Smooth Spectral Operators for Indefinite Hessians: Relative Resolution, Stability, and Matrix-Free Approximation

Anonymous authors
Paper under double-blind review

Abstract

Curvature-aware optimization in high-dimensional non-convex problems relies on matrix-free spectral transformations of indefinite Hessians. A fundamental difficulty in designing these operators is that a regularization parameter measured in absolute curvature units changes its operational meaning when the overall objective scale changes. Under homogeneous curvature rescaling, a fixed absolute parameter alters the relative transition geometry and arbitrarily changes the matrix-free computational difficulty.

To resolve this ambiguity, we introduce relative resolution, parameterizing the spectral transition scale relative to an admissible reference spectral scale. This dimensionless formulation yields a normalized spectral geometry whose polynomial approximation complexity is independent of the absolute curvature scale under admissible normalization. Our analysis connects relative resolution to fixed-τ operator-perturbation stability, finite polynomial-approximation budgets, and the operational scale mismatch. Experiments support the predicted approximation-complexity scaling, quantify sensitivity to operational scale mismatch, and characterize the distinct realized-action regimes observed across controlled non-convex and stiff physics-informed optimization settings.

1 Introduction

In high-dimensional non-convex optimization, exploiting indefinite curvature often requires transforming the Hessian spectrum rather than treating curvature as uniformly positive or negative. However, explicitly forming and decomposing the Hessian is computationally prohibitive at scale. Consequently, spectral curvature methods rely on matrix-free approximations, realizing the target spectral operator exclusively through Hessian-vector products (HVPs).

A fundamental challenge in matrix-free spectral transformations is the treatment of near-zero curvature. To address this, we replace the singular ideal inverse-absolute-curvature response by a smooth operator $M_\tau(H)$. Conceptually, this operator regularizes the spectrum via a scalar response such as $1/\sqrt{\lambda^2+\tau^2}$, where a positive scale parameter τ determines the width of the near-zero transition. However, a fixed numerical τ lacks an invariant spectral meaning: under homogeneous curvature rescaling ($H \mapsto cH$), the same τ occupies a different relative position in the spectrum, altering both the operator's shape and its effective spectral resolution.

Crucially, in a matrix-free setting, this scale ambiguity extends beyond operator design to become a computational bottleneck. Matrix-free methods realize nonlinear operators via polynomial or Krylov subspace approximations, whose difficulty is governed by the operator's transition geometry relative to the accessible spectrum. When the absolute curvature scale changes, a fixed τ inadvertently alters this relative geometry. As a result, the same absolute parameter changes both the spectral transformation being targeted and the approximation difficulty—and therefore the matrix-free computational budget required to realize it at a prescribed fidelity.

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.

To resolve this entanglement, we introduce *relative resolution*, the dimensionless ratio $\alpha = \tau/L^{\text{th}}$, where L^{th} denotes an admissible theoretical normalization scale. This parameterization separates normalized spectral geometry from absolute curvature magnitude. Under homogeneous curvature rescaling, a fixed α, together with a consistently rescaled normalization scale, preserves the normalized operator geometry and its relative transition resolution. In practice, however, L^{th} is not directly available in practice, and its normalization role is realized by a matrix-free estimate $\widehat{L}$; this gap between theoretical normalization and empirical scale realization becomes a central part of the approximation problem.

Establishing this principle requires validation beyond improved optimization trajectories. As outlined in Fig. 1, we study its consequences for operator stability, finite-budget matrix-free approximation, and empirical scale realization, before evaluating its macroscopic consequences for end-to-end optimization.

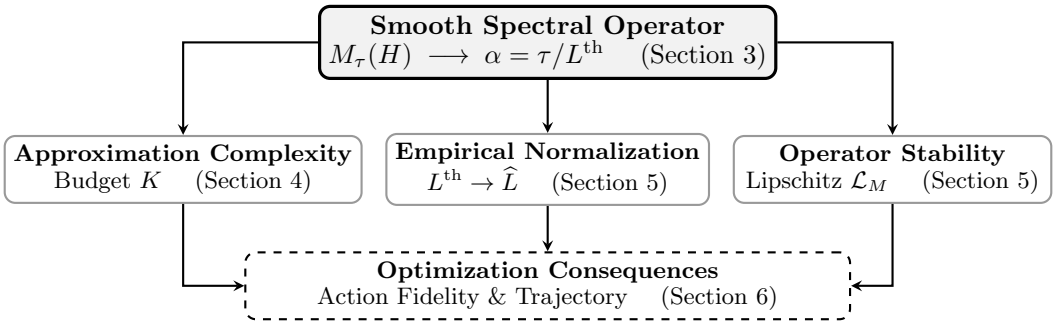


Figure 1: **The validation hierarchy of the relative-resolution framework.**

Contributions. We introduce a scale-consistent framework for analyzing and realizing regularized spectral transformations under finite matrix-free budgets, instantiated through Smooth Saddle-Free Newton.

- **Scale-consistent parameterization.** We introduce the relative resolution $\alpha = \tau/L^{\text{th}}$, yielding a dimensionless normalized formulation whose spectral geometry and polynomial approximation complexity are independent of the absolute curvature scale under admissible normalization.
- **Operator stability and finite-budget approximation.** We characterize the operator's fixed-regularization perturbation sensitivity and show that finite-degree Chebyshev approximation is governed by the dimensionless complexity coordinate $K\,\text{arsinh}(\alpha)$.
- **Empirical realization limits.** We distinguish the admissible normalization assumed by the approximation theory from the empirical spectral-scale estimation required in matrix-free optimization, and isolate the realized-action error induced by spectral-scale mismatch.
- **Mechanistic and transfer validation.** Experiments support the predicted approximation-complexity scaling, quantify sensitivity to operational scale mismatch, and examine how their realized finite-budget actions are associated with end-to-end optimization behavior across controlled non-convex and physics-informed settings.

2 Related Work and Positioning

2.1 Curvature-Aware Optimization

Curvature-aware optimization seeks to exploit second-order information in non-convex objectives without explicitly forming and factorizing the Hessian. Hessian-Free methods

(Martens, 2010) use Hessian–vector products (HVPs) with iterative solvers, whereas Saddle-Free Newton (SFN) (Dauphin et al., 2014) modifies indefinite curvature through a nonlinear spectral transformation. Because the ideal inverse-absolute response is singular at zero curvature, practical SFN variants introduce regularization near the origin, thereby specifying the spectral response that must ultimately be realized computationally.

Other scalable second-order methods reduce cost through structured curvature approximations. AdaHessian (Yao et al., 2021) and Sophia (Liu et al., 2024) use inexpensive diagonal curvature estimates, while Low-Rank Saddle-Free Newton (O'Leary-Roseberry et al., 2020) approximates a dominant curvature subspace. Our setting is complementary: rather than introducing an explicit diagonal or low-rank structural surrogate, we access the Hessian implicitly through HVPs and study finite-budget realization of its prescribed spectral transformation.

2.2 Matrix-Free Realization of Spectral Operators

For large problems, spectral transformations can be evaluated as matrix-function actions $f(H)v$ without explicit eigendecomposition. Polynomial and Krylov methods are classical tools for approximating such actions using matrix–vector products (Druskin & Knizhnerman, 1989; Saad, 1992; Higham, 2008), and are therefore compatible with automatic-differentiation HVPs.

Within SFN, Simpson (2022) studies regularized spectral transformations, while Oldewage et al. (2024) develops a series-based HVP realization of the saddle-free update. These constructions expose a finite-budget trade-off: sharper variation in the target spectral response generally requires greater approximation effort. We study this trade-off for Chebyshev realization of the normalized regularized response.

2.3 Scale-Aware Parameterization and Positioning

Existing regularized SFN formulations, such as Simpson (2022), introduce dimensionful regularization scales to smooth or control the near-zero spectral response. Such an absolute parameter changes its relative spectral meaning when the objective and Hessian are homogeneously rescaled.

We instead study how the spectral transition should be parameterized and realized when its resolution is defined relative to an admissible normalization scale. The relative resolution $\alpha = \tau/L^{\text{th}}$ separates normalized spectral geometry from absolute curvature magnitude, yielding under admissible normalization a fixed dimensionless approximation problem. In the practical matrix-free setting, this formulation further distinguishes theoretical spectral normalization from its empirical realization using finite HVPs.

3 The Relative-Resolution Spectral Operator

This section develops a spectral parameterization whose resolution is independent of absolute curvature scale.

3.1 Assumptions and the Singular Ideal

We consider objective functions $f : \mathbb{R}^n \to \mathbb{R}$ of class C^2, with gradient $g(x) = \nabla f(x)$ and Hessian $H(x) = \nabla^2 f(x) \in \mathbb{S}^n$. The operator-level analysis treats H as an element of $\mathbb{S}^n$; no regularity assumption on the map $x \mapsto H(x)$ is needed unless explicitly stated.

In high-dimensional non-convex optimization, Saddle-Free Newton replaces the signed Hessian spectrum by its absolute magnitude, yielding the idealized inverse-absolute-curvature response $|\lambda|^{-1}$ (Dauphin et al., 2014). However, this idealized spectral mapping is singular at $\lambda = 0$: arbitrarily small curvature magnitudes yield unbounded spectral gain, making the ideal operator increasingly sensitive along near-flat eigendirections. We therefore consider a smooth regularization of this spectral response over the symmetric matrix space $\mathbb{S}^n$ that removes this origin singularity while preserving the Hessian eigenspaces.

3.2 Regularized Spectral Operator

At the scalar level, this simply replaces the singular gain $1/|\lambda|$ by the bounded gain $1/\sqrt{\lambda^2+\tau^2}$, while applying the same transformation independently along each Hessian eigendirection. To remove the singularity while maintaining continuous spectral behavior, we consider a positive regularization parameter $\tau > 0$ and define the **smooth spectral operator**

$$M_\tau(H) := (H^2 + \tau^2 I)^{-1/2}. \tag{1}$$

Viewed as a matrix function in the standard functional-calculus sense (Higham, 2008, Chap. 1), M_τ is induced by the scalar response $f_\tau(\lambda) := (\lambda^2+\tau^2)^{-1/2}$. Hence, for a spectral decomposition $H = Q\Lambda Q^\top$, $M_\tau(H) = Q f_\tau(\Lambda) Q^\top$. The operator preserves the eigenspaces of H and, since $f_\tau(\lambda) > 0$ for every $\lambda \in \mathbb{R}$ when $\tau > 0$, remains strictly positive definite even when H is indefinite.

