# GAUSS–NEWTON IS BLIND TO EMBEDDED SADDLES: VALUE-TRAINED SURROGATES AS A DERIVATIVE ORACLE FOR AMORTIZED INVERSION

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

## ABSTRACT

Amortized inversion—one nonlinear least-squares solve per target, across many targets sharing one forward model—commonly initialises the full model from the fit of a reduced one. By the embedding principle of nested models, such a fit, when interior and stationary, is an exact critical point of the richer model's landscape; we characterise it in closed form as a *strict saddle* whenever a sign test on the residual holds, and for coalesced Gaussian components generically. The saddle is invisible to the Gauss–Newton matrix $J^\top J$: it carries *exactly zero* curvature in the escape subspace, so Gauss–Newton, Levenberg–Marquardt and secant-corrected methods report convergence at the non-minimiser. The escape lives in the residual-weighted term $C = \sum_i r_i \nabla^2 f_i$ that Gauss–Newton discards; taking it from the true model costs $O(n^2)$ evaluations per finite-difference probe or an interpolation model, restarts multiply the cost, and under evaluation noise a single-sample stencil faces a bias–variance trade-off. We show the term can be *learned*: a neural surrogate of the single-component shape, trained on function values alone with input dimension independent of $n$, supplies the Jacobian and the curvature at zero true-evaluation cost, and the fidelity that matters is low leakage into the well-determined subspace, not value accuracy. On Voigt-profile families with $n = 8$–$32$, Surrogate-Guided LM matches the strongest classical escapes to within a few points (twelve behind at transmission $n$=8) at a cost flat in $n$, 12–$36\times$ below them at $n$=32; an ablation attributes the cost to the free Jacobian and the first-attempt escape to the curvature (a single split perturbation reaches 63–$81\%$, the curvature step from it 79–$96\%$). With $0.1\%$ evaluation noise it meets the accuracy criterion on 73–$88\%$ of targets against at most $52\%$ for single-sample alternatives. On an external stellar-synthesis code at 2 s per call, initialised at the single-star fit of a synthetic binary, it meets the success criterion on all 52 non-trivial targets at 3–$9\times$ fewer syntheses than the tested classical escapes—a saving the free Jacobian supplies, since a perturbed split without the curvature step does as well—the surrogate's cost repaid after about seven solves.

## 1 INTRODUCTION

*Parametric* optimisation—one instance $\min_x \frac{1}{2}\|f(x) - y\|^2$ per target $y$, across very many targets sharing one forward model $f$—is the regime of this paper: one small nonlinear least-squares (NLS) problem per pixel in fluorescence-lifetime imaging (Smith et al., 2019; Zickus et al., 2020) and relaxometry, per spectrum in chromatography, per curve in pharmacokinetics. *Amortized* optimisation (Amos, 2023) learns the solution map $y \mapsto x^*(y)$ once; we amortize a component derivative model instead and keep the per-instance solve, which preserves a genuine residual and an evaluated misfit.

The paper is about how those solves are *initialised*. When the model belongs to a nested family, a natural practice is to fit the reduced member first and seed the richer one from it by *merging*: the added component is initialised on top of an existing one, with equal shape parameters and the amplitude split between them. By the embedding principle of nested models (Fukumizu & Amari, 2000; Zhang et al., 2021), that seed, from an interior stationary reduced fit, is not an arbitrary starting guess but an exact critical point of the wider model's landscape, and a solver is placed on it once



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.

per target (appending a zero-amplitude component instead has no such saddle, its curvature being $\propto a_1 a_2$, Rem. 2; merging is the initialisation we study).

*A running example.* Throughout, $f(x) = a_1\varphi(\theta_1) + a_2\varphi(\theta_2)$ with $x = (a_1, \theta_1, a_2, \theta_2) \in \mathbb{R}^n$, sampled at $m > n$ points, so $f : \mathbb{R}^n \to \mathbb{R}^m$; the component $\varphi(\theta) \in \mathbb{R}^m$ is a shared shape—a decay $e^{-kt}$, a spectral peak, an Extended Phase Graph echo train—with an amplitude $a_i$ and a shape parameter $\theta_i$ (a rate, a $T_2$, a centre or width; $n = 2(1 + p)$ for $\theta \in \mathbb{R}^p$). The model is linear in the amplitudes, nonlinear in the shapes, and the components are interchangeable. The reduced model $A\varphi(\theta)$ is fitted first; its optimum $(A^*, \theta^*)$ is split into $x_0 = (A^*/2, \theta^*, A^*/2, \theta^*)$, which reproduces the single-component signal exactly and sits on the *coalescence manifold* $\theta_1 = \theta_2$.

We establish what kind of critical point $x_0$ is and why the standard solver cannot leave it (Fig. 4). *(i)* For two-component amplitude-linear models it is a strict saddle—not a plateau—whenever a sign test on the data holds, which it does at essentially every target of our benchmark families in the misspecified or noisy regime that warrants the richer fit, with a closed-form split curvature (§3): the negative direction pulls the merged components apart, and for coalesced Gaussians no sign condition is needed outside a codimension-one set (Cor. 1). *(ii)* The Gauss–Newton matrix $J^\top J$—the Fisher information under Gaussian noise, the curvature model of Levenberg–Marquardt (LM) and much of second-order practice—has *exactly zero* curvature in that direction, because the entire escape lives in the term $C = \sum_i r_i \nabla^2 f_i$ that Gauss–Newton discards; started on the exact critical point, LM, TRF, VarPro-LM and NL2SOL stop there on 99–100% of targets (§5). *(iii)* The discarded term can be taken from the true model at a price that grows with $n$ ($2n^2{+}2n$ per finite-difference Hessian, $n{+}1$ to initialise an interpolation model; Prop. 2 bounds a single-sample estimate under noise), or supplied by a neural surrogate of the single-component shape trained on *function values alone*, with input dimension independent of $n$, at no cost at solve time—and the fidelity it needs is not the one it would be trained for (§4.1).

**Contributions.**

1. *Gauss–Newton curvature is structurally blind to embedded saddles* (Prop. 1, Cor. 1). That the merged point is a critical point, and a saddle or a minimum according to the sign of a residual-weighted second-derivative matrix, is the embedding principle (Fukumizu & Amari, 2000, Thm. 3), (Zhang et al., 2021), and the same matrix drives splitting steepest descent (Liu et al., 2019). New here is the NLS reading: the split curvature in closed form, $\kappa = a_1 a_2 A^*(\varphi'' \cdot r^*)/\left\|u_2\right\|^2$, removed for Gaussians with no sign condition; the Gauss–Newton model's deficit along the split being *total*—zero, not small—so that no damping, trust region or tolerance recovers it; and what evaluation noise does to a finite-difference estimate of it (Prop. 2).
2. *A learned derivative oracle*, Surrogate-Guided LM (SG-LM; §4). A values-only surrogate of the single-component shape, whose input dimension is $p$ whatever $n$ is, is assembled into the full model and supplies the Jacobian and the discarded term $C$; the escape direction costs no true evaluations, and the line search that validates it ten. Steps are accepted on the true residual, so with exact evaluations an inaccurate surrogate wastes evaluations but cannot raise the true objective.
3. *Value accuracy alone does not determine derivative usefulness* (Lemma 1, §4.1). Descent is governed by leakage into the stiff subspace, not by approximation error or alignment: a ReLU surrogate matches a smooth one's MSE with none of the usable curvature, and the lowest-MSE variant loses $2.1\times$ on curvature to one with $1.2\times$ worse MSE.
4. *What escaping costs classically, measured* (§5). On the exact critical point, LM, TRF, VarPro-LM and NL2SOL succeed on 0–1% of targets. The methods that do escape—finite-difference probes, trust-region Newton and ARC, DFO-LS (whose secant model is *not* blind and beats SG-LM outright at $n{=}4$), restarts—all cost more as $n$ grows; SG-LM does not, and by $n{=}32$ it is 12–14$\times$ cheaper than DFO-LS with restarts, $30\times$ than jitter multistart and 36–94$\times$ than finite-difference second-order methods, at comparable observed success rates; an ablation shows the scaling is bought by the free Jacobian and the escape by the curvature.
5. *Stochastic forward models* (Prop. 2, §5). The best central-difference estimate of the split-mode curvature has error $\propto \sigma^{1/2}$, and Prop. 2 gives a sufficient sign-recovery condition set by the direction of the misfit and local derivative bounds, not by its size; a value-

trained surrogate averages the noise out. At $0.1\%$ evaluation noise SG-LM meets the $1.05\times$ criterion on 73–$88\%$ of targets, the tested single-sample baselines on at most $52\%$ ($n$=12) and $3\%$ ($n$=24), and DFO-LS with 16-fold averaging on $71\%$ at $47\times$ SG-LM's cost.

6. *An external simulator* (§5, §F.5). On synthetic binary spectra from the stellar-synthesis code Korg (2 s per call), initialised at the single-star fit, the saddle is present, a values-only surrogate reproduces its escape curvature against the code's own finite differences, and SG-LM with restarts meets the criterion on all $52$ non-trivial targets at $154$ syntheses against $513$–$1450$ for the classical escapes and a ten-start multistart (the surrogate Jacobian alone, from a perturbed split, at $53$), the surrogate repaid after about seven solves.

