the 12 horizons, giving effective dimension

$$12 \times 49 = 588.$$

The quotient preserves the predictor class while removing the exact structural nullspace. The reconstruction and orthonormality errors are 7.11×10^{-15} and 2.44×10^{-15}, respectively.

For each smoothing scale, 64 equal-distance initializations are placed in a fixed two-dimensional plane around a reference optimum using $r/S = 0.01$. Full inverse BFGS starts from $H_0 = I_{588}$ and uses the same weak-Wolfe line search with $(c_1, c_2) = (10^{-4}, 0.9)$. Runs terminate when

$$F(\theta) \le F^\star + 10^{-6} \max(1, |F^\star|).$$

The evaluation budgets are 8,192 combined objective/gradient requests, 4,096 line-search trials, and 1,000 outer iterations. All 192 runs reach the common target.

Table 3: **Smooth-loss transfer across pseudo-Huber scales.** Cost is $N_f + N_g$ to the common target.

δ	Success	Min	Q1	Median	Q3	P90	Max	Max/Min
10^{-3}	64/64	49	55.75	59	61	62	66	1.347×
10^{-2}	64/64	47	53	54.5	59	60.7	64	1.362×
10^{-1}	64/64	26	34	35	40	42.7	46	1.769×

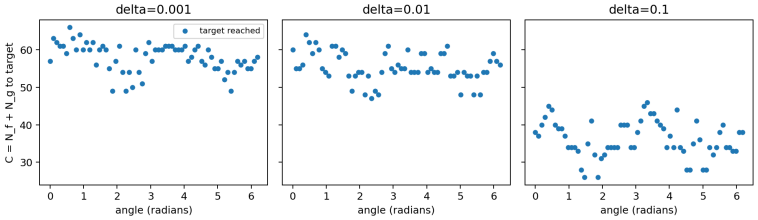


Figure 5: **Initialization-dependent computation under smooth quantile losses.** Optimization cost across 64 equal-distance initialization directions for each pseudo-Huber smoothing scale.

Learned-representation transfer. The representation-transfer experiment uses a fixed PatchTST representation and optimizes a 9,228-dimensional convex quantile head on the same ETTm1 task. The affine head is full-column-rank, so this setting has no exact parameter nullspace.

The initialization protocol follows the real-forecasting experiment: 64 equal-distance directions are placed at each of three relative radii around a fixed reference, with $H_0 = I$, the same weak-Wolfe parameters, and the same $N_f + N_g$ accounting. Full BFGS reaches the common target for all 64 small-radius starts, with min/median/max cost

$$3932/4053.5/4179.$$

The $r/S = 0.03$ and $r/S = 0.10$ configurations reach the outer evaluation budget; their terminal gaps are retained as finite-budget diagnostics.

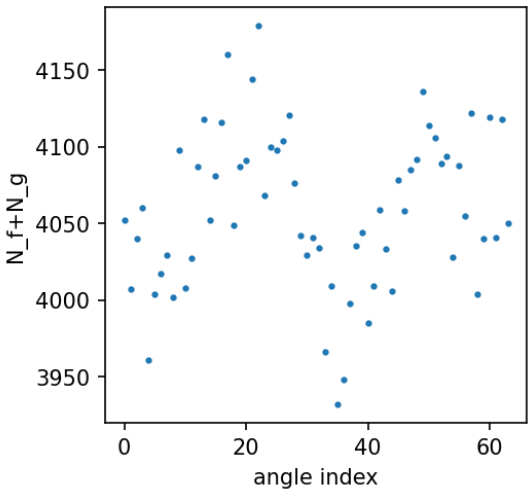


Figure 6: **Learned-representation transfer at the small radius.** Oracle cost $N_f + N_g$ across 64 equal-distance initialization directions at $r/S = 0.01$.

Q5: transfer to native strong-Wolfe L-BFGS. The final transfer uses the same learned-representation reference and the same 64 small-radius initialization directions as Q4. The optimizer is native `torch.optim.LBFGS(lr=1, history_size=100, line_search_fn='strong_wolfe')`, with its native tolerances and the same common target.

The initial evaluation budget is $4{,}096$ closure calls. A follow-up run increases the budget to $6{,}144$ while keeping the reference, initializations, optimizer, target, and stopping rule unchanged. Within this window, time-to-target is observed for $51/64$ directions. Their closure costs have

$$4265/4307/4370/4464.5/4517/4821.5/5358$$

for min/Q1/median/Q3/P90/P95/max, corresponding to a max/min spread of $1.25627\times$. The remaining 13 directions are right-censored at the $6{,}144$-closure budget and finish with objective gaps between 1.00047×10^{-6} and 1.04558×10^{-6}.