3.3 Relative Parameterization and Dimensionless Spectral Geometry

A fixed regularization parameter τ does not preserve its relative spectral meaning under changes in the absolute curvature scale. Let $L^{\text{th}} \geq \|H\|_2$ denote an admissible theoretical normalization scale, and define the dimensionless **relative resolution**

$$\alpha := \frac{\tau}{L^{\text{th}}}. \tag{2}$$

To express the same spectral transformation in normalized coordinates, define the dimensionless counterpart of M_τ by $\mathcal{M}_\alpha(X) := (X^2+\alpha^2 I)^{-1/2}$. With $X = H/L^{\text{th}}$ and $\tau = \alpha L^{\text{th}}$, the dimensional and normalized forms are related by

$$M_{\alpha L^{\text{th}}}(H) = \frac{1}{L^{\text{th}}} \mathcal{M}_\alpha\left(\frac{H}{L^{\text{th}}}\right). \tag{3}$$

In normalized spectral coordinates $x = |\lambda|/L^{\text{th}} \in [0,1]$, the corresponding scalar response is $(x^2+\alpha^2)^{-1/2}$, which exhibits two limiting regimes (Fig. 2). For $x \gg \alpha$, it approaches x^{-1}, recovering the asymptotic inverse-curvature behavior of the ideal saddle-free response. For $x \ll \alpha$, it flattens toward α^{-1}, producing approximately uniform gradient-like scaling across near-flat directions. Thus, α sets the dimensionless transition scale between resolved curvature adaptation and the regularized near-flat regime.

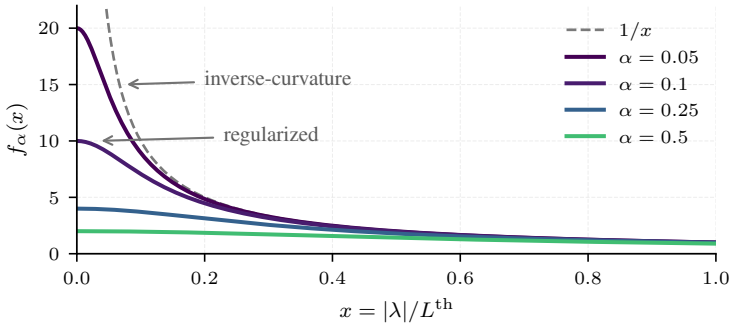


Figure 2: **Normalized spectral response** $(x^2+\alpha^2)^{-1/2}$ **on** $[0,1]$**.** The parameter α controls the transition between the inverse-curvature and regularized near-flat responses.

3.4 Scale Consistency of the Target Direction

Under positive rescaling of the objective ($f \mapsto cf$, $c > 0$), gradient and curvature scales transform as $g \mapsto cg$ and $H \mapsto cH$. Under the corresponding scale transformation $L^{\text{th}} \mapsto cL^{\text{th}}$ and $\tau \mapsto c\tau$, the relative resolution $\alpha = \tau/L^{\text{th}}$ remains invariant. Because the operator satisfies $M_{c\tau}(cH) = c^{-1} M_\tau(H)$, the target direction satisfies

$$M_{c\tau}(cH)(cg) = \left(c^{-1} M_\tau(H)\right)(cg) = M_\tau(H) g. \tag{4}$$

Thus relative parameterization makes the regularized target direction invariant to positive homogeneous rescaling.

4 Scale-Consistent Matrix-Free Approximation

We next analyze finite-budget Chebyshev realization under admissible spectral normalization.

4.1 Normalized Chebyshev Approximation

Using the admissible theoretical scale L^{th} from Section 3.3, define $\widetilde{H} := H/L^{\mathrm{th}}$ and $\tau := \alpha L^{\mathrm{th}}$. Then $\sigma(\widetilde{H}) \subseteq [-1, 1]$. Here, L^{th} fixes the dimensionless approximation interval and need not be available to the practical matrix-free implementation.

By the scale-equivariant factorization in Eq. (3), $M_\tau(H) = (L^{\mathrm{th}})^{-1}\mathcal{M}_\alpha(\widetilde{H})$, where the normalized scalar response is $f_\alpha(x) := (x^2 + \alpha^2)^{-1/2}$. Hence the absolute Hessian scale is separated from the approximation geometry: matrix-free realization reduces to approximating the normalized operator $\mathcal{M}_\alpha$ over $\sigma(\widetilde{H}) \subseteq [-1, 1]$.

We approximate $\mathcal{M}_\alpha$ through a degree-K Chebyshev approximation of its scalar response f_α:

$$P_{K,\alpha}(x) := \sum_{k=0}^{K} c_k(\alpha) T_k(x),$$

with coefficients given by the standard Chebyshev projection (Trefethen, 2013, Chap. 8). Accordingly, $P_{K,\alpha}(\widetilde{H})$ is the finite-degree matrix-function approximation to $\mathcal{M}_\alpha(\widetilde{H})$, yielding the matrix-free action $(L^{\mathrm{th}})^{-1}P_{K,\alpha}(\widetilde{H})g$. Using the standard three-term recurrence, evaluating the degree-K action requires exactly K HVPs. This count excludes HVPs used for empirical scale estimation.

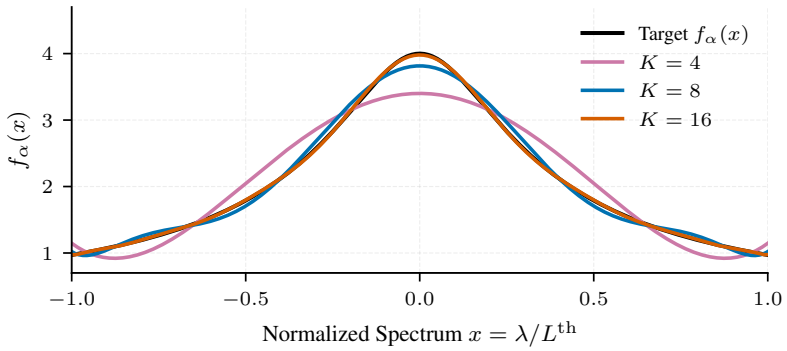


Figure 3: **Chebyshev approximation of the normalized spectral response $f_\alpha(x)$ on $[-1, 1]$.** Increasing the approximation degree K progressively improves the polynomial approximation of the target normalized response.

4.2 Bernstein Geometry and Relative Approximation Complexity

For analytic functions on $[-1, 1]$, Chebyshev approximation errors decay geometrically at a rate governed by the Bernstein ellipses in the domain of analytic continuation (Trefethen, 2013, Chap. 8). For $f_\alpha(z) = (z^2 + \alpha^2)^{-1/2}$, the nearest complex singularities are branch points at $z_\star = \pm i\alpha$, defining the limiting Bernstein ellipse parameter

$$\rho(\alpha) = \alpha + \sqrt{1 + \alpha^2} = e^{\operatorname{arsinh}(\alpha)}. \tag{5}$$

For a nonzero reference y, let $E_y(x) := \|x - y\|_2/\|y\|_2$ denote the relative error under the Euclidean vector norm or spectral matrix norm. By scale equivariance, the absolute normal-

ization factor $(L^{\mathrm{th}})^{-1}$ cancels in relative error, allowing operator fidelity to be evaluated entirely in normalized coordinates by $E_{\mathcal{M}_\alpha(\widetilde{H})}(P_{K,\alpha}(\widetilde{H}))$. Thus, conditional on $\sigma(\widetilde{H}) \subseteq [-1,1]$, the uniform approximation bound can be expressed entirely in normalized coordinates and is independent of the absolute curvature magnitude.

For the compact operating regime $\mathcal{A} = [\alpha_{\min}, \alpha_{\max}]$ $(0 < \alpha_{\min} \le \alpha_{\max} < \infty)$ and any proof-only interior factor $\theta \in (0,1)$, let $\rho_\theta(\alpha) := \rho(\alpha)^\theta = e^{\theta \operatorname{arsinh}(\alpha)}$. The ellipses $\mathcal{E}_{\rho_\theta(\alpha)}$ remain uniformly separated from the branch points and lie strictly within the domain of analyticity. Standard Bernstein bounds yield a finite constant $C_{\mathcal{A},\theta} > 0$, independent of α and K, such that

$$E_{\mathcal{M}_\alpha(\widetilde{H})}\Big(P_{K,\alpha}(\widetilde{H})\Big) \le C_{\mathcal{A},\theta} e^{-\theta K \operatorname{arsinh}(\alpha)}, \qquad \alpha \in \mathcal{A}. \tag{6}$$

The bound therefore gives the sufficient degree condition $K \ge [\theta \operatorname{arsinh}(\alpha)]^{-1} \log(C_{\mathcal{A},\theta}/\epsilon)$, and hence $K = \mathcal{O}_{\mathcal{A},\theta}\Big(\frac{\log(1/\epsilon)}{\operatorname{arsinh}(\alpha)}\Big)$. Thus relative approximation complexity is organized by $K \operatorname{arsinh}(\alpha)$: larger values correspond to better-resolved polynomial realizations, whereas small values identify the difficult finite-budget regime.