## 2 Related work

The strict-saddle property (Ge et al., 2015; Lee et al., 2016) and second-order saddle evasion (Dauphin et al., 2014; Jin et al., 2017; Paternain et al., 2019) assume curvature that is either met by chance along a trajectory or available on demand; zeroth-order methods escape with function values only, by $O(n)$-evaluation Hessian–vector estimates that find negative curvature (Zhang et al., 2022) or by perturbed two-point gradient estimates with no explicit curvature oracle (Ren et al., 2023); our finite-difference probes are of the former kind (Prop. 2 bounds a fixed stencil under noise). The Gauss–Newton matrix, equivalently the generalised Gauss–Newton (Schraudolph, 2002) and, under Gaussian noise, the Fisher information (Amari, 1998; Martens, 2020; Martens & Grosse, 2015), is positive semidefinite by construction, so dropping $C$ is normally treated as a benign large-residual error. Embedding-principle results (Fukumizu & Amari, 2000; Zhang et al., 2021) establish that such critical points exist, Fukumizu & Amari (2000, Thm. 3) classify them by the sign of a residual-weighted second-derivative matrix, the learning dynamics near them are classical (Amari et al., 2006; Wei et al., 2008), and splitting steepest descent (Liu et al., 2019) grows a network from that same matrix $M$ (Eq. (6)) with exact gradients; none reads $M$ as the term Gauss–Newton discards, or asks which second-order *model* sees the saddle and at what evaluation cost. Variable projection (Golub & Pereyra, 2003) inherits the critical point and the basis rank drop; measured, VarPro-LM stops there and VarPro-TRF escapes by accident on $40$–$53\%$ (§5); NL2SOL's secant correction (Dennis et al., 1981), initialised at zero on a stationary point, has no step to learn from; trust-region Newton and ARC (Cartis et al., 2011) escape with finite-difference Hessians at $2n^2{+}2n$ evaluations; model-based derivative-free solvers (Conn et al., 2009; Cartis et al., 2019; Powell, 2009) are *not* blind but pay $n{+}1$ evaluations per initial model and, at one sample per call, are fragile under noise. Geodesic-accelerated LM (Transtrum et al., 2011) finite-differences the true curvature; derivative-carrying surrogates (Czarnecki et al., 2017; O'Leary-Roseberry et al., 2024; Cao et al., 2025) and surrogate-management frameworks (Booker et al., 1999) do not study surrogate *negative* curvature as the escape oracle; learned warm starts (Sambharya et al., 2024) inherit the saddle if they land on it, and greedy insertion (Denoyelle et al., 2020) or the clinical EPG-NNLS fit (Whittall et al., 1997) avoid it by not splitting. Appendix B has the full discussion.

## 3 Theory

Let $\Phi(x) = \frac{1}{2}\|r(x)\|^2$ with $r = f(x) - y$, $f : \mathbb{R}^n \to \mathbb{R}^m$, $J = \partial f/\partial x$, and exact Hessian $\nabla^2\Phi = J^\top J + C$, $C = \sum_i r_i \nabla^2 f_i$. A point $\bar{x}$ is a *strict saddle* if $\nabla\Phi(\bar{x}) = J^\top r = 0$ and $\nabla^2\Phi(\bar{x})$ has a strictly negative eigenvalue. At the coalescence point the columns of $J$ collapse onto $\mathrm{span}\{\varphi(\theta^*), \varphi'(\theta^*)\}$, so $J$ loses rank and, for a scalar shape parameter, $\ker(J)$ is two-dimensional: one direction (amplitude swap) is an exact symmetry, the other (shape split) is degenerate only to first order, and it is there that the residual-weighted term can curve the objective downward (details in the appendix).

**Proposition 1** (Embedded-coalescence spectrum)**.** *For a two-component model $f = a_1\varphi(\theta_1) + a_2\varphi(\theta_2)$ ($\theta \in \mathbb{R}$, $n = 4$ unknowns $(a_1, \theta_1, a_2, \theta_2)$, $m$ data points) at the single-component fit $\theta_1 = \theta_2 = \theta^*$, $a_1 + a_2 = A^*$, write $g = \varphi(\theta^*)$, $g' = \partial_\theta\varphi(\theta^*)$, $g'' = \partial_\theta^2\varphi(\theta^*) \in \mathbb{R}^m$, and let $r^* = f(x_0^*) - y$ be the residual of the reduced fit $x_0^* = (A^*, \theta^*)$. Then $r^* \perp \{g, g'\}$, by stationarity of the reduced fit in $(A, \theta)$. Define*

$$\gamma := g'' \cdot r^*,$$

*the dot product of the shape function's curvature with the reduced-fit residual. Assume $a_1, a_2 > 0$, and that $g$ and $g'$ are linearly independent. Then*

*(a)* $\ker(J) = \operatorname{span}\{u_1, u_2\}$ *with* $u_1 = (1, 0, -1, 0)$ (amplitude swap) *and* $u_2 = (0, a_2, 0, -a_1)$ (shape split)*;*

*(b) in the orthonormal basis* $P = [u_1/\|u_1\|, u_2/\|u_2\|]$ *of* $\ker(J)$, $P^\top \nabla^2 \Phi P = \operatorname{diag}(0, \kappa)$ *with* $\kappa = \dfrac{a_1 a_2 A^* (g'' \cdot r^*)}{\|u_2\|^2}$*, the curvature of* $\Phi$ *along the split direction;*

*(c)* $u_1$ *is exactly flat, and the point is a strict saddle if and only if* $\gamma < 0$*. When* $a_1 = a_2 = a$*,* $u_2$ *is an eigenvector of* $\nabla^2 \Phi$ *with eigenvalue* $a\gamma = \kappa$*, its smallest; for* $a_1 \neq a_2$ *the escape eigenvector leaves* $\operatorname{span}\{u_1, u_2\}$ *and* $u_2$ *merely certifies negativity, giving the bound* $\lambda_{\min} \leq \kappa < 0$.[1]

**Corollary 1** (Coalesced Gaussians are a strict saddle with no sign condition)**.** *Let* $\varphi(t; \mu, \sigma) = \exp\big(-((t-\mu)/\sigma)^2\big)$*, so* $p = 2$ *and* $\theta = (\mu, \sigma)$*, and let* $M$ *be the* $2 \times 2$ *split matrix* $M_{k\ell} = \partial^2_{\theta_k \theta_\ell} \varphi(\theta^*) \cdot r^*$ *of Rem. 3. Then the coalescence point of the two-component model is a strict saddle for* every *single-fit residual* $r^*$ *with* $M_{\mu\sigma} \neq 0$*. No sign condition on the data is required. On the excluded set* $M_{\mu\sigma} = 0$ *one has* $\operatorname{eig}(M) = \{0, M_{\sigma\sigma}\}$*, and the point is a strict saddle if and only if* $M_{\sigma\sigma} < 0$*.*

**Lemma 1** (Escape with an approximate direction)**.** *Let* $\bar{x}$ *be a strict saddle with eigenpair* $(\lambda_{\min} < 0, v)$ *and largest eigenvalue* $\lambda_{\max}$*. Split the spectrum into a benign part* $\mathcal{S}$ *(eigenvalues* $\leq \lambda_+$*, containing* $v$*) and a stiff complement (eigenvalues* $> \lambda_+$*, the well-determined directions). For a unit surrogate direction* $\hat{v}$*, let* $c_v = |\hat{v}^\top v|$ *(alignment) and* $\beta^2 = \|P_{\text{stiff}} \hat{v}\|^2$ *(stiff leakage). Then*

$$\hat{v}^\top \nabla^2 \Phi\, \hat{v} < 0 \quad \textit{whenever} \quad |\lambda_{\min}|\, c_v^2 > \lambda_+ (1 - \beta^2 - c_v^2) + \lambda_{\max} \beta^2.$$

*In the coalescence case* $\lambda_+ = 0$ *and this reduces to* $|\lambda_{\min}| c_v^2 > \lambda_{\max} \beta^2$*. The condition is sufficient, not necessary.*

The governing quantity is stiff-space leakage relative to alignment: at coalescence the condition reads $\beta^2 < (|\lambda_{\min}|/\lambda_{\max}) c_v^2$, and the ratio is small ($0.26$ against $19$ in the biexponential example; $0.006$ against $3.2$ for EPG), because $\lambda_{\min}$ comes from $C$ and the stiff eigenvalues from $J^\top J$.

**Neural networks are component models.** A one-hidden-layer network $f(x) = \sum_k a_k \sigma(w_k^\top x + b_k) + c$ is the model of this section with shape $\theta_k = (w_k, b_k) \in \mathbb{R}^{d+1}$, and growing it by splitting a unit into two half-amplitude copies (Chen et al., 2016; Liu et al., 2019) is the merged initialisation. Prop. 1 and the matrix test of Rem. 3 apply verbatim: with the unsplit units at a minimum of the reduced network, the split point is a strict saddle iff the splitting matrix $M = a \sum_i r_i \nabla^2_\theta \sigma(w^\top x_i + b)$—Liu et al.'s splitting matrix, the unit's output weight $a$ included—has a negative eigenvalue, and along the split the Gauss–Newton matrix has exactly zero curvature. For a Gaussian likelihood that matrix is the Fisher information $J^\top J/\sigma^2$, so natural gradient and K-FAC (Amari, 1998; Martens & Grosse, 2015) precondition with the blind model; §5 tests this on grown networks with exact derivatives.

**Curvature under evaluation noise.** For a stochastic forward model (Monte Carlo transport, an inner iterative solve) the question is not what curvature costs but whether finite differences resolve it at all.

**Proposition 2** (Finite-difference curvature under evaluation noise)**.** *Let every evaluation return* $\tilde{f}(x) = f(x) + \varepsilon$*,* $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_m)$ *drawn independently at each call; let* $u$ *be a unit direction,* $M_3 \geq \sup \|\partial^3_u f\|$*,* $M_4 \geq \sup \|\partial^4_u f\|$ *along the segment used, and* $c_u := r^{*\top} \partial^2_u f(\bar{x})$ *the curvature of* $C$ *along* $u$ *(for* $u = u_2/\|u_2\|$ *this is exactly* $\kappa$ *of Prop. 1(b)); one sample per stencil point.*

*(a) The central-difference estimate* $\hat{c}_u = r^{*\top}\big(\tilde{f}(\bar{x} + hu) - 2\tilde{f}(\bar{x}) + \tilde{f}(\bar{x} - hu)\big)/h^2$ *has bias at most* $\|r^*\| M_4 h^2/12$ *and standard deviation* $\sqrt{6}\, \|r^*\| \sigma/h^2$*; the bias-plus-one-standard-deviation bound is minimised at* $h_* = (12\sqrt{6}\,\sigma/M_4)^{1/4}$*, where it equals*

[1] The content is not that a first-order method stops at a critical point, but that the Gauss–Newton model $\frac{1}{2} p^\top J^\top J\, p$—second-order in form—has *exactly zero* curvature along $u_2$: no damping, trust region or tolerance recovers the escape that the exact Hessian, or any model carrying $C$, sees at once.