Finite-degree approximation does not automatically preserve the strict positive definiteness of the ideal operator. Nevertheless, the same approximation bound yields a sufficient positivity regime. Since $f_\alpha(x) \ge (1+\alpha^2)^{-1/2}$ on $[-1,1]$, Weyl's inequality and $\|\mathcal{M}_\alpha(\widetilde{H})\|_2 \le 1/\alpha$ imply that any relative operator error $\epsilon < \alpha/\sqrt{1+\alpha^2}$ guarantees $P_{K,\alpha}(\widetilde{H}) \succ 0$. Combining this condition with Eq. (6), strict positivity is guaranteed whenever

$$C_{\mathcal{A},\theta} e^{-\theta K \operatorname{arsinh}(\alpha)} < \frac{\alpha}{\sqrt{1+\alpha^2}}.$$

Thus positivity is not guaranteed for arbitrary (K,α), but is ensured once the approximation error is sufficiently smaller than the ideal operator's positive spectral margin. On compact α-ranges bounded away from zero, the action error inherits the same exponential dependence; derivations are collected in Appendix A.

5 Empirical Scale Realization and Operator Sensitivity

The analysis above relies on the admissible theoretical scale L^{th}, whereas a practical matrix-free implementation must estimate an operational scale from HVPs. Let $\widehat{\rho}$ denote the dominant-magnitude spectral estimate obtained from finite persistent power iteration. We define the empirical scale as $\widehat{L} := \gamma\widehat{\rho}$ for a safety margin $\gamma \ge 1$.

Because finite power iteration does not certify $\widehat{\rho} = \|H\|_2$, the resulting empirical scale need not satisfy $\widehat{L} \ge \|H\|_2$. Consequently, part of the normalized spectrum may extend outside $[-1,1]$, the interval on which the uniform approximation guarantee is established. Beyond the loss of this guarantee, such leakage introduces an extrapolation risk: for fixed $|x| > 1$, Chebyshev polynomials satisfy $|T_k(x)| = \cosh\big(k \operatorname{arcosh}|x|\big)$ and therefore grow exponentially with k. The polynomial realization can thus become increasingly sensitive to out-of-interval spectral components. Choosing $\gamma > 1$ provides an empirical buffer against such leakage, but does not make spectral containment certified.

Empirical scale realization affects more than spectral coverage. Since $\tau = \alpha\widehat{L}$, uncertainty in the realized scale also changes the dimensional smoothing of the target operator. We therefore first isolate the Hessian-argument sensitivity of $M_\tau(H)$ at fixed τ. The analysis leverages the spectral matrix-function perturbation framework of Carlsson (2021).

Theorem 1 (Global Frobenius-Lipschitz Bound)**.** *For any fixed $\tau > 0$, $H \in \mathbb{S}^n$, and symmetric perturbation $\Delta H \in \mathbb{S}^n$, the smooth spectral operator $M_\tau : \mathbb{S}^n \to \mathbb{S}^n_{++}$ satisfies:*

$$\|M_\tau(H+\Delta H) - M_\tau(H)\|_F \le \frac{2}{3\sqrt{3}\tau^2}\|\Delta H\|_F \tag{7}$$

Evaluating this fixed-τ bound at $\tau = \alpha\widehat{L}$ gives a worst-case sensitivity scaling of $\mathcal{O}((\alpha\widehat{L})^{-2})$. Thus, for fixed α, a smaller realized scale corresponds to a larger fixed-τ Hessian-argument

sensitivity bound. This does not characterize perturbations of the scale estimator itself or imply a convergence guarantee for the resulting optimization map. The proof of the fixed-τ bound and a restricted-domain extension accounting for scale-induced variation in τ are collected in Appendix B.

6 Experimental Evaluation

Following the analytical structure above, we first isolate finite-degree and scale-realization effects on neural-network Hessian checkpoints, then examine realized spectral actions and their association with end-to-end optimization behavior.

6.1 Mechanism Validation

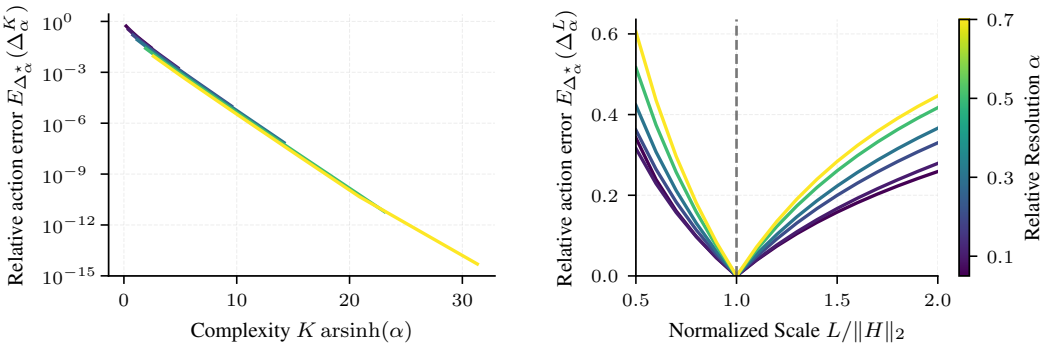
We study two complementary realization mechanisms on a common population of neural-network Hessian checkpoints obtained from Adam (Kingma & Ba, 2015) trajectories of an MLP trained on MNIST. At each checkpoint, we record g, the explicit Hessian H, and the oracle scale $L^\star := \|H\|_2$ for offline analysis (Yao et al., 2020), corresponding to the tight admissible choice of L^{th}. The explicit Hessian is used only for this oracle evaluation. For each (H, g, α), the exact reference action is

$$\Delta_\alpha^\star := M_{\alpha L^\star}(H)g.$$

We perform two controlled interventions on the same checkpoint population using the relative action error.

Finite-degree approximation. We fix $L = L^\star$ and vary the polynomial degree K and relative resolution α, using $\Delta_\alpha^K := (L^\star)^{-1} P_{K,\alpha}(H/L^\star)g$. Across checkpoints, the relative action error $E_{\Delta_\alpha^\star}(\Delta_\alpha^K)$ is largely organized by the dimensionless complexity coordinate $K\,\mathrm{arsinh}(\alpha)$.

Scale mismatch. We replace $L^\star$ by a mismatched scale L and vary $L/L^\star$, holding (H, g, α) fixed, with $\Delta_\alpha^L := M_{\alpha L}(H)g$. The relative action error $E_{\Delta_\alpha^\star}(\Delta_\alpha^L)$ vanishes at $L = L^\star$ and increases as the operational scale moves away from the oracle over the evaluated range.



(a) Finite-degree approximation

(b) Scale mismatch

Figure 4: **Controlled mechanism validation on neural-network Hessian checkpoints.** (a) Relative finite-degree action error $E_{\Delta_\alpha^\star}(\Delta_\alpha^K)$ against the complexity coordinate $K\,\mathrm{arsinh}(\alpha)$ under oracle normalization. (b) Relative scale-mismatch error $E_{\Delta_\alpha^\star}(\Delta_\alpha^L)$ against the normalized operational scale $L/L^\star$. Colors denote the relative resolution α.

6.2 From Spectral Realization to Optimization Behavior

Our primary matrix-free spectral-realization comparator is a controlled series-based realization derived from the truncated inverse-square-root construction of Oldewage et al. (2024),

which we denote Series-HVP. Smooth-SFN uses the empirical scale $L_t = \widehat{L}_t$ obtained from matrix-free spectral estimation. At iteration t, the realized action is

$$\Delta_t := \frac{1}{L_t} P_{K,\alpha}\left(\frac{H_t}{L_t}\right) g_t. \tag{8}$$

The complete matrix-free update, including empirical scale estimation and the Chebyshev recurrence, is specified in Appendix C.

We evaluate end-to-end optimization on two neural-network objectives: the reconstructed CURVES autoencoder benchmark (Hinton & Salakhutdinov, 2006) and a reaction–diffusion Physics-Informed Neural Network (PINN) (Raissi et al., 2019; Karniadakis et al., 2021; Krishnapriyan et al., 2021). Optimization progress is measured against cumulative Gradient-Equivalent Computations (GEC), including HVPs used for empirical scale estimation and polynomial actions. Task construction, tuning, and computational-accounting details are given in Appendix D.

On CURVES, Smooth-SFN reduces the training objective faster than Series-HVP within the evaluated GEC budget. This ordering does not transfer to the PINN, where Adam and L-BFGS (Liu & Nocedal, 1989) reach substantially lower objectives and both second-order spectral methods exhibit slower late-stage progress (Fig. 5, left column).

To characterize the realized Smooth-SFN action, we monitor

$$A_t := \frac{\langle \Delta_t, g_t \rangle}{\|\Delta_t\|_2 \|g_t\|_2}, \qquad \Delta_t^{(0)} := \frac{P_{K,\alpha}(0)}{L_t} g_t,$$

together with

$$E_{\Delta_t^{(0)}}(\Delta_t) = \frac{\|[P_{K,\alpha}(H_t/L_t) - P_{K,\alpha}(0)I]g_t\|_2}{|P_{K,\alpha}(0)|\,\|g_t\|_2}.$$

Here $|A_t|$ measures gradient collinearity, whereas the center-action error measures agreement with the specific polynomial-center response. Thus $|A_t| \approx 1$ alone does not imply $\Delta_t \approx \Delta_t^{(0)}$.

On CURVES, the realized action remains nearly gradient-collinear through most of optimization while the center-action error remains appreciable until late in the trajectory. The PINN shows a clearer regime transition: the action begins less gradient-collinear and far from the center response, then moves toward $|A_t| \approx 1$ while the center-action error decreases by several orders of magnitude.

These diagnostics characterize the realized finite-degree action rather than the ideal Smooth-SFN operator. In particular, $P_{K,\alpha}(0)$ need not equal $f_\alpha(0) = 1/\alpha$. Accordingly, the benchmark trajectories should be interpreted as finite-budget realizations under the compute-constrained tuning protocol detailed in Appendix D, rather than as high-fidelity realizations of the ideal spectral operator. The observed regime transition should not be interpreted as a causal explanation of the optimization gap.

Appendix E probes residual dependence on spectral geometry, while Appendix F reports seed-wise realized-action and operational-curvature diagnostics together with PINN endpoint and hard-initial-condition controls.

7 Conclusion

We introduced a scale-consistent framework for matrix-free spectral transformations of indefinite Hessians. The central observation is that an absolute regularization parameter does not have an invariant spectral interpretation: under curvature rescaling, it changes both the relative geometry of the target operator and the computational difficulty required for its matrix-free realization.

Relative resolution provides a dimensionless parameterization that separates normalized spectral geometry from absolute curvature magnitude. Within this formulation, we connected spectral operator sensitivity, finite polynomial approximation complexity, and the practical realization of the normalization scale. The end-to-end results further emphasize

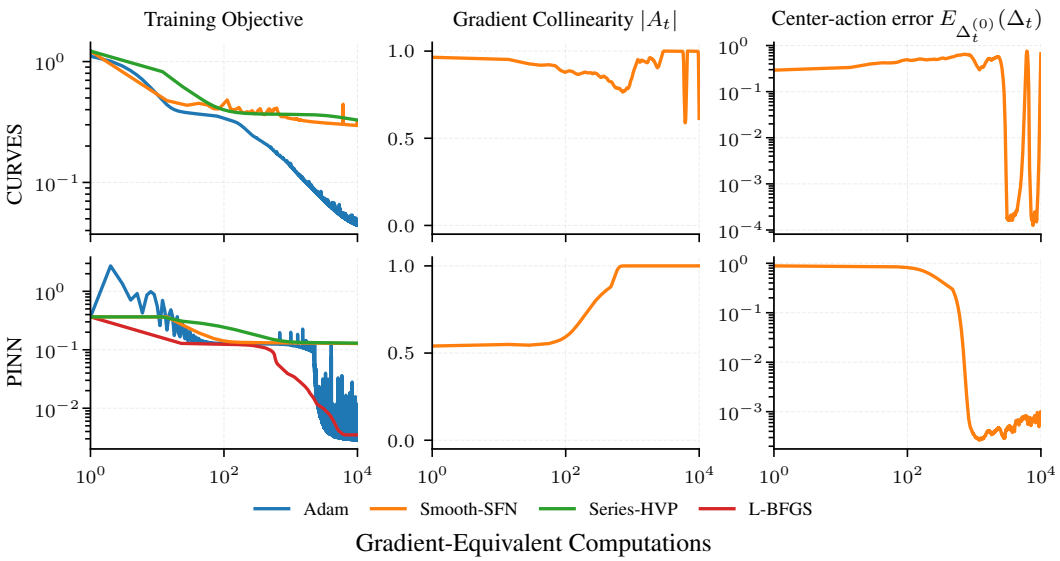


Figure 5: **Optimization behavior and realized update geometry.** Rows correspond to CURVES (top) and the reaction–diffusion PINN (bottom). Left: task-specific training objective versus cumulative GEC for the evaluated optimizers. Middle and right: Smooth-SFN realized-action diagnostics, showing gradient collinearity $|A_t|$ and center-action relative error $E_{\Delta_t^{(0)}}(\Delta_t)$, respectively. The center-action error is shown on a logarithmic scale; lower values indicate closer agreement with the polynomial-center reference $\Delta_t^{(0)}$. The main panels show the seed-67 trajectories.

that the operating point favored under a fixed computational budget need not coincide with a high-fidelity approximation of the target spectral operator, separating operator-realization accuracy from downstream compute-constrained optimization performance.

Controlled mechanism experiments support the predicted approximation-complexity scaling and demonstrate sensitivity of realized action fidelity to operational scale mismatch. In the end-to-end study, the fixed-seed CURVES trajectory improves more rapidly than the controlled Series-HVP realization within the evaluated GEC budget, while this behavior does not transfer to the reaction–diffusion PINN.

The PINN controls narrow, but do not identify, the source of the weaker late-stage behavior; disentangling approximation budget, scale realization, and stiff curvature geometry remains an open empirical question. More generally, relative scaling also exposes a potential optimization-level limitation: when substantial gradient mass lies in eigendirections satisfying the normalized condition $|\lambda|/L^{\mathrm{th}} \ll \alpha$, the exact spectral response approaches its near-flat center regime and the resulting action becomes increasingly gradient-like. Whether this scale separation contributes materially to late-stage optimization slowdown remains to be established.

More broadly, normalization should be treated as part of the approximation problem rather than as an implementation detail. This relative-resolution perspective may extend to other nonlinear matrix functions realized under finite matrix-free budgets.

AI Use Statement

The authors used generative AI tools, including ChatGPT, Gemini, during research development and manuscript preparation. The tools were used to assist with language editing, improving clarity, restructuring sections, and exploring alternative presentations of technical material. They were also used to search for and discuss relevant literature, provide feedback on theoretical arguments and proof presentation, discuss experimental methodology and diagnostic design, review code and pseudocode, and assist in interpreting author-generated experimental results.

Generative AI tools were not used to execute the reported experiments or to produce the numerical results presented in this work. All reported experiments were run using the authors' implementations. Mathematical arguments, implementation details, cited sources, empirical interpretations, and AI-assisted suggestions were reviewed and independently verified by the authors before inclusion. The authors made all final research decisions and take full responsibility for the content of the submission.

References

Marcus Carlsson. Lipschitz continuity for isotropic matrix functions. *Linear Algebra and its Applications*, 624:259–266, 2021. doi: 10.1016/j.laa.2021.04.006.

Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2933–2941, 2014.

Vladimir L. Druskin and Leonid A. Knizhnerman. Two polynomial methods of calculating functions of symmetric matrices. *USSR Computational Mathematics and Mathematical Physics*, 29(6):112–121, 1989. doi: 10.1016/S0041-5553(89)80020-5.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008. doi: 10.1137/1.9780898717778.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. doi: 10.1038/s42254-021-00314-5.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26548–26560, 2021.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989. doi: 10.1007/BF01589116.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *International Conference on Learning Representations*, 2024.

James Martens. Deep learning via Hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 735–742, 2010.

Elre Talea Oldewage, Ross M. Clarke, and José Miguel Hernández-Lobato. Series of Hessian-vector products for tractable saddle-free Newton optimisation of neural networks. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=qBZeQBEDIW`.

Thomas O'Leary-Roseberry, Nick Alger, and Omar Ghattas. Low rank saddle free Newton: A scalable method for stochastic nonconvex optimization. *arXiv preprint arXiv:2002.02881*, 2020. doi: 10.48550/arXiv.2002.02881.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Yousef Saad. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM Journal on Numerical Analysis*, 29(1):209–228, 1992. doi: 10.1137/0729014.

Cooper Simpson. Regularized saddle-free Newton: Saddle avoidance and efficient implementation. Master's thesis, University of Colorado Boulder, 2022. URL `https://stephenbeckr.github.io/assets/docs/CooperSimpson_Summer2022_MAthesis.pdf`.

Lloyd N. Trefethen. *Approximation Theory and Approximation Practice*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2013. doi: 10.1137/1.9781611972825.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. PyHessian: Neural networks through the lens of the Hessian. In *IEEE International Conference on Big Data*, pp. 581–590, 2020. doi: 10.1109/BigData50022.2020.9378171.

Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael W. Mahoney. ADAHESSIAN: An adaptive second order optimizer for machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10665–10673, 2021. doi: 10.1609/aaai.v35i12.17275.

A Matrix-Free Approximation Details

We suppress the iteration index t throughout this appendix.

A.1 Dimensionless Normalization and Analyticity

Let $L^{\text{th}} \geq \|H\|_2$ be an admissible theoretical normalization scale and define $\widetilde{H} := H/L^{\text{th}}$ and $\alpha := \tau/L^{\text{th}}$. Then $\sigma(\widetilde{H}) \subseteq [-1, 1]$, and by Eq. (3), the associated scalar response is $f_\alpha(x) := (x^2 + \alpha^2)^{-1/2}$.

The nearest complex singularities of f_α are the branch points $z_\star = \pm i\alpha$. For every Bernstein ellipse $\mathcal{E}_r$ with $1 < r < \rho(\alpha)$, the function is analytic on and inside $\mathcal{E}_r$, with the limiting ellipse parameter $\rho(\alpha)$ from Eq. (5). Thus, the boundary of $\mathcal{E}_{\rho(\alpha)}$ passes exactly through the nearest branch points.

A.2 Chebyshev Polynomial Construction

We instantiate the approximation using the truncated Chebyshev expansion of degree K. The polynomial is constructed as $P_{K,\alpha}(x) = \sum_{k=0}^{K} c_k(\alpha) T_k(x)$, where $T_k(x)$ is the k-th Chebyshev polynomial of the first kind. The coefficients $c_k(\alpha)$ are determined by the standard Chebyshev projection:

$$c_k(\alpha) = \frac{2 - \delta_{k,0}}{\pi} \int_{-1}^{1} \frac{f_\alpha(x) T_k(x)}{\sqrt{1 - x^2}} dx$$

Since f_α is even, all odd Chebyshev coefficients vanish. We therefore restrict the terminal degree to even values, avoiding a redundant final HVP associated with an odd terminal degree.

These coefficients depend only on the dimensionless resolution α, not on the absolute curvature scale L^{th}. In matrix-free execution, $P_{K,\alpha}(\tilde{H})v$ is evaluated via the standard three-term

Chebyshev recurrence using rotating vector buffers, requiring $\mathcal{O}(d)$ working memory and exactly K Hessian-vector products without forming matrix powers.