$c_* \|r^*\| (M_4\sigma)^{1/2}$, $c_* = 2(\sqrt{6}/12)^{1/2} \approx 0.90$. *The central-difference Jacobian column likewise has a per-entry error bound* $\propto M_3^{1/3}\sigma^{2/3}$ *at* $h_* \propto (\sigma/M_3)^{1/3}$.

*(b) Hence the sign of* $c_u$*—the strict-saddle test of Prop. 1(c)—is recovered with probability at least* $1-\delta$ *whenever* $\left|\hat{r}^{*\top}\partial_u^2 f\right| > z_{1-\delta}^{1/2}\, c_*(M_4\sigma)^{1/2}$, $\hat{r}^* = r^*/\|r^*\|$, $z_{1-\delta}$ *the standard normal quantile (*$0 < \delta < \frac{1}{2}$, $M_4 > 0$, $r^* \neq 0$ *fixed and exact, the step* $h_\delta = z_{1-\delta}^{1/4} h_*$ *inside the segment on which the bounds hold): a threshold on the noise level set by the* shape *of the misfit, not its size. The condition is sufficient, not necessary, for a fixed stencil with one sample per call (what the finite-difference baselines of §5 use, Table 17 excepted); averaging* $q$ *calls per point lowers the threshold by* $q^{-1/4}$ *at* $q$ *times the cost.*

## 4 METHOD: SURROGATE-GUIDED LM

Surrogate-Guided LM (SG-LM) is an LM solver that takes its derivatives from a learned model of the forward map and its residuals from the true one: the LM steps use the surrogate Jacobian, and when LM stops at a saddle the surrogate supplies the curvature term $C$ that Gauss–Newton discards, from which a negative-curvature eigenstep is line-searched on the true model. The key invariant: step acceptance and residual evaluation always use the true forward model, so with exact evaluations surrogate errors cannot raise the true objective—they can only waste evaluations or stop the solve early; under evaluation noise the acceptance test compares noisy values and carries no such pathwise guarantee.

**Surrogate.** We learn the *single-component shape*, not the full map: a values-only network $\hat{\varphi} : \mathbb{R}^p \to \mathbb{R}^m$ approximates $\theta \mapsto \varphi(\theta)$ over the shape box, and the model's surrogate is assembled from it: $\hat{f}(x) = \sum_k a_k \hat{\varphi}(\theta_k)$ for an additive family, $\exp(-\sum_k a_k \hat{\varphi}(\theta_k))$ for the transmission family of §5. Its Jacobian and correction $\hat{C} = \sum_i r_i \nabla^2 \hat{f}_i$ are therefore block-structured; the amplitude structure is exact (the approximation error sits in the shape's values and derivatives); and the network's input dimension is $p$ whatever $n$ and $m$ are. For a peak family, a *coordinate* network $\hat{\varphi}(t-\mu, \text{shape})$, exact in the translation and evaluated at every grid point, serves every axis length. A monolithic surrogate $\hat{f} : \mathbb{R}^n \to \mathbb{R}^m$ is kept as an ablation, and is what fails at $n$=6. A monolithic ReLU surrogate has $\nabla^2 \hat{f} \equiv 0$ and never escapes (Rem. 9); a smooth (GELU) activation is our design choice for faithful component curvature (§G), not a claimed necessity.

**Solver.** A box-constrained LM on the true residual $r = f(x) - y$ with the surrogate Jacobian: each step solves $(\hat{J}^\top \hat{J} + \mu I)\delta = -\hat{J}^\top r$ at zero true evaluations, clamps to the box, and is accepted on the true objective by the usual gain-ratio test; LM stops at $\left\|\hat{J}^\top r\right\|_\infty < 10^{-7}$ or when no damping gives an accepted step. At a stop, $\hat{H} = \hat{J}^\top \hat{J} + \sum_i r_i \nabla^2 \hat{f}_i$ is formed from the *true* residual and *surrogate* derivatives; if $\lambda_{\min}(\hat{H}) < 0$, the true objective is evaluated at $x \pm \alpha \hat{v}$ along its eigenvector for five fixed $\alpha$ (ten evaluations), the best point is taken, and LM resumes. Two options extend this: a *secant correction* of the surrogate Jacobian from the residual differences the LM steps produce anyway, which reduces the bias an inexact Jacobian leaves in the stationary point; and *restarts*, each an SG-LM solve of $\sim$100 evaluations from a jittered split pair, either triggered by a noise-floor test (the "$\chi^2$" test of the box: final $\|r\|^2$ above twice its expected value at a correct fit; suits additive models, where a failed escape lands far above the floor) or unconditional (suits near-miss failures; §E.1).

**Cost.** True forward-model evaluations are the currency. The surrogate's Jacobian and Hessian cost none; a central-difference Jacobian costs $2n$; a finite-difference Hessian of $\Phi$ costs $2n^2+2n$ with our symmetric stencil (2112 at $n$=32); and the $p \times p$ split matrix $M$ of Rem. 3 costs $2p^2$ for a practitioner who knows which pair is coalesced (the cheapest finite-difference escape, a baseline). Training is accounted for in §F.2; wall-clock is a different currency (limitation (a)).

**Algorithm: Surrogate-Guided LM (SG-LM), one target.** *Input:* target $y$; true model $f$ (evaluation only); shape surrogate $\hat{\varphi}$ assembled into $\hat{f}$; box $[\ell, u]$; noise floor $\Phi_0$.

1. $x \leftarrow$ converged reduced fit of $y$, worst-fit component split into two coincident halves.
2. **repeat** ($\leq$ 3 rounds; 40 with the secant correction): LM on $r = f(x) - y$ with steps from $\hat{J}(x)$ [$+E$], accepted on $\Phi$ *(1 true eval per trial step)*; $\hat{H} \leftarrow \hat{J}^\top \hat{J} + \sum_i r_i \nabla^2 \hat{f}_i(x)$ *(free)*; **if** $\lambda_{\min}(\hat{H}) < 0$: $\hat{v} \leftarrow$ its eigenvector, $x \leftarrow \mathrm{clamp}(x \pm \alpha^* \hat{v})$, $\alpha^*$ the best of a fixed schedule *(10 true evals)*, if it lowers $\Phi$; **else break**.
3. **if** $\|r\|^2 > 2\Phi_0$, $\Phi_0 = m(\sigma_{\text{obs}}^2 + \sigma^2)$ [$\chi^2$ policy] **or** always [jitter policy]: jitter the split pair at the initial point and rerun step 2, at most 12 times.
4. **return** the best $x$.

Figure 1: Surrogate-Guided LM. Only the true-model steps spend forward-model evaluations; the surrogate Hessian and its eigenstep are free (cf. Fig. 6).

### 4.1 What surrogate fidelity has to mean

Any learned model whose *derivatives* are consumed—a surrogate Jacobian, a derivative-informed operator, a network differentiated inside a solver—raises the question of what fidelity means for it; this subsection answers it for the curvature the escape consumes. The natural way to build the surrogate is to minimise forward error on sampled values; for this use, value accuracy alone does not determine what matters. Lemma 1 says what governs success: with escape mode $v$, alignment $c_v = |\hat{v}^\top v|$ and leakage $\beta^2 = \|P_{\text{stiff}} \hat{v}\|^2$ into the stiff, well-determined subspace, the surrogate direction descends when $|\lambda_{\min}| c_v^2 > \lambda_{\max} \beta^2$. The binding quantity is $\beta$, not $c_v$ and not $\left\|\hat{f} - f\right\|$: a poorly aligned surrogate still escapes if it does not spill into the stiff walls (`gelu256`: $c_v = 0.66$, $\beta^2 = 0.005$, descends; `base`: $c_v = 0.44$, $\beta^2 = 0.087$, ascends; Fig. 5). Approximation error and leakage are distinct, so optimising the first can move the second the wrong way, and in our sweep it does (Table 4, §G; monolithic biexponential surrogates, $\hat{f} : \mathbb{R}^4 \to \mathbb{R}^{40}$), read on the correction $C = \sum_i r_i \nabla^2 f_i$ the solver consumes:

a ReLU (monolithic) surrogate matches the smooth baseline's forward MSE within $5\%$ with relative $C$ error exactly $1.00$ (Rem. 9); the best-MSE variant beats `GELU + width` on forward error by $1.20\times$ and loses to it on consumed curvature by $2.10\times$; and $8\times$ more data moves the $C$ error from $0.86$ to $0.87$ where smoothness and width move it fourfold.

A surrogate destined to be differentiated should therefore be selected—and, we expect, trained—against a curvature-aware criterion rather than a value one. The bar is low: SG-LM consumes only the sign and rough subspace of one direction, and a $\sim 0.2$ relative Hessian error still gives a median escape-eigenvector alignment of $0.99$, so curvature-aware validation, or selective Sobolev-style derivative supervision (Czarnecki et al., 2017; O'Leary-Roseberry et al., 2024), is the natural remedy—though the value-only second-difference penalty of Table 4 did not move the $C$ error ($0.84$ vs $0.86$), so it is not a free one. The same distinction makes the surrogate robust to evaluation noise (§5): its derivative error is approximation and leakage, not a difference quotient.