A.3 Relative-Error Decoupling and Uniform Finite-Budget Bounds

Using the relative-error functional $E_y(x)$ from Section 4.2, scale equivariance causes the multiplicative factor $(L^{\text{th}})^{-1}$ to cancel from the relative error:

$$E_{M_\tau(H)}\left(\frac{1}{L^{\text{th}}}P_{K,\alpha}(\widetilde{H})\right) = E_{\mathcal{M}_\alpha(\widetilde{H})}\Big(P_{K,\alpha}(\widetilde{H})\Big).$$

Since $\sigma(\widetilde{H}) \subseteq [-1,1]$, spectral calculus gives $\|\mathcal{M}_\alpha(\widetilde{H})\|_2 \geq \min_{x\in[-1,1]} f_\alpha(x) = (1+\alpha^2)^{-1/2}$. We therefore bound the relative operator error by the uniform scalar error on $[-1,1]$:

$$E_{\mathcal{M}_\alpha(\widetilde{H})}\Big(P_{K,\alpha}(\widetilde{H})\Big) \leq \sqrt{1+\alpha^2}\max_{x\in[-1,1]}|f_\alpha(x) - P_{K,\alpha}(x)|.$$

To establish a uniform geometric rate, fix an interior analysis parameter $\theta \in (0,1)$ and set $\rho_\theta(\alpha) := e^{\theta\,\operatorname{arsinh}(\alpha)}$. For every $\alpha \in \mathcal{A} = [\alpha_{\min}, \alpha_{\max}]$, the Bernstein ellipse $\mathcal{E}_{\rho_\theta(\alpha)}$ has imaginary semiaxis $\sinh(\theta\,\operatorname{arsinh}(\alpha)) < \alpha$, remaining strictly separated from the branch points $z_\star = \pm i\alpha$. By the uniform separation above and compactness, $M_{\mathcal{A},\theta} := \sup_{\alpha\in\mathcal{A},\, z\in\mathcal{E}_{\rho_\theta(\alpha)}}|f_\alpha(z)| < \infty$. Applying the standard Chebyshev truncation bound (Trefethen, 2013, Chap. 8) and using $\rho_\theta(\alpha) \geq \rho_\theta(\alpha_{\min}) > 1$ yields

$$\begin{aligned}\max_{x\in[-1,1]}|f_\alpha(x) - P_{K,\alpha}(x)| &\leq \frac{2M_{\mathcal{A},\theta}}{\rho_\theta(\alpha)-1}\rho_\theta(\alpha)^{-K} \\ &\leq \frac{2M_{\mathcal{A},\theta}}{\rho_\theta(\alpha_{\min})-1}e^{-\theta K\,\operatorname{arsinh}(\alpha)}.\end{aligned}$$

Combining the scalar estimate with the operator relation recovers Eq. (6), where $C_{\mathcal{A},\theta} := \sqrt{1+\alpha_{\max}^2}\left(\frac{2M_{\mathcal{A},\theta}}{\rho_\theta(\alpha_{\min})-1}\right)$ is independent of K and α.

Finally, for any nonzero vector g, the action-level quantity used in Sec. 6.1 is controlled by the operator error:

$$E_{\mathcal{M}_\alpha(\widetilde{H})g}\Big(P_{K,\alpha}(\widetilde{H})g\Big) \leq \kappa_2\Big(\mathcal{M}_\alpha(\widetilde{H})\Big)\,E_{\mathcal{M}_\alpha(\widetilde{H})}\Big(P_{K,\alpha}(\widetilde{H})\Big).$$

Since $\kappa_2(\mathcal{M}_\alpha(\widetilde{H})) \leq \sqrt{1+\alpha^2}/\alpha$, the condition-number factor is uniformly bounded on compact α-regimes bounded away from zero. Thus, over compact α-regimes bounded away from zero, the action error inherits the same exponential dependence on $K\,\operatorname{arsinh}(\alpha)$.

B Operator Stability

B.1 Proof of the Fixed-τ Frobenius–Lipschitz Bound

Theorem 1 (Global Frobenius-Lipschitz Bound – Restated). *For any fixed $\tau > 0$, $H \in \mathbb{S}^n$, and symmetric perturbation $\Delta H \in \mathbb{S}^n$, the smooth spectral operator $M_\tau : \mathbb{S}^n \to \mathbb{S}^n_{++}$ satisfies:*

$$\|M_\tau(H+\Delta H) - M_\tau(H)\|_F \leq L_M(\tau)\|\Delta H\|_F$$

where $L_M(\tau) := \frac{2}{3\sqrt{3}\tau^2}$.

Proof. Let $H \in \mathbb{S}^n$ be a real symmetric matrix with spectral decomposition $H = Q\Lambda Q^\top$. For the dimensionful scalar spectral response $f_\tau(x) := (x^2+\tau^2)^{-1/2}$, the spectral operator satisfies $M_\tau(H) = Qf_\tau(\Lambda)Q^\top$.

Following Carlsson (2021), we define the component-wise eigenvalue vector field $\vec{f_\tau} : \mathbb{R}^n \to \mathbb{R}^n$ as $\vec{f_\tau}(\lambda_1,\ldots,\lambda_n) = (f_\tau(\lambda_1),\ldots,f_\tau(\lambda_n))$. This field satisfies the point-symmetry condition $\Pi^{-1}\vec{f_\tau}(\Pi\lambda) = \vec{f_\tau}(\lambda)$ for any permutation matrix Π. By Carlsson (2021, Theorem 3.3),

a point-symmetric Lipschitz eigenvalue field induces a Frobenius-Lipschitz isotropic matrix function. Consequently, the operator Lipschitz constant is bounded by the supremum of the operator 2-norm of the vector field's diagonal Jacobian $J_{\vec{f}_\tau}(z)$:

$$\mathrm{Lip}_F(M_\tau) \le \mathrm{Lip}(\vec{f_\tau}) = \sup_{z\in\mathbb{R}^n} \|J_{\vec{f}_\tau}(z)\|_2 = \sup_{z\in\mathbb{R}} |f'_\tau(z)|$$

Differentiating the scalar mapping yields $f'_\tau(z) = -z(z^2+\tau^2)^{-3/2}$. Maximizing $|f'_\tau(z)|$ identifies critical points at $z = \pm\tau/\sqrt{2}$, establishing the upper bound $\mathrm{Lip}_F(M_\tau) \le \frac{2}{3\sqrt{3}\tau^2} = L_M(\tau)$.

To prove sharpness independently, evaluating scalar matrices $A = aI$ and $B = bI$ yields:

$$\frac{\|M_\tau(A) - M_\tau(B)\|_F}{\|A-B\|_F} = \frac{|f_\tau(a) - f_\tau(b)|}{|a-b|}$$

Taking the limit as $a, b \to \tau/\sqrt{2}$ attains exactly $\frac{2}{3\sqrt{3}\tau^2}$. Thus, the constant in the established bound is exact and sharp. □

B.2 Restricted-Domain Joint Perturbation

To isolate the intrinsic stability consequences of dynamic scaling, we specialize the admissible theoretical normalization to its tight choice, $L^{\mathrm{th}} = \|H\|_2$.

This defines the dynamically scaled operator

$$\Phi_\alpha(H) := M_{\alpha\|H\|_2}(H) \tag{9}$$

on the domain $\mathbb{S}^n \setminus 0$.

While relative resolution preserves scale equivariance, dynamically tying the regularization to the matrix norm introduces an inverse dependence on the absolute curvature scale: $\Phi_\alpha(cH) = |c|^{-1}\Phi_\alpha(H)$ for $c \neq 0$. This exposes two mathematically distinct singularity mechanisms.

First, the smooth regularization removes the eigenvalue-level singularity for every fixed $\tau > 0$. Tying τ dynamically to $\|H\|_2$, however, introduces a distinct inverse-scale singularity as the entire matrix approaches the origin. Because $\|\Phi_\alpha(cH)\|_2 = |c|^{-1}\|\Phi_\alpha(H)\|_2 \to \infty$ as $c \to 0$ for any fixed $H \neq 0$, no global Lipschitz bound can hold for Φ_α on $\mathbb{S}^n \setminus \{0\}$.

Nevertheless, stability is retained on any domain strictly bounded away from the matrix-level origin, $D_{b_0} = \{H \in \mathbb{S}^n : \|H\|_2 \ge b_0 > 0\}$. Any perturbation ΔH changes both the Hessian argument and the induced regularization scale $\tau = \alpha\|H\|_2$. We bound this joint variation via an endpoint decomposition, circumventing the domain's non-convexity:

$$\|\Phi_\alpha(H+\Delta H) - \Phi_\alpha(H)\|_F \le \underbrace{\|M_{\tau'}(H+\Delta H) - M_{\tau'}(H)\|_F}_{\text{Hessian-Argument Perturbation}} + \underbrace{\|M_{\tau'}(H) - M_\tau(H)\|_F}_{\text{Induced-Scale Perturbation}}$$

where $\tau = \alpha\|H\|_2$ and $\tau' = \alpha\|H+\Delta H\|_2$. Applying Theorem 1 to the Hessian-argument term and the Mean Value Theorem to the scalar parameter τ yields the following uniform bound.

Theorem 2 (Restricted Pairwise Joint Perturbation Bound)**.** *For every pair $H, H+\Delta H \in D_{b_0}$, the dynamically scaled operator bounds the simultaneous perturbation of geometry and scale by:*

$$\|\Phi_\alpha(H+\Delta H) - \Phi_\alpha(H)\|_F \le \frac{1}{\tau_{\min}^2}\left(\alpha\sqrt{n} + \frac{2}{3\sqrt{3}}\right)\|\Delta H\|_F \tag{10}$$

where $\tau_{\min} = \alpha\min(\|H\|_2, \|H+\Delta H\|_2) \ge \alpha b_0 > 0$.

Proof. For any fixed $H \neq 0$ and $c > 0$, the exact-norm dynamic operator satisfies $\Phi_\alpha(cH) = c^{-1}\Phi_\alpha(H)$ and $\Phi_\alpha(2cH) = (2c)^{-1}\Phi_\alpha(H)$. Taking $c \to 0$ yields:

$$\frac{\|\Phi_\alpha(cH) - \Phi_\alpha(2cH)\|_F}{\|cH - 2cH\|_F} = \frac{\|\Phi_\alpha(H)\|_F}{2c^2\|H\|_F} \longrightarrow \infty.$$

This divergence confirms that a global Lipschitz constant cannot exist. We therefore establish stability strictly on the restricted non-singular domain $D_{b_0} = \{H \in \mathbb{S}^n : \|H\|_2 \geq b_0 > 0\}$.