## 5 Experiments

**Protocol (in full in §F.4).** Four families, evaluation-only for every method: **biexponential** ($n$=4, $m$=40; misspecified stretched-exponential targets; 200 targets), **two-Gaussian** ($n$=6, $m$=50; 150), $K$**-Voigt** $\sum_{k=1}^{K} a_k V(t - \mu_k; \sigma_k, \gamma_k)$ ($n$=4$K$, $K$ = 2–8: $K-2$ resolved peaks and one *unresolved pair* about a line width apart, on an axis growing with $K$, Lorentzian tails coupling everything; 100 per $K$), and the non-separable **transmission** family $\exp(-\sum_k a_k V)$ on the same targets; observation noise $0.01$. *Initialisation*: the converged reduced fit (exact, uncounted), with the worst-fit component split into coincident halves—the exact critical point when the reduced optimum is interior to the box (median $\left\|J^\top r\right\|_\infty \sim 10^{-8}$); Table 7 uses those targets, and the at-bound ones are in §F.4. *Scoring*: true evaluations counted uniformly; success = final true $\|r\|^2$ within $1.05\times$ the best value found by any method or an uncounted reference; budgets $4000$ ($n \leq 6$), $1000n$ (Voigt); means are over all $100$ targets per cell, failures included at the evaluations they spent (budgets are nominal, §F.4); rates are finite-sample estimates over $100$ targets and we do not read differences of

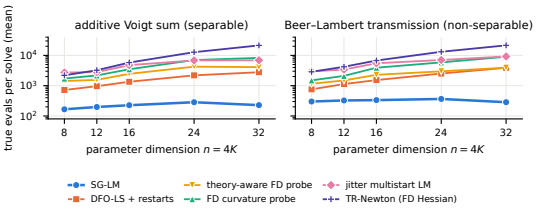


Figure 2: **Classical evaluation counts grow with $n$; SG-LM's stay flat here.** Mean true evaluations per solve against $n = 4K$, additive Voigt family (left) and non-separable transmission family (right), 100 targets per point. SG-LM: $\chi^2$ restarts with the secant correction (left), jitter restarts (right); success rates 83–100% for the classical escapes shown, 97–99% (additive) and 85–100% (transmission) for SG-LM (Tables 11–12).

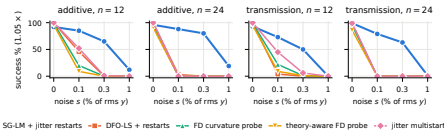


Figure 3: **Stochastic forward model.** Success within $1.05\times$ against the evaluation-noise level $s$ (fresh $\mathcal{N}(0, (s \cdot \mathrm{rms}\, y)^2)$ noise on every evaluation; surrogate trained on equally noisy component values, component-level access for transmission, §F.4; scored on the noiseless misfit), 100 targets per point; finite-difference steps noise-adapted (Prop. 2), DFO-LS in its noisy-objective mode with restarts. Above 0.3% the noise floor of Rem. 5 (6–34%) dominates the criterion.

a few points as established. *Surrogates*: one $256 \times 4$ GELU network of the single-component shape per family, values only; the Voigt one trained once on 4000 single-peak spectra and reused at every $K$ (Jacobian error 2.6%, $C$ 3.5%); the $n \le 6$ surrogates used $4 \times 10^5$–$10^6$ uncounted evaluations (§F.4). *Restart policy*: fixed per family by §4's failure criterion ($\chi^2$ additive, jitter transmission); both are tabulated for both families in §F.4. *Baselines*: those of §2, listed in §F.4.

**Blindness, exactly (Table 7).** The initialisation is a strict saddle on all $10^4$ target/seed/noise combinations tested (Fig. 9, §F.4). Started exactly on it, FD-LM, scipy-TRF, VarPro-LM and the NL2SOL secant method succeed on 0–1% of targets at both $n$, stopping after one Jacobian (zero gradient, a Gauss–Newton model flat along the split, a secant with nothing to learn from). (Escapes from a loosely converged reduced fit are accidental, Table 18.) The methods that escape pay for second-order information—the finite-difference probes, trust-region Newton and ARC, multistart—at 192–3049 median evaluations. DFO-LS, whose interpolation model carries curvature by construction, is not blind (100% at 66 evaluations at $n$=4, beating every method including ours; 83% at $n$=6, 95% with restarts): its model over points spread across 10% of the box carries the second-order term along the split. SG-LM with the component surrogate reaches 85–86% per attempt at 69–98 median evaluations, the same cost as DFO-LS, and 93–95% with the $\chi^2$ restarts.

**The classical cost grows with $n$; the surrogate's does not (Fig. 2, Tables 11–12).** The $K$-Voigt families raise the dimension with one unresolved pair fixed and every component coupled through the Lorentzian tails. The Gauss–Newton-type methods stay blind at every $n$ (Tables 11–12), and every method that escapes pays more as $n$ grows, at a rate set by how it obtains its derivatives: from $n$=8 to 32 the finite-difference probe rises from 1715 to 8273 mean evaluations, trust-region Newton to 21356, and DFO-LS with restarts from 721 to 2781. SG-LM's cost is set by iteration counts, not by $n$: 166–284 mean evaluations at 97–99% across the range with the $\chi^2$ restarts, a ratio to DFO-LS with restarts that grows from $4.3\times$ to $12\times$, to jitter multistart and the $2p^2$ probe from 8–16$\times$ to 18–30$\times$. On the transmission family, with the jitter policy: 285–365 mean evaluations at 85–100% success (99–100% for $n \ge 24$), $14\times$ cheaper than DFO-LS with restarts and 14–75$\times$ than finite differences at $n$=32.

**Stochastic forward model (Fig. 3, Tables 13–14).** With 0.1% evaluation noise the single-attempt finite-difference methods lose the escape (LM, TRF and both probes: optimum on 0–22%, pair

Table 1: **Which surrogate derivative buys what.** LM Jacobian from finite differences ($2n$ each) or the surrogate (free), crossed with the escape curvature (none, the $2p^2$ probe, the surrogate's); one LM core, line search and restart policy; success % / mean true evaluations, 100 targets per cell. Lower block: the no-curvature control against SG-LM. Full grids: §F.4.

| Jacobian | escape | add. $n$=8 | | add. $n$=32 | | trans. $n$=32 | | $n$=24, $s$=0.1% | |
|---|---|---|---|---|---|---|---|---|---|
| | | % | mean | % | mean | % | mean | % | mean |
| *single attempt* | | | | | | | | | |
| FD | $2p^2$ probe | 95 | 1423 | 94 | 4095 | 89 | 3932 | 0 | 217 |
| surrogate | none | 0 | 25 | 0 | 28 | 1 | 26 | 0 | 28 |
| surrogate | $2p^2$ probe | 76 | 112 | 72 | 113 | 77 | 100 | 8 | 58 |
| surrogate | surrogate (SG-LM) | 84 | 105 | 85 | 115 | 85 | 111 | 73 | 95 |
| *with restarts ($\chi^2$ additive, jitter transmission)* | | | | | | | | | |
| surrogate | none | 93 | 149 | 78 | 100 | 97 | 228 | 85 | 100 |
| surrogate | surrogate (SG-LM) | 94 | 195 | 96 | 151 | 100 | 283 | 91 | 125 |
| *single 2% split perturbation, no restarts* | | | | | | | | | |
| surrogate | none | 63 | 64 | 73 | 68 | 79 | 57 | 49 | 45 |
| surrogate | surrogate (SG-LM) | 86 | 97 | 91 | 104 | 96 | 87 | 70 | 70 |

separated on 0–61% additive, 5–85% transmission), and the restart methods keep the escape but lose the refinement: jitter multistart and DFO-LS with restarts still separate the pair on 65–100% of targets but reach the optimum on 52%/46% at $n$=12 and 0%/3% at $n$=24 (additive), consistent with noise-sensitive finite-difference Jacobians for multistart LM (Prop. 2) and degraded interpolation models for DFO-LS. Averaging $q$ noisy calls per evaluation, all counted (additive, $n$=24, $s$=0.1%; Table 17), lifts DFO-LS with restarts to 51% at $q$=4 and 71% at $q$=16, at 12–47× SG-LM's cost; the finite-difference methods stay at 0%. SG-LM with jitter restarts stays at 85–88% (additive) and 73–79% (transmission) at $\sim 250$ evaluations (a surrogate trained on 0.3–3% noisy values has the noiseless one's derivative errors, Rem. 5(iii)); at 0.3%, 0–6% against 50–80%. At 1% the floor (19–34%) dominates the criterion, but SG-LM's median misfit is 1.09–1.34× the optimum with the pair separated on 85–96% of targets, against 3–34× for every classical method.

**Which derivative buys what (Table 1).** SG-LM takes two things from the surrogate, the LM Jacobian and the escape curvature; the table separates them. *(i) A surrogate Jacobian alone is blind*: LM on $\hat{J}^\top \hat{J}$ stops at the point on 99–100% of targets (Prop. 1). *(ii) The cost is the Jacobian's.* Every FD-Jacobian variant with the $2p^2$ or the surrogate escape costs 1.1–4.4 thousand evaluations per attempt at $n$=8–32 (§F.4), 87–97% of them in Jacobians; every variant with the free Jacobian and a cheap escape costs 100–120 per attempt. Fig. 2's scaling is bought by the free Jacobian; the surrogate curvature saves the $2n^2+2n$ of a generic Hessian probe ($445 \to 5146$), not a practitioner's $2p^2$. *(iii) The curvature buys reliability, and under noise the escape.* Per attempt the surrogate curvature escapes on 79–90% of targets against 69–84% for the theory-aware probe at the same cost; with 0.1% evaluation noise the probe's single attempt drops to 8–33% while the surrogate keeps 73%, and every FD-Jacobian row falls to 0–41%—noisy Jacobians impair the refinement even when the curvature is supplied; with $\chi^2$ restarts the probe recovers to 85–88% against 87–91% (Table 22). *(iv) Symmetry breaking does most of the work; the curvature does the rest.* Surrogate-Jacobian LM with the same restart policy and *no* curvature step reaches 93, 78, 97 and 85% in the four columns against 94, 96, 100 and 91% (76–97 vs. 85–100% over every $n$); from a single 2% split perturbation with no restarts it escapes on 63–81% of additive and 68–79% of transmission targets against 86–92 and 79–96% for SG-LM from the same start (Table 1, last block; under 0.1% noise 49–58 vs 70–75%). The free Jacobian plus a perturbation buys most of the reliability; the curvature step buys 11–23 points per attempt for 25–36 more evaluations and, with restarts, the last 1–18, most at $n \geq 16$. *(v) The surrogate need not be neural.* A Gaussian RBF of the same data, tuned for smoothness, matches the network's derivative fidelity and escapes too, 5–11 points behind it per attempt at $n$=8–16 (1–8 with restarts, more at $n$=32; Table 19); a narrower kernel beats the network's value error with a 70% $C$ error—smoothness is the lever, whichever approximator supplies it. *Other ablations* (§F.4): monolithic surrogate 54 vs. component 86% at $n$=6 (but 94 vs 85% at $n$=4); ±7 points over seeds; a wider box changes nothing.

**An external simulator: unresolved spectroscopic binaries (Table 2).** A controlled synthetic problem on an external forward model, motivated by the unresolved-binary degeneracy of survey spectroscopy (El-Badry et al., 2018): the target is $a_1\varphi(\theta_1) + a_2\varphi(\theta_2)$ with $\varphi$ the rectified single-

Table 2: **An external simulator: synthetic unresolved binaries with Korg.** Success within $1.05\times$ of the pooled best, pair separated, median rest-frame syntheses and median measured wall-clock per solve (including the surrogate's autodiff), on the 52 non-trivial targets (§F.5). Budget 1500 assembled-model evaluations; one synthesis $\approx 2.2$ s.

| method | % | split | synth. | wall-clock min |
|---|---|---|---|---|
| FD-LM | 8 | 8 | 30 | 1.5 |
| theory-aware FD probe ($2p^2$) | 79 | 79 | 251 | 9.5 |
| jitter multistart LM | 100 | 100 | 546 | 21.3 |
| light-ratio/velocity multistart (10 starts) | 100 | 100 | 513 | 20.3 |
| DFO-LS + restarts | 100 | 100 | 1450 | 57.6 |
| surrogate Jacobian, no curvature, one 2% perturbation | 100 | 100 | 53 | 2.0 |
| surrogate Jacobian, no curvature, jitter restarts | 100 | 100 | 146 | 5.5 |
| SG-LM, single attempt | 79 | 79 | 54 | 2.2 |
| SG-LM + jitter restarts | 100 | 100 | 154 | 6.1 |

star spectrum from Korg (Wheeler et al., 2023) over 15 700–16 000 Å at $R$=22 500 ($m$=1392), $\theta = (T_{\rm eff}, \log g, [\mathrm{M/H}], v_{\rm rad})$, $n$=10 in box-normalised units, 2.2 s per synthesis (median), 0.5% observation noise. Velocity is an exact Fourier shift, so the component surrogate ($256\times4$ GELU, 2 000 syntheses, values only) is three-dimensional. Sixty targets are near-twins with $|\Delta v| = 3$–12 km/s, forty have $\Delta v = 0$; each starts from its single-star fit (uncounted; costs incremental from it), split. The $\Delta v = 0$ block is trivial at this noise (39 of 40 split points already within $1.05\times$), so the test is the near-twin block. On a separate 100-fit pilot cohort the surrogate's split curvature matches Korg's own finite differences within 2% (§F.5). FD-LM is blind (8%); the $2p^2$ probe reaches 79%; jitter multistart, a ten-start light-ratio/velocity multistart and DFO-LS with restarts reach 100% at 513–1450 syntheses (20–58 min); SG-LM reaches 79% in one attempt at 54 syntheses and 100% with jitter restarts at 154 (6 min), the training repaid after six to seven solves against the multistarts, two against DFO-LS (Fig. 12). The surrogate-Jacobian control without the curvature step succeeds on 3 of 52 from the exact split point against SG-LM's 41, but from a single 2% perturbation of the split it meets the criterion on all 52 at 53 syntheses (2.0 min), as does SG-LM from the same start, and with jitter restarts on all 52 at 146 (SG-LM: 154): here the saving over the classical escapes is the free Jacobian's, and the curvature step matters only from the exact point. Two caveats: Korg is accessed value-only (its native autodiff (Wheeler et al., 2024) is not compared), and the dominant escape is the velocity split, which a 2% perturbation escapes without curvature (§F.5).

**The same saddle in a grown network (Table 5).** The mechanism is not physics: on one-hidden-layer $\tanh$ networks grown by splitting a unit (§3; $n = 10$–25, exact derivatives), the split point is a strict saddle on 100% of 506 instances; LM, natural gradient, K-FAC, GD and Adam stop there, a Hessian eigenstep or Liu et al.'s splitting direction escapes on 97–100%, and Adam from a 1% jitter only after a 275–864-step plateau (Amari et al., 2006).

## 6 Discussion and limitations

The blindness theorem holds for exact and infinitesimal-step Gauss–Newton, not for every linear residual model (DFO-LS is not blind); the advantage is in the cost of derivatives, paid per solve classically and once by the surrogate (Table 1), and in noise robustness.

Beyond inverse problems, the result is a statement about the Fisher curvature model: a split-and-continue growth scheme trained with natural gradient, K-FAC or Gauss–Newton (Müller & Zeinhofer, 2023) starts where those optimisers cannot move (Table 5); splitting steepest descent (Liu et al., 2019) supplies the missing term.

Limitations. (a) Evaluation count is the right currency only for expensive models: the surrogate's autodiff makes SG-LM faster in wall-clock only once an evaluation costs $\gtrsim$ 0.2–5 ms (§F.2); at 2 s per call the measured saving is 3–9$\times$ (Table 2). (b) At a reduced fit on a bound (Tables 9–10) single attempts lose heavily ($86 \to 10$% at $n$=6), restarts recover 96–100%; Lemma 1's guarantee is unconstrained. (c) The component surrogate assumes the shape and composition are known (so do the baselines) and is trained from single-component calls the baselines do not get. (d) Reference optima are pooled bests, the theory is for two components of one shape, and the binary experiment is synthetic and relaxed, its dominant escape direction analytic (§F.5).

### AI USE STATEMENT

In this work we used generative AI tools for the following tasks: generating the synthetic benchmark data sets (the target generators, parameter boxes and the $K$-Voigt and transmission families of Section 5 were designed and implemented with AI assistance); helping to develop the theoretical framework and formulate mathematical claims (the matrix form of the split criterion, Corollary 1 and Proposition 2 were first proposed by the tool during a review of an earlier draft, and the proofs in the appendix were rewritten for clarity and expanded with its assistance); proposing and refining hypotheses (for example that a secant-model solver would not be blind at the critical point, and the diagnosis of the at-bound targets); designing and giving feedback on the experimental methodology; implementing methods (the classical baselines, the surrogate variants and restart policies, the cluster runner and the analysis scripts were written with AI assistance); and interpreting results. Additionally, we used generative AI tools for the tasks of identifying and summarizing related literature, drafting and editing sections of the paper for readability and length, suggesting the paper's structure and title, creating figures, editing code, and formatting references. We also used the OpenReview Paper Assistant Tool, whose automated pre-review feedback we used as a checklist for revision.

We have reviewed all AI-assisted work. All theorem statements and proofs were checked line by line by two authors; the numerical checks quoted after each result were run by the authors independently of the tool. Every number in the tables and figures is produced by a named script from stored result logs, and the prose was checked against those tables in a separate pass. AI-written code was run under a self-test that checks each solver against brute-force or reference solutions, and its results on our earlier benchmark implementation were reproduced before it was used. AI-suggested related work was read by the authors before being cited, and bibliographic entries produced by the tool were verified against the publishers' records. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

Every theoretical claim is stated with its hypotheses in the result that carries it (§3) and proved in full in Appendix C. The solver, its restart policies, the surrogate architectures and training recipes, and the true-evaluation accounting convention are specified in §4 and Appendix E; the benchmark families, their generators, parameter boxes, initialisation, success criterion and budgets are in §5 and Appendix F, together with the full tables behind every figure, the at-bound split, three surrogate training seeds and evaluation-noise levels of $0.1$–$1\%$. The anonymised code, the result logs of every run reported (one row per target and method), the surrogate caches, the stellar syntheses and the scripts that regenerate every table and figure from those logs are provided as supplementary material; the solver and baseline code carries a self-test against brute-force and reference solutions.

## REFERENCES

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Shun-ichi Amari, Hyeyoung Park, and Tomoko Ozeki. Singularities affect dynamics of learning in neuromanifolds. *Neural Computation*, 18(5):1007–1065, 2006.

Brandon Amos. Tutorial on amortized optimization. *Foundations and Trends in Machine Learning*, 16(5):592–732, 2023.

Jordan T. Ash and Ryan P. Adams. On warm-starting neural network training. In *NeurIPS*, 2020.

Andrew J. Booker, J. E. Dennis, Paul D. Frank, David B. Serafini, Virginia Torczon, and Michael W. Trosset. A rigorous framework for optimization of expensive functions by surrogates. *Structural Optimization*, 17(1):1–13, 1999.

Mary Ann Branch, Thomas F. Coleman, and Yuying Li. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing*, 21(1):1–23, 1999.

Lianghao Cao, Thomas O'Leary-Roseberry, and Omar Ghattas. Derivative-informed neural operator acceleration of geometric MCMC for infinite-dimensional Bayesian inverse problems. *Journal of Machine Learning Research*, 26(78):1–68, 2025.

Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

Coralia Cartis, Jan Fiala, Benjamin Marteau, and Lindon Roberts. Improving the flexibility and robustness of model-based derivative-free optimization solvers. *ACM Transactions on Mathematical Software*, 45(3):32:1–32:41, 2019.

Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2Net: Accelerating learning via knowledge transfer. In *ICLR*, 2016.

Andrew R. Conn, Katya Scheinberg, and Luís N. Vicente. *Introduction to Derivative-Free Optimization*. SIAM, 2009.

Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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 (NeurIPS)*, 2014.

John E. Dennis, David M. Gay, and Roy E. Welsch. An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software*, 7(3):348–368, 1981.

Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1): 014001, 2020.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In *NeurIPS*, 2001.

Kareem El-Badry, Yuan-Sen Ting, Hans-Walter Rix, Eliot Quataert, Daniel R. Weisz, Phillip Cargile, Charlie Conroy, David W. Hogg, Maria Bergemann, and Chao Liu. Discovery and characterization of 3000+ main-sequence binaries from APOGEE spectra. *Monthly Notices of the Royal Astronomical Society*, 476(1):528–553, 2018.

Kenji Fukumizu and Shun-ichi Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13(3):317–327, 2000.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pp. 797–842, 2015.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *AISTATS*, 2011.

Gene Golub and Victor Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems*, 19(2):R1–R26, 2003.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning (ICML)*, 2017.

Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, pp. 1246–1257, 2016.

Qiang Liu, Lemeng Wu, and Dilin Wang. Splitting steepest descent for growing neural architectures. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shaihan J. Malik, Rui P. A. G. Teixeira, and Joseph V. Hajnal. Extended phase graph formalism for systems with magnetization transfer and exchange. *Magnetic Resonance in Medicine*, 80(2): 767–779, 2018.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2408–2417, 2015.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *NeurIPS*, 2014.