For any pair $H, H + \Delta H \in D_{b_0}$, let $\tau_1 = \alpha\|H\|_2$ and $\tau_2 = \alpha\|H + \Delta H\|_2$. We avoid path-integrals over the non-convex domain D_{b_0} by employing an endpoint decomposition via the triangle inequality:

$$\|\Phi_\alpha(H + \Delta H) - \Phi_\alpha(H)\|_F \leq \underbrace{\|M_{\tau_2}(H + \Delta H) - M_{\tau_1}(H + \Delta H)\|_F}_{\text{Induced-scale perturbation}} + \underbrace{\|M_{\tau_1}(H + \Delta H) - M_{\tau_1}(H)\|_F}_{\text{Hessian-argument perturbation}}$$

For the induced-scale perturbation, we isolate the variation of the dimensional regularization scale. By the reverse triangle inequality on the spectral norm, the scale difference is bounded by $|\tau_2 - \tau_1| = \alpha\big|\|H + \Delta H\|_2 - \|H\|_2\big| \leq \alpha\|\Delta H\|_2 \leq \alpha\|\Delta H\|_F$. For the fixed matrix $A := H + \Delta H$, the map $\tau \mapsto M_\tau(A)$ is continuously differentiable for $\tau > 0$. By the fundamental theorem of calculus:

$$M_{\tau_2}(A) - M_{\tau_1}(A) = \int_{\tau_1}^{\tau_2} \frac{\partial}{\partial \tau} M_\tau(A)\, d\tau.$$

Differentiating the matrix function with respect to the scalar parameter τ yields $\partial_\tau M_\tau(A) = -\tau(A^2 + \tau^2 I)^{-3/2}$. Taking the Frobenius norm gives $\|\partial_\tau M_\tau(A)\|_F \leq \frac{\sqrt{n}}{\tau^2}$. Therefore:

$$\|M_{\tau_2}(A) - M_{\tau_1}(A)\|_F \leq |\tau_2 - \tau_1| \sup_{\tau \in [\tau_{\min}, \tau_{\max}]} \left\| \frac{\partial}{\partial \tau} M_\tau(A) \right\|_F \leq \frac{\sqrt{n}}{\tau_{\min}^2} \alpha \|\Delta H\|_F$$

where $\tau_{\min} = \min(\tau_1, \tau_2) \geq \alpha b_0 > 0$.

For the Hessian-argument perturbation, Theorem 1 directly bounds the fixed-τ_1 variation by:

$$\|M_{\tau_1}(H + \Delta H) - M_{\tau_1}(H)\|_F \leq L_M(\tau_1)\|\Delta H\|_F \leq L_M(\tau_{\min})\|\Delta H\|_F = \frac{2}{3\sqrt{3}\tau_{\min}^2}\|\Delta H\|_F$$

Combining both terms yields the stated pairwise bound. Furthermore, because $\tau_{\min} \geq \alpha b_0$, the operator admits the uniform domain consequence:

$$\|\Phi_\alpha(H + \Delta H) - \Phi_\alpha(H)\|_F \leq \frac{\alpha\sqrt{n} + \frac{2}{3\sqrt{3}}}{(\alpha b_0)^2}\|\Delta H\|_F$$

valid for all pairs in D_{b_0}. □

The explicit $\sqrt{n}$ dependence enters through the Frobenius-norm bound on the scalar regularization-scale perturbation across all eigendirections. It therefore reflects a worst-case matrix-level dependence of the present Frobenius analysis and should not be interpreted as implying that the realized action sensitivity necessarily grows as $\sqrt{n}$ in practice.

C Matrix-Free Realization of Smooth-SFN

This appendix specifies the matrix-free realization used in all Smooth-SFN experiments. At each iteration, the method computes the gradient $g = \nabla_\theta z$, estimates a local spectral scale from Hessian–vector products, and applies a degree-K Chebyshev polynomial to the Hessian normalized by that realized scale.

The operating relative resolution $\alpha > 0$ and the even polynomial degree $K \in 2\mathbb{N}_+$ are fixed optimizer hyperparameters. At iteration t, Smooth-SFN estimates an empirical operational scale $\widehat{L}_t$ from HVPs and realizes the dimensional smoothing parameter as $\tau_t = \alpha\widehat{L}_t$. No Hessian matrix or eigendecomposition is formed.

Algorithm 1 Smooth-SFN Update

Input: Parameters θ, objective z
Hyperparameters: relative resolution α, even Chebyshev degree $K \in 2\mathbb{N}^+$, learning rate η, power-iteration steps J, empirical margin $\gamma \geq 1$, scale floor $\varepsilon_L > 0$
Persistent State: Power-iteration vector $\widetilde{v}$
Output: Updated parameters θ

procedure SmoothSFN_Step(θ, z)
  $g \leftarrow \nabla_\theta z$
  for $j = 1, \ldots, J$ **do**
    $w \leftarrow \nabla_\theta \langle \nabla_\theta z, \widetilde{v} \rangle$
    $\widehat{\rho} \leftarrow \|w\|_2$
    if $\gamma \widehat{\rho} \leq \epsilon_L$ **then**
      $\widehat{L} \leftarrow \epsilon_L$
      break
    else
      $\widetilde{v} \leftarrow w / \widehat{\rho}$
      $\widehat{L} \leftarrow \gamma \widehat{\rho}$
    end if
  end for
  $\Delta \leftarrow$ ChebyshevApprox($\theta, z, g, \widehat{L}, \alpha, K$)
  $\theta \leftarrow \theta - \eta \Delta$
  return θ
end procedure

function ChebyshevApprox($\theta, z, g, L, \alpha, K$)
  $v_0 \leftarrow g$
  $v_1 \leftarrow (\frac{1}{L}) \nabla_\theta \langle \nabla_\theta z, v_0 \rangle$ ▷ $v_1 = T_1(H/L)g$
  $p \leftarrow c_0(\alpha) v_0$
  for $k = 2$ **to** K **do**
    $v_k \leftarrow \frac{2}{L} \nabla_\theta \langle \nabla_\theta z, v_{k-1} \rangle - v_{k-2}$ ▷ $v_k = T_k(H/L)g$
    $p \leftarrow p + c_k(\alpha) v_k$
  end for
  return $\frac{1}{L} p$ ▷ Restore dimensional scaling
end function

C.1 Hessian–Vector Product

At iterate θ, let $g := \nabla_\theta z$ and $H := \nabla_\theta^2 z$. For a vector v treated as constant with respect to θ, we evaluate

$$\nabla_\theta \langle \nabla_\theta z, v \rangle = Hv.$$

Vectors produced by the power iteration and Chebyshev recurrence are treated as stop-gradient inputs in subsequent evaluations. Hence the recurrences below repeatedly apply the current Hessian to vectors, without differentiating through the procedure that generated those vectors.

C.2 Empirical Spectral Scale

The theoretical admissible normalization $L^{\text{th}} \geq \|H\|_2$ is not directly available in the matrix-free realization. We therefore construct an empirical operational scale from at most J warm-started power-iteration steps.

At initialization, the persistent vector $\widetilde{v}$ is sampled from a standard Gaussian distribution and normalized to unit norm. Subsequent power-iteration steps update this vector in place, so the final direction from one optimization step is reused to warm-start the next.

For each power-iteration step $j = 1, \ldots, J$, we evaluate and update the persistent direction as

$$w_j = H\widetilde{v}_{j-1}, \qquad \widehat{\rho}_j = \|w_j\|_2, \qquad \widetilde{v}_j = \frac{w_j}{\widehat{\rho}_j}.$$

We apply an empirical safety margin $\gamma \geq 1$ and a positive scale floor $\varepsilon_L > 0$. If $\gamma\widehat{\rho}_j \leq \varepsilon_L$, the power-iteration loop terminates early and the operational scale is set to $\widehat{L} = \varepsilon_L$. Otherwise, the procedure continues up to the prescribed budget J, and the final operational scale becomes $\widehat{L} = \gamma\widehat{\rho}_J$.

Because the persistent direction is normalized, each HVP-norm estimate satisfies $\widehat{\rho}_j = \|H\widetilde{v}_{j-1}\|_2 \leq \|H\|_2$. Finite power iteration seeks to align the persistent direction with a dominant-magnitude eigendirection. We use the HVP norm rather than a signed Rayleigh quotient, yielding a nonnegative magnitude estimate that is insensitive to the sign of the dominant-curvature direction. The resulting scale estimate remains empirical rather than a certified upper bound on $\|H\|_2$.

The floor ϵ_L is used only to keep the empirical normalization well-defined when the estimated spectral scale becomes numerically degenerate. It is not intended as a trust-region or update-magnitude safeguard.

Consequently, the floor by itself does not guarantee bounded update magnitudes in a genuinely floor-active near-flat regime.

C.3 Chebyshev Matrix–Vector Action

Given the realized operational scale $\widehat{L}$, Smooth-SFN evaluates the degree-K Chebyshev polynomial action. The coefficients $c_k(\alpha)$ are determined by the operating relative resolution α and are precomputed as described in Section 4.1.

As detailed in Algorithm 1, the action is computed via the standard Chebyshev recurrence. The final factor $1/\widehat{L}$ restores the dimensional scaling of the normalized spectral action. The uniform approximation guarantee derived for the normalized interval $[-1, 1]$ additionally requires the realized scale to contain the Hessian spectrum, i.e., $\widehat{L} \geq \|H\|_2$.

D Experimental Protocol and Reproducibility

Mechanism-validation details. The neural-Hessian checkpoints evaluated in Section 6.1 are extracted from full-batch Adam trajectories (learning rate 10^{-3}) on a 2000-image subset of the MNIST dataset (downsampled to 14×14). To keep the exact explicit eigendecomposition tractable, the model is a small Multilayer Perceptron with layers $196 \to 64 \to 32 \to 10$ and ReLU activations, yielding 15,018 parameters.