Johannes Müller and Marius Zeinhofer. Achieving high accuracy with PINNs via energy natural gradient descent. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25471–25485, 2023.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

Thomas O'Leary-Roseberry, Peng Chen, Umberto Villa, and Omar Ghattas. Derivative-informed neural operator: An efficient framework for high-dimensional parametric derivative learning. *Journal of Computational Physics*, 496:112555, 2024.

Santiago Paternain, Aryan Mokhtari, and Alejandro Ribeiro. A newton-based method for nonconvex optimization with fast evasion of saddle points. *SIAM Journal on Optimization*, 29(1):343–368, 2019.

Michael J. D. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical Report DAMTP 2009/NA06, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, 2009.

Thomas Prasloski, Burkhard Mädler, Qing-San Xiang, Alex MacKay, and Craig Jones. Applications of stimulated echo correction to multicomponent $T_2$ analysis. *Magnetic Resonance in Medicine*, 67(6):1803–1814, 2012.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *ICML*, 2019.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks. *Journal of Computational Physics*, 378:686–707, 2019.

Zhaolin Ren, Yujie Tang, and Na Li. Escaping saddle points in zeroth-order optimization: the power of two-point estimators. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28914–28975, 2023.

Rajiv Sambharya, Georgina Hall, Brandon Amos, and Bartolomeo Stellato. Learning to warm-start fixed-point optimization algorithms. *Journal of Machine Learning Research*, 25(166):1–46, 2024.

Nicol N. Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7):1723–1738, 2002.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020.

Jason T. Smith, Ruoyang Yao, Nattawut Sinsuebphon, Alena Rudkouskaya, Nathan Un, Joseph Mazurkiewicz, Margarida Barroso, Pingkun Yan, and Xavier Intes. Fast fit-free analysis of fluorescence lifetime imaging via deep learning. *Proceedings of the National Academy of Sciences*, 116(48):24019–24030, 2019.

Mark K. Transtrum and James P. Sethna. Geodesic acceleration and the small-curvature approximation for nonlinear least squares. *arXiv preprint arXiv:1207.4999*, 2012.

Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Why are nonlinear fits to data so challenging? *Physical Review Letters*, 104(6):060201, 2010.

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

Calvin Tsay. Sobolev trained neural network surrogate models for optimization. *Computers & Chemical Engineering*, 153:107419, 2021.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

Haikun Wei, Jun Zhang, Florent Cousseau, Tomoko Ozeki, and Shun-ichi Amari. Dynamics of learning near singularities in layered networks. *Neural Computation*, 20(3):813–843, 2008.

Matthias Weigel. Extended phase graphs: Dephasing, RF pulses, and echoes — pure and simple. *Journal of Magnetic Resonance Imaging*, 41(2):266–295, 2015.

Adam J. Wheeler, Matthew W. Abruzzo, Andrew R. Casey, and Melissa K. Ness. Korg: A modern 1D LTE spectral synthesis package. *The Astronomical Journal*, 165(2):68, 2023.

Adam J. Wheeler, Andrew R. Casey, and Matthew W. Abruzzo. Korg: Fitting, model atmosphere interpolation, and Brackett lines. *The Astronomical Journal*, 167(2):83, 2024.

Kenneth P. Whittall, Alex L. MacKay, Douglas A. Graeb, Robert A. Nugent, David K. B. Li, and Donald W. Paty. In vivo measurement of $T_2$ distributions and water contents in normal human brain. *Magnetic Resonance in Medicine*, 37(1):34–43, 1997.

Hualin Zhang, Huan Xiong, and Bin Gu. Zeroth-order negative curvature finding: escaping saddle points without gradients. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yaoyu Zhang, Zhongwang Zhang, Tao Luo, and Zhi-Qin John Xu. Embedding principle of loss landscape of deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Vytautas Zickus, Ming-Lo Wu, Kazuhiro Morimoto, et al. Fluorescence lifetime imaging with a megapixel SPAD camera and neural network lifetime estimation. *Scientific Reports*, 10:20986, 2020.

# A SCOPE

**Scope.** The method targets the setting in which the forward model can be *evaluated* but its derivatives are not cheaply available: either a *black-box* model (a legacy simulator, instrument model, or non-differentiable code that admits no automatic differentiation) or an *expensive* model whose derivatives are unavailable or prohibitively costly (each call is a PDE or physics solve without an adjoint or second-order autodiff; expense alone does not preclude autodiff, and a differentiable expensive model needs no surrogate). In both cases the curvature needed to escape the saddle cannot

be obtained by autodiff of the true model; finite differencing it costs $O(n^2)$ true evaluations per Hessian and is noise-fragile, whereas a differentiable surrogate fit to input–output pairs supplies smooth analytic derivatives essentially for free at solve time. Where the true model is both cheap and differentiable, no surrogate is needed—one autodifferentiates it to second order (the Hessian, or Hessian–vector products) and hands that to a solver that can use negative curvature. First-order autodiff alone does not help: it yields an exact $J$, and an exact $J^\top J$ is exactly as blind (Prop. 1). Our benchmark forward models (biexponential, Gaussian sums) are in fact cheap and analytic, chosen so that ground truth is available for scoring; we therefore *restrict all methods to evaluation-only access* (finite-difference derivatives for the classical baselines, surrogate autodifferentiation for ours) to emulate the black-box setting and make the true-evaluation accounting meaningful.

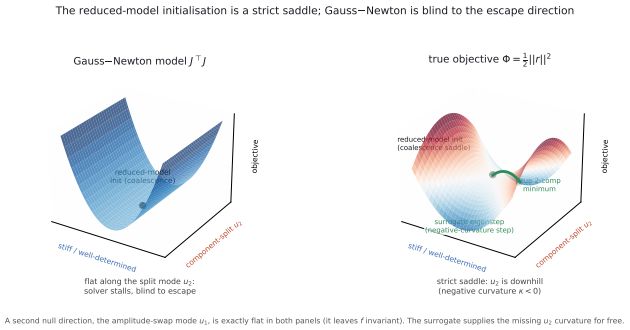


Figure 4: **The mechanism** (schematic). Seeding a two-component fit from its single-component fit lands the solver on the *coalescence* point. *Left:* the Gauss–Newton quadratic model $\frac{1}{2}p^\top(J^\top J)p$ is flat along the direction that would pull the components apart, so LM stalls there, blind to any escape. *Right:* the true objective at the same point is a strict saddle: that direction is downhill. The gap is the term Gauss–Newton drops, $C = \sum_i r_i \nabla^2 f_i$; a smooth surrogate supplies it. Idealised quadratics for intuition; the directions are made precise in §3.

## B Related work in full

**Escaping saddles with negative curvature.** The strict-saddle property (Ge et al., 2015) and the fact that first-order methods almost surely avoid strict saddles asymptotically (Lee et al., 2016) underpin most saddle analysis in learning. Saddle-free Newton (Dauphin et al., 2014) and fast Newton-type saddle evasion (Paternain et al., 2019) use the *true* Hessian's negative eigenvalues; perturbed gradient methods (Jin et al., 2017) escape with gradient information only, and zeroth-order methods escape with function values only, by $O(n)$-evaluation Hessian–vector estimates that find negative curvature (Zhang et al., 2022) or by perturbed two-point gradient estimates with no explicit curvature oracle (Ren et al., 2023)—the finite-difference probes of §5 are of the former kind, and Prop. 2 bounds what a fixed stencil with one sample per call resolves under evaluation noise. Two things separate our setting. The saddle here is not encountered by chance along a trajectory—the standard initialisation lands the solver near it, endemic across the misspecified/noisy regime—so asymptotic-avoidance guarantees do not apply. And the curvature that would escape it is unavailable: the true Hessian is what the black-box or expensive setting denies. None of these substitute a learned forward-model surrogate.

**Gauss–Newton and Fisher curvature models.** The Gauss–Newton matrix $J^\top J$, equivalently the generalised Gauss–Newton (Schraudolph, 2002) and, for Gaussian noise, the Fisher information (Amari, 1998; Martens, 2020), is the curvature model underneath Levenberg–Marquardt and a large body of second-order and natural-gradient practice, including its scalable approximations (Martens & Grosse, 2015). It is positive semidefinite by construction, so it can never certify a descent direction that the residual term $C = \sum_i r_i \nabla^2 f_i$ alone provides. That is usually treated as a benign large-residual approximation error. Proposition 1 shows it is not benign at an embedded critical point: there the deficit is not small but *total*—exactly zero curvature in exactly the escape direction. We prove this for the nested NLS family studied here; whether the same degeneracy occurs at the

embedded critical points of deep networks, where Gauss–Newton/Fisher curvature is most heavily relied on, is a question this analysis raises and does not settle.

**Embedded critical points.** Fukumizu & Amari (2000) and the modern Embedding Principle (Zhang et al., 2021) show that a narrower model's minimum embeds as a saddle in a wider one, and the dynamics of gradient learning near these overlap singularities—plateaus, Milnor-type attractors, and the slow escape—are analysed in detail for layered networks by Amari et al. (2006) and Wei et al. (2008). Theorem 3 of Fukumizu & Amari (2000) already classifies the embedded point as a saddle or a minimum by the sign of a residual-weighted second-derivative matrix, and Liu et al. (2019) use the same matrix to decide how to split a unit. What those results do not do is read the matrix as the term Gauss–Newton discards: they say nothing about which *second-order model* sees the saddle, nor about the noise or the evaluation cost of obtaining the matrix from an evaluation-only model. Our Proposition 1 is this phenomenon in an NLS *forward-model* parameter space, made fully explicit and closed-form, with Cor. 1 supplying a family for which the point is a strict saddle (has a negative eigenvalue, Prop. 1(c)) with no sign condition on the data.