We generate trajectories across 3 independent random seeds, extracting the first 10 optimization steps from each to form a population of 30 distinct checkpoints. At each checkpoint, the exact 15018×15018 dense Hessian is explicitly constructed column-by-column, symmetrized, and decomposed to obtain the oracle scale $L^\star$. All operations use double-precision arithmetic. The finite-degree interventions are evaluated over the parameter grid: relative resolution $\alpha \in \{0.05, 0.1, 0.2, 0.3, 0.5, 0.7\}$ and polynomial degree $K \in \{4, 8, 12, 16, 24, 32, 48\}$. The scale-mismatch interventions are evaluated over a uniform 16-point normalized scale grid $L/L^\star \in [0.5, 2.0]$.

D.1 End-to-End Benchmark Tasks

CURVES Autoencoder. The CURVES objective trains a deep autoencoder to reconstruct 784-dimensional synthetic images of curves. Because the original generator is unavailable, we explicitly reconstruct the dataset geometry: each 28×28 image renders a natural cubic spline passing through three ordered control points with bounded spatial constraints, rendered via Gaussian accumulation. The dataset is generated once using dataset seed 34, yielding a 20,000/10,000 train/test split. The architecture is a multi-layer perceptron with

layers $784-400-200-100-50-25-6-25-50-100-200-400-784$. The 6-dimensional bottleneck layer is linear, while all other hidden layers use sigmoid activations. The objective is the pixel-wise binary cross-entropy loss applied to the output logits and evaluated on full batches.

Reaction–Diffusion PINN. The PINN optimizes the solution to the 1D reaction–diffusion equation $u_t - \nu u_{xx} - \rho u(1-u) = 0$, with $\rho = 5$ and $\nu = 2$, subject to the exact initial condition $u(x,0) = \exp(4(\cos(x-\pi)-1))$. The model is a fully connected network structured as $2 \to 50 \to 50 \to 50 \to 50 \to 1$, using `Tanh` activations and Xavier initialization.

The soft-penalty training objective combines the mean-squared PDE residual ($\mathcal{L}_{\text{PDE}}$), initial-condition error ($\mathcal{L}_{\text{IC}}$), and periodic value- and derivative-matching boundary errors ($\mathcal{L}_{\text{BC,value}}$ and $\mathcal{L}_{\text{BC,deriv}}$). The continuous spatio-temporal domain $(x,t) \in [0, 2\pi] \times [0,1]$ is discretized using 2560 interior PDE collocation points, 256 initial-condition points, and 256 boundary points. All points are sampled uniformly using dataset seed 34 and held fixed throughout optimization.

To provide a high-resolution physical ground truth for Appendix F.2, the exact reference solution is computed using a Fourier spectral method for spatial discretization ($N_x = 512$ modes) and a 5th-order implicit Radau integrator for time stepping, solved under strict relative and absolute tolerances (10^{-10} and 10^{-12}, respectively).

D.2 Optimizer Selection and Computational Accounting

Hyperparameters are selected using search seed 24 under a tuning budget of 2000 GECs. The configuration attaining the lowest training objective is frozen. Benchmark trajectories are then evaluated across three independent initialization seeds (67, 68, and 69) for 10,000 GECs. The hyperparameter search spaces and the final selected configurations are summarized in Table 1 (CURVES) and Table 2 (PINN).

Table 1: **CURVES Autoencoder Hyperparameters.** Parameters with a dash (–) in the search space were held fixed across all configurations based on their algorithm defaults. Series-HVP refers to the controlled series-based realization defined in Section 6.2 and detailed below.

Optimizer	Parameter	Search Space	Selected
Adam	Learning rate	$\{10^{-4}, 10^{-3}, 10^{-2}\}$	10^{-3}
	β_1	$\{0.9, 0.999\}$	0.9
Series-HVP	Damping	$\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$	0.01
	Margin γ	$\{1.0, 1.2, 1.5\}$	1.0
	Series terms K	$\{4, 6, 11\}$	4
	Learning rate η	–	1.0
	Power iterations J	–	5
Smooth-SFN	Resolution α	$\{0.002, 0.005, 0.01, 0.02, 0.05\}$	0.01
	Margin γ	$\{1.0, 1.2, 1.5\}$	1.5
	Chebyshev degree K	$\{4, 8, 16\}$	8
	Learning rate η	–	1.0
	Power iterations J	–	5
	Scale floor ε_L	–	10^{-8}

Selection is based on downstream training-objective performance under the fixed GEC budget rather than on a prescribed operator-approximation tolerance. The selected configuration therefore reflects a compute–fidelity trade-off. In particular, increasing K can improve spectral approximation while reducing the number of optimization steps available within the same budget. Consequently, the selected configuration should be interpreted as the operating point favored by compute-constrained downstream optimization, rather than as evidence of a high-fidelity realization of the target spectral operator.

Table 2: **Reaction–Diffusion PINN Hyperparameters.** Parameters with a dash (–) in the search space were held fixed across all configurations based on their algorithm defaults.

Optimizer	Parameter	Search Space	Selected
Adam	Learning rate	$\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$	3×10^{-3}
	Weight decay	$\{0, 10^{-5}\}$	0
	β_1	–	0.9
L-BFGS	Learning rate	$\{0.9, 1.0, 1.1\}$	0.9
	History size	$\{50, 100\}$	100
	Line search	–	Strong Wolfe
Series-HVP	Damping	$\{0.002, 0.005, 0.01, 0.02, 0.05\}$	0.002
	Margin γ	$\{1.0, 1.2, 1.5\}$	1.0
	Series terms K	$\{4, 6\}$	4
	Learning rate η	–	1.0
	Power iterations J	–	5
Smooth-SFN	Resolution α	$\{0.002, 0.005, 0.01, 0.02, 0.05\}$	0.002
	Margin γ	$\{1.0, 1.2, 1.5\}$	1.5
	Chebyshev degree K	$\{4, 8\}$	8
	Learning rate η	–	1.0
	Power iterations J	–	5
	Scale floor ε_L	–	10^{-8}

Because the tuning horizon is shorter than the final benchmark horizon, the selected operating point may favor early-budget progress and is not guaranteed to be optimal over the full 10,000-GEC evaluation horizon.

For the two matrix-free spectral methods, we use the canonical unit step $\eta = 1$ throughout and do not include the learning rate in the tuning search. This choice treats the realized spectral direction as a Newton-like full-step update rather than introducing an additional first-order step-size degree of freedom. Consequently, the end-to-end comparison should be interpreted as a controlled evaluation of the spectral realizations under this design constraint, rather than as a fully tuned optimizer leaderboard.

Computational budget is accounted through Gradient-Equivalent Computations (GEC). One full-batch gradient evaluation counts as 1 GEC, and each exact Hessian–vector product (HVP) contributes 1 additional GEC. For L-BFGS, every gradient-evaluating closure call invoked by the line search is explicitly counted.

For both matrix-free second-order methods, empirical spectral-scale estimation uses $J = 5$ persistent power-iteration HVPs per optimization step. Smooth-SFN subsequently requires exactly K HVPs for its degree-K Chebyshev action. For the controlled Series-HVP realization, we retain the series construction while using the same persistent HVP-based power-iteration estimator, followed by the scale construction $V = (\gamma\widehat{\rho})^2 + \delta$, where $\widehat{\rho}$ is the resulting estimate of $\|H\|_2$. The K-term truncated action is evaluated as

$$\Delta_{\text{series}} = \frac{1}{\sqrt{V}} \sum_{k=0}^{K-1} c_k \left(I - \frac{H^2 + \delta I}{V} \right)^k g, \qquad c_k = 4^{-k} \binom{2k}{k}.$$

We retain the truncated series construction while replacing the original scale-adaptation mechanism with the same persistent HVP-based spectral estimator used for Smooth-SFN. We additionally omit the momentum, learning-rate adaptation, and sequence-acceleration machinery used in the full optimizer of Oldewage et al. (2024). In terms of computation, the $k = 0$ term is the gradient itself, while each of the remaining $K - 1$ terms requires two HVPs to apply H^2. The Series-HVP action therefore requires exactly $2(K - 1)$ HVPs.

Consequently, excluding early termination in scale estimation, the per-step accounting is

$$\text{GEC}_{\text{Smooth-SFN}} = 1 + J + K, \qquad \text{GEC}_{\text{Series-HVP}} = 1 + J + 2(K - 1).$$

For both methods, the gradient graph used for scale estimation is reused by the subsequent matrix-free action, so no additional full-gradient evaluation is incurred.

The scale floor $\varepsilon_L = 10^{-8}$ is used only as a numerical safeguard for degenerate empirical scale estimates and is not intended as an update-magnitude safeguard. It was not activated in the reported benchmark runs: across seeds 67–69, the minimum realized scales were 6.48×10^{-2} on CURVES and 1.32×10^{2} on the PINN. Accordingly, the reported trajectories do not probe the floor-active regime.

E Residual Spectral-Geometry Dependence

To test whether the finite-degree action error retains substantial dependence on spectral geometry after conditioning on the predicted complexity coordinate, we normalize the realized error by the pooled reference trend. Specifically, let $s = K \operatorname{arsinh}(\alpha)$, and let $\widetilde{E}(s)$ denote the median of $E_{\Delta_\alpha^\star}(\Delta_{K,\alpha})$ pooled across all spectral families evaluated at the coordinate s. We repeat the finite-degree intervention on three controlled synthetic matrix families:

- **Origin-clustered**: 80% of eigenvalues drawn uniformly from $[-0.05, 0.05]$, 20% from $[-1, 1]$.
- **Highly-diffused**: All eigenvalues drawn uniformly from $[-1, 1]$.
- **Bimodal-indefinite**: 50% drawn uniformly from $[-1.0, -0.5]$, 50% from $[0.5, 1.0]$.

All 1000-dimensional matrices are normalized to unit spectral norm, and normalized Gaussian probe vectors determine the action direction. We evaluate the exact reference action and the finite-degree approximation over the same (K, α) parameter grid used in Section 6.1, aggregating 100 independent trials per configuration.