**Second-order / geometric NLS.** Geodesic-accelerated LM (Transtrum et al., 2010, 2011; Transtrum & Sethna, 2012) adds a directional second-derivative correction to LM for sloppy models; its curvature is the true model's (finite-differenced), computed per solve, not learned or amortized. Classical large-residual corrections (Dennis et al., 1981; Nocedal & Wright, 2006)—NL2SOL maintains a secant approximation $S \approx C$ of the discarded term and switches between $J^\top J$ and $J^\top J + S$—assume that the iterates move; a secant update initialised at $S_0 = 0$ on a stationary point receives no step from which to learn curvature, so it cannot generate descent without an external probe. We benchmark an NL2SOL-style method (§5): on the exact critical point it stops with the others (0–1%); started off it, it escapes on 0–13% of targets. Methods that compute and exploit negative curvature explicitly—trust-region Newton with an exact subproblem solver and adaptive cubic regularisation (ARC; Cartis et al., 2011)—do escape; run with finite-difference Hessians, which is all an evaluation-only model allows, they pay $2n^2+2n$ evaluations per Hessian and we include them as baselines.

**Variable projection.** All our additive families are separable in the sense of Golub & Pereyra (2003): linear in the amplitudes, nonlinear in the shapes. Variable projection (VarPro) eliminates the amplitudes by linear least squares and optimises the shapes alone, and is the textbook solver for multi-exponential and spectral fits. It does not obviously remove the trap: the reduced objective $\Phi_{\mathrm{red}}(\theta) = \min_a \Phi(a, \theta)$ is the full objective on the manifold of amplitude-optimal points, which contains the coalescence point (the single-component fit is amplitude-optimal), but at $\theta_1 = \theta_2$ the basis matrix $[\varphi(\theta_1), \varphi(\theta_2)]$ loses rank, so the differentiable-projector treatment, which assumes locally constant rank, does not apply there and we make no general blindness claim for VarPro. Measured (§5), VarPro-LM stops on the point exactly as FD-LM does, while VarPro-TRF escapes by accident on 40–53%. The transmission family of §5 is non-separable, so VarPro does not apply there at all; our surrogate, which learns the single-component shape and composes it through the known nonlinearity, does.

**Model-based derivative-free optimisation.** When derivatives are unavailable the classical answer is a local interpolation model rebuilt from function values (Conn et al., 2009): DFO-LS (Cartis et al., 2019) fits a linear model to each residual from $n+1$ points spread over a trust region and takes Gauss–Newton steps on it, with restarts for noisy problems; BOBYQA (Powell, 2009) fits a quadratic to the scalar objective. These are the natural competitors for expensive or noisy models, and DFO-LS is the strongest classical method in our experiments at small $n$. It is *not* blind: its "Jacobian" is a secant over a trust region of $O(1)$ radius, which carries the second-order term along the split that an exact or small-step Jacobian discards, and its initial sample set is itself a 10% jitter of the pair. What it cannot avoid is the $O(n)$ evaluations of its initial interpolation set (points are reused thereafter) and the loss of reliability as $n$ grows, which its restarts repair at growing cost, and the loss of its secant model's accuracy under evaluation noise at one sample per call, which averaging repairs only at a multiple of the cost (§5, Table 17). Our surrogate is, in these terms, a global model fitted once for all targets rather than a local one refitted per solve.

**Growing networks by splitting.** The closest analysis in the learning literature is splitting steepest descent (Liu et al., 2019), which grows a network by splitting a neuron into two and decides the split direction from a *splitting matrix*—for a squared loss exactly our $M = (\partial_\theta^2 \varphi) \cdot r^*$ (Eq. (6))—whose negative eigenvalue certifies that the split decreases the loss. Their setting is neural-architecture growth with exact gradients; ours is the same critical point met by a least-squares solver that cannot see the splitting matrix, and the contribution is what the solver's curvature model misses and how to supply it.

**Derivative-carrying surrogates.** Sobolev-trained surrogates (Tsay, 2021; Czarnecki et al., 2017) and derivative-informed neural operators (O'Leary-Roseberry et al., 2024; Cao et al., 2025) use surrogate first/second derivatives to accelerate gradient-based optimisation or geometric MCMC. None, to our knowledge, study surrogate *negative curvature* as the saddle-escape oracle. Surrogate-management frameworks (Booker et al., 1999) share SG-LM's invariant of accepting steps on the true function and admit arbitrary surrogate-based searches; their convergence guarantee, however, comes from a pattern-search poll with mesh refinement, which SG-LM's finite-budget procedure does not have, and they do not study a value-trained component oracle for the residual-weighted split curvature.

**Direct-inverse learning, learned warm starts, and greedy insertion.** FLIM/relaxometry deep nets (Smith et al., 2019; Zickus et al., 2020) regress parameters directly from data, discarding the forward model and its residual; learned warm starts (Sambharya et al., 2024; Amos, 2023) keep the solver and learn its initial point instead. Neither addresses the merged initialisation: a learned start that lands near the coalescence point inherits the saddle. Greedy insertion methods—the sliding Frank–Wolfe algorithm for sparse spike recovery (Denoyelle et al., 2020)—add a component at the maximiser of a dual certificate rather than by splitting an existing one, which is the zero-amplitude alternative of Rem. 2: it avoids the saddle by construction, at the price of a global search over the position of every new component. Our approach is complementary: it keeps the true forward model in the loop (genuine residuals, an evaluated misfit) and uses the surrogate only for derivatives. In multi-component $T_2$ relaxometry the clinical standard is a different formulation again: a regularised non-negative least-squares fit of a $T_2$ *spectrum* over a fixed EPG dictionary (Whittall et al., 1997; Prasloski et al., 2012), a convex problem that has no coalescence saddle because the component positions are not free. It trades the saddle for a regularisation choice and a resolution limit set by the dictionary; the parametric few-component fit of §F.1, where the positions are estimated, is the formulation in which the saddle—and our remedy—arise, and we do not compare the two.

# C PROOFS

**Proof of Lemma 2.**

*Proof.* Throughout, $\bar{x}$ is a critical point of $\Phi$, and $J$, $r$ and $C = \sum_i r_i \nabla^2 f_i$ are evaluated there. For every $u \in \mathbb{R}^n$,

$$u^\top \nabla^2 \Phi u = \|Ju\|^2 + u^\top C u, \tag{1}$$

which is just the quadratic form of $\nabla^2 \Phi = J^\top J + C$. It holds at every $x$, not only at critical points; what is special about a critical point is that the sign of this form is what decides the nature of the point.

*(a)* Stationarity reads $J^\top r = 0$, i.e. $r \in \ker(J^\top) = \operatorname{range}(J)^\perp$. If $r = 0$ then $\Phi(\bar{x}) = 0$, the smallest value $\Phi$ can take, so $\bar{x}$ is a global minimiser and not a saddle. Hence a strict saddle has $r \in \ker(J^\top) \setminus \{0\}$.

*(b)* If $\bar{x}$ is a strict saddle, there is a $u$ with $u^\top \nabla^2 \Phi u < 0$ (an eigenvector of the negative eigenvalue will do). By (1), $u^\top C u = u^\top \nabla^2 \Phi u - \|Ju\|^2 < 0$: the negative curvature has to come from $C$, because the Gauss–Newton term contributes $\|Ju\|^2 \geq 0$ in every direction. On $\ker(J)$ the Gauss–Newton term is not merely non-negative but zero, $\nabla^2 \Phi u = J^\top (Ju) + Cu = Cu$, so there $\nabla^2 \Phi$ and $C$ coincide. This is the precise sense in which Gauss–Newton is blind on $\ker(J)$: whatever curvature $\nabla^2 \Phi$ has there, $J^\top J$ reports none of it. We do not claim that the most negative direction of $\nabla^2 \Phi$ lies in $\ker(J)$; Prop. 1(c) shows that in general it does not.

*(c)* Since $\dim\ker(J^\top) = m - \operatorname{rank}(J)$, condition (a) can be met only if $\operatorname{rank}(J) < m$. When $m > n$ this always holds, because $\operatorname{rank}(J) \le n < m$: an overdetermined problem always leaves room for the residual. When $J$ has full row rank at $\bar{x}$, which requires $m \le n$, $\ker(J^\top) = \{0\}$, stationarity forces $r = 0$, and $\bar{x}$ is a zero-residual global minimiser by (a). □

**Proof of Proposition 1.**

*Proof.* Coordinates are ordered $x = (a_1, \theta_1, a_2, \theta_2)$ throughout. The coalescence point is $\bar{x} = (a_1, \theta^*, a_2, \theta^*)$.

*The residual at $\bar{x}$ is $r^*$, and $\bar{x}$ is a critical point.* Since $f(\bar{x}) = (a_1 + a_2)\, g = A^* g = f(x_0^*)$, the two-component model reproduces the single-component fit exactly, so $r(\bar{x}) = r^*$. The single-component fit $x_0^* = (A^*, \theta^*)$ is a critical point of $\frac{1}{2}\left\|A\varphi(\theta) - y\right\|^2$, whose gradient is $(g \cdot r^*,\ A^*\, g' \cdot r^*)$; hence $g \cdot r^* = 0$ and, as $A^* = a_1 + a_2 > 0$, $g' \cdot r^* = 0$. The four columns of $J(\bar{x})$,

$$\partial_{a_1} f = g, \qquad \partial_{\theta_1} f = a_1 g', \qquad \partial_{a_2} f = g, \qquad \partial_{\theta_2} f = a_2 g', \tag{2}$$

are therefore all orthogonal to $r^*$, so $\nabla\Phi(\bar{x}) = J^\top r^* = 0$.

*The correction matrix $C$ in closed form.* Differentiating (2) once more, the only nonzero second derivatives of $f$ at $\bar{x}$ are

$$\partial^2_{a_i\theta_i} f = \partial^2_{\theta_i a_i} f = g', \qquad \partial^2_{\theta_i\theta_i} f = a_i\, g'', \qquad i = 1, 2;$$

every $\partial^2_{a_i a_j} f$ vanishes because $f$ is linear in the amplitudes, and every derivative mixing component 1 with component 2 vanishes because $f$ is a sum of a function of $(a_1, \theta_1)$ and a function of $(a_2, \theta_2)$. By definition $C_{jk} = \sum_i r_i^*\, \partial^2_{jk} f_i = (\partial^2_{jk} f) \cdot r^*$, so

$$C_{a_i\theta_i} = g' \cdot r^* = 0, \qquad C_{\theta_i\theta_i} = a_i\, g'' \cdot r^* = a_i \gamma,$$

and every other entry is zero:

$$C = \operatorname{diag}\big(0,\ a_1\gamma,\ 0,\ a_2\gamma\big). \tag{3}$$

The off-diagonal entries $C_{a_i\theta_i}$ vanish for the same reason $\bar{x}$ is stationary, $r^* \perp g'$. Parts (a)–(c) are read off (2) and (3).