Figure 6 reports the normalized action error $E_{\Delta_\alpha^\star}(\Delta_{K,\alpha})/\widetilde{E}(s)$, where values near one indicate agreement with the pooled complexity trend. Across the numerically resolved regime, the normalized errors remain concentrated near one, with only modest variation across spectral families. Within these controlled spectral families, this indicates that residual spectral-geometry effects are secondary to the dominant organization by $K \operatorname{arsinh}(\alpha)$. Configurations with $\widetilde{E}(s) < 10^{-12}$ are omitted to avoid floating-point dominated ratios.

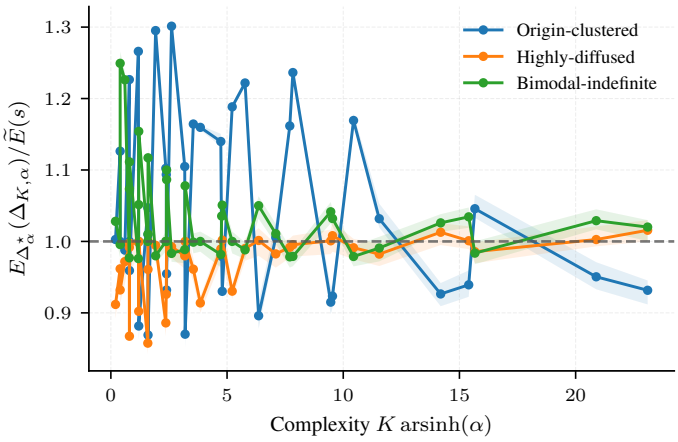


Figure 6: **Residual dependence on spectral geometry.** Normalized action error $E_{\Delta_\alpha^\star}(\Delta_{K,\alpha})/\widetilde{E}(s)$ against the complexity coordinate $s = K \operatorname{arsinh}(\alpha)$. The horizontal baseline at one denotes the pooled median trend at the same complexity coordinate. Solid lines show medians across repeated spectrum and probe realizations; shaded bands denote interquartile ranges. The vertical axis is logarithmic.

F Optimization Robustness and Controls

F.1 Seed-Wise Trajectory Robustness

To assess whether the realized-action behavior identified in Section 6.2 persists across initialization, Fig. 7 reports individual trajectories for seeds 67, 68, and 69. In addition to the objective and the realized-action diagnostics used in the main text, we monitor the gradient-weighted normalized curvature

$$r_t := \frac{\|H_t g_t\|_2}{L_t \|g_t\|_2},$$

where $L_t := \widehat{L}_t$ is the empirical operational scale used by Smooth-SFN.

If $H_t q_{i,t} = \lambda_{i,t} q_{i,t}$ and $w_{i,t} = (q_{i,t}^\top g_t)^2 / \|g_t\|_2^2$, then

$$r_t^2 = \sum_i w_{i,t} \left(\frac{|\lambda_{i,t}|}{L_t} \right)^2,$$

so the operational curvature $r_t := \|H_t g_t\|_2 / (L_t \|g_t\|_2)$ measures the gradient-weighted RMS curvature relative to the operational normalization. The relative resolution α provides a natural reference level for this aggregate quantity. Because L_t is obtained from finite empirical scale estimation, however, r_t is not a spectral-containment certificate; moreover, values above α may be influenced by a small high-curvature component of the gradient.

Across seeds, CURVES remains predominantly gradient-collinear, while both r_t and the center-action error exhibit seed-dependent transitions. Periods in which r_t approaches or falls below the operating relative-resolution level broadly coincide with reduced center-action error, although the correspondence is not pointwise. This distinction is expected because r_t summarizes normalized curvature through an RMS aggregate, whereas $E_{\Delta_t^{(0)}}(\Delta_t)$ depends on the full realized polynomial response.

The PINN exhibits a clearer common pattern across seeds. The operational curvature decreases substantially during optimization, and reductions toward or below the relative-resolution level are accompanied by increasing gradient collinearity and sharply reduced center-action error. These trajectories provide complementary evidence that the late center-action behavior is associated with a change in the operational curvature regime, without establishing that this change uniquely determines the realized action or causes the observed optimization behavior.

F.2 PINN Physical-Space Endpoint Diagnostic

As a complementary diagnostic to the training objective, Table 3 reports the relative L_2 error of the final PINN solution against a high-resolution numerical reference. This provides a direct physical-space assessment of the endpoint reached within the common GEC budget, exposing substantial seed dependence that is not fully summarized by the training objective alone.

Table 3: **Physical-Solution Error.** Relative L_2 error against the numerical reference solution at the conclusion of the 10000-GEC benchmark budget.

Optimizer	**Seed 67**	**Seed 68**	**Seed 69**
Adam	6.83×10^{-1}	6.91×10^{-1}	6.89×10^{-1}
L-BFGS	6.91×10^{-1}	6.84×10^{-1}	2.89×10^{-3}
Series-HVP	9.97×10^{-1}	9.94×10^{-1}	2.80×10^{-1}
Smooth-SFN	9.94×10^{-1}	9.93×10^{-1}	9.91×10^{-1}

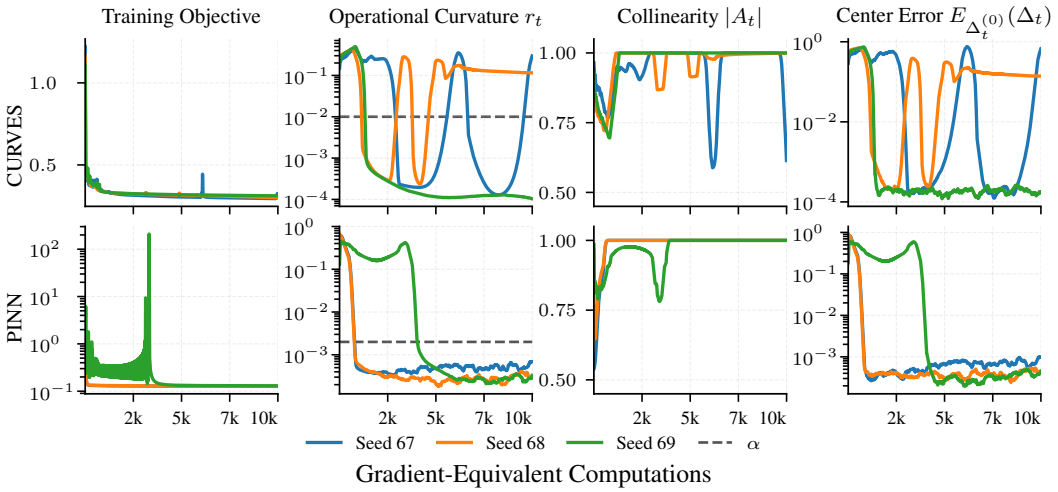


Figure 7: **Seed-wise optimization and realized-action diagnostics.** Rows correspond to CURVES (top) and the reaction–diffusion PINN (bottom). Columns show training-objective progress, operational curvature $r_t = \|H_t g_t\|_2/(L_t\|g_t\|_2)$ (gradient-weighted RMS curvature), gradient collinearity $|A_t|$, and center-action relative error $E_{\Delta_t^{(0)}}(\Delta_t)$. The dashed horizontal line in the operational-curvature panels denotes the operating relative-resolution level α. The quantity r_t is an aggregate diagnostic relative to the empirically realized scale L_t and does not certify spectral containment of the Chebyshev approximation interval.

F.3 Hard-Initial-Condition Control

The main PINN formulation enforces the initial condition through a soft penalty. To test whether competition involving this penalty is sufficient to explain the weaker late-stage Smooth-SFN trajectory, we repeat the benchmark with an exact initial-condition parameterization while leaving the governing PDE, collocation realization, optimizer protocol, and computational accounting unchanged.

Let $u_0(x)$ denote the prescribed initial condition. We parameterize the solution as

$$u_\theta^{\mathrm{hard}}(x,t) = u_0(x) + tN_\theta(x,t),$$

so that $u_\theta^{\mathrm{hard}}(x,0) = u_0(x)$ identically. The corresponding training objective therefore removes the initial-condition penalty and retains only the PDE-residual and boundary-condition terms.

Figure 8 reports optimization progress under this Hard-IC formulation. For this fixed seed-67 control, removing the soft initial-condition penalty does not eliminate the weaker late-stage Smooth-SFN behavior. The result therefore shows that soft initial-condition penalty competition is not sufficient to explain the observed gap in this trajectory; it does not identify the remaining cause.

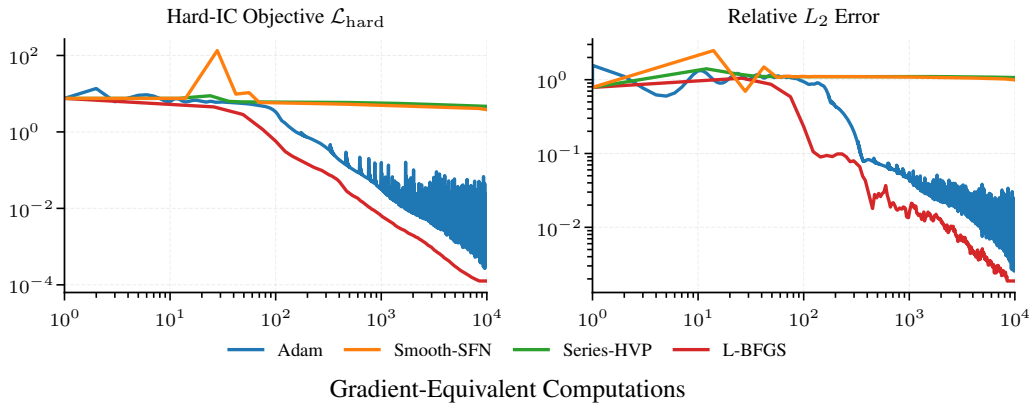


Figure 8: **Hard-IC formulation control for the fixed seed-67 trajectory.** Optimization trajectories under exact initial-condition enforcement, removing the soft IC penalty from the PINN objective.Throughout this appendix, Series-HVP denotes the controlled realization defined in Section 6.2 and Appendix D.