*(a)* By (2), for $u = (u_{a_1}, u_{\theta_1}, u_{a_2}, u_{\theta_2})$,

$$Ju = (u_{a_1} + u_{a_2})\, g + (a_1 u_{\theta_1} + a_2 u_{\theta_2})\, g'.$$

Since $g$ and $g'$ are linearly independent, $Ju = 0$ if and only if $u_{a_1} + u_{a_2} = 0$ and $a_1 u_{\theta_1} + a_2 u_{\theta_2} = 0$. These are two independent linear conditions on $\mathbb{R}^4$, so $\ker(J)$ is two-dimensional; $u_1 = (1, 0, -1, 0)$ and $u_2 = (0, a_2, 0, -a_1)$ satisfy both and are linearly independent, hence $\ker(J) = \operatorname{span}\{u_1, u_2\}$. Equivalently, $\operatorname{range}(J) = \operatorname{span}\{g, g'\}$ has dimension 2.

*(b)* For $u, v \in \ker(J)$ the Gauss–Newton term drops out, $u^\top \nabla^2\Phi v = (Ju)^\top (Jv) + u^\top C v = u^\top C v$, so on $\ker(J)$ the quadratic form of $\nabla^2\Phi$ is that of the diagonal matrix (3). Because $u_1$ has zero entries in both $\theta$ slots,

$$u_1^\top C u_1 = 0, \qquad u_1^\top C u_2 = 0, \qquad u_2^\top C u_2 = a_2^2\,(a_1\gamma) + a_1^2\,(a_2\gamma) = a_1 a_2 (a_1 + a_2)\gamma = a_1 a_2 A^* \gamma.$$

The vectors $u_1, u_2$ are orthogonal, with $\left\|u_1\right\|^2 = 2$ and $\left\|u_2\right\|^2 = a_1^2 + a_2^2$, so in the orthonormal basis $P = [u_1/\left\|u_1\right\|,\ u_2/\left\|u_2\right\|]$ of $\ker(J)$,

$$P^\top \nabla^2\Phi P = \operatorname{diag}(0, \kappa), \qquad \kappa = \frac{u_2^\top \nabla^2\Phi u_2}{\left\|u_2\right\|^2} = \frac{a_1 a_2 A^* \gamma}{a_1^2 + a_2^2}.$$

Thus $\kappa$ is the Rayleigh quotient of $\nabla^2\Phi$ at $u_2$, i.e. the curvature of $\Phi$ along the split direction. Writing it as $\kappa = \big(a_2^2 \cdot a_1\gamma + a_1^2 \cdot a_2\gamma\big)/(a_1^2 + a_2^2)$ shows that it is a weighted mean of the two nonzero eigenvalues $a_1\gamma$, $a_2\gamma$ of $C$; it is an eigenvalue of $\nabla^2\Phi$ itself only in the balanced case, see (c). Note that $\nabla^2\Phi$ does not map $\ker(J)$ into itself when $a_1 \neq a_2$, so $P^\top \nabla^2\Phi P$ is a compression of $\nabla^2\Phi$, not a restriction.

*(c) $u_1$ is exactly flat.* $Ju_1 = 0$ by (a) and $Cu_1 = 0$ by (3), so $\nabla^2\Phi u_1 = J^\top(Ju_1) + Cu_1 = 0$: $u_1$ is an eigenvector of $\nabla^2\Phi$ with eigenvalue exactly 0, whatever the residual.

*Strict saddle if* $\gamma < 0$. Then $\kappa < 0$ by (b), and by the Courant–Fischer characterisation of the smallest eigenvalue,

$$\lambda_{\min}(\nabla^2\Phi) = \min_{u\neq 0} \frac{u^\top \nabla^2\Phi u}{\|u\|^2} \leq \frac{u_2^\top \nabla^2\Phi u_2}{\|u_2\|^2} = \kappa < 0,$$

so $\nabla^2\Phi$ has a negative eigenvalue and the critical point $\bar{x}$ is a strict saddle, with $u_2$ an explicit direction of negative curvature.

*Only if.* The null-space computation cannot give the converse, since it says nothing about directions outside $\ker(J)$; (3) can. If $\gamma \geq 0$ then, as $a_1, a_2 > 0$, every diagonal entry of $C$ is non-negative, so $C \succeq 0$ and $\nabla^2\Phi = J^\top J + C \succeq 0$: no direction of negative curvature exists and $\bar{x}$ is not a strict saddle. Hence $\bar{x}$ is a strict saddle if and only if $\gamma < 0$.

*When is* $u_2$ *the escape eigenvector?* From (3),

$$\nabla^2\Phi u_2 = Cu_2 = \big(0,\ a_1a_2\gamma,\ 0,\ -a_1a_2\gamma\big) = a_1a_2\gamma\,(0,1,0,-1).$$

If $a_1 = a_2 = a$ this equals $a\gamma\,(0,a,0,-a) = a\gamma\,u_2$, so $u_2$ is an eigenvector with eigenvalue $a\gamma$, which is exactly $\kappa$ in that case ($a \cdot a \cdot 2a\,\gamma/(2a^2) = a\gamma$), and it is the smallest eigenvalue: $C = \mathrm{diag}(0, a\gamma, 0, a\gamma) \succeq a\gamma I$ and $J^\top J \succeq 0$, so $\nabla^2\Phi \succeq a\gamma I$. If $a_1 \neq a_2$, no eigenvector of a nonzero eigenvalue lies in $\ker(J)$ at all: suppose $v = \alpha u_1 + \beta u_2$ satisfied $\nabla^2\Phi v = \lambda v$ with $\lambda \neq 0$. Then $\nabla^2\Phi v = Cv = \beta a_1 a_2 \gamma\,(0,1,0,-1)$, while $\lambda v = \lambda(\alpha,\ \beta a_2,\ -\alpha,\ -\beta a_1)$. Comparing first components gives $\lambda\alpha = 0$, so $\alpha = 0$ and $\beta \neq 0$; comparing the second and fourth then gives $a_1\gamma = \lambda = a_2\gamma$, which for $\gamma \neq 0$ forces $a_1 = a_2$. So for $a_1 \neq a_2$ and $\gamma < 0$ the eigenvector of every negative eigenvalue has a component outside $\ker(J)$, and $u_2$ certifies negativity without being an eigenvector: $\lambda_{\min} \leq \kappa < 0$, with strict inequality in general. □

**Proof of Corollary 1.**

*Proof.* Write $u = (t-\mu)/\sigma$, so $\varphi = e^{-u^2}$ with $\partial_\mu u = -1/\sigma$ and $\partial_\sigma u = -u/\sigma$. Differentiating,

$$\partial_\mu\varphi = \frac{2u}{\sigma}\,\varphi, \qquad \partial_\sigma\varphi = \frac{2u^2}{\sigma}\,\varphi, \qquad \partial_\mu^2\varphi = \frac{2}{\sigma}\big(\partial_\mu u\ \varphi + u\,\partial_\mu\varphi\big) = \frac{4u^2-2}{\sigma^2}\,\varphi,$$

and since $\frac{2}{\sigma}\partial_\sigma\varphi - \frac{2}{\sigma^2}\varphi$ is the same expression $\frac{4u^2-2}{\sigma^2}\varphi$, we have the pointwise identity

$$\partial_\mu^2\varphi \ = \ \frac{2}{\sigma}\,\partial_\sigma\varphi \ - \ \frac{2}{\sigma^2}\,\varphi, \tag{4}$$

valid at every sample point $t$ and hence as an identity between vectors in $\mathbb{R}^m$: the second $\mu$-derivative of the shape is a linear combination of the shape and its first $\sigma$-derivative.

The reduced fit $(A^*, \mu^*, \sigma^*)$ is a critical point of $\frac{1}{2}\,\|A\varphi(\cdot;\mu,\sigma) - y\|^2$ in three variables, so its residual $r^*$ is orthogonal to $\varphi$, $A^*\partial_\mu\varphi$ and $A^*\partial_\sigma\varphi$ at $(\mu^*, \sigma^*)$; with $A^* > 0$ this is $r^* \perp \mathrm{span}\{\varphi, \partial_\mu\varphi, \partial_\sigma\varphi\} = \mathrm{range}(J)$, the same stationarity that makes the coalescence point critical (Rem. 3). Pairing (4) with $r^*$ therefore annihilates both terms on the right:

$$M_{\mu\mu} = \partial_\mu^2\varphi \cdot r^* = \frac{2}{\sigma^*}\,(\partial_\sigma\varphi \cdot r^*) - \frac{2}{\sigma^{*2}}\,(\varphi \cdot r^*) = 0,$$

identically in $r^*$: no admissible residual can make the pure $\mu$-split curve upward. The matrix of Eq. (6) is then

$$M = \begin{pmatrix} 0 & M_{\mu\sigma} \\ M_{\mu\sigma} & M_{\sigma\sigma} \end{pmatrix}, \qquad \det M = -M_{\mu\sigma}^2 \leq 0.$$

The product of the two eigenvalues of $M$ is $\det M \leq 0$, so at least one is $\leq 0$; if $M_{\mu\sigma} \neq 0$ the product is strictly negative and $\lambda_{\min}(M) < 0$, whatever the sign of $M_{\sigma\sigma}$. By the matrix criterion of Rem. 3 the coalescence point is then a strict saddle, with escape direction $(0, a_2w, 0, -a_1w)$ for $w$ the negative eigenvector of $M$, a direction that mixes the $\mu$- and $\sigma$-splits. If $M_{\mu\sigma} = 0$ the eigenvalues are 0 and $M_{\sigma\sigma}$, and the same criterion gives a strict saddle exactly when $M_{\sigma\sigma} < 0$. □

**Proof of Lemma 1.**

*Proof.* Let $\nabla^2\Phi = \sum_i \lambda_i q_i q_i^\top$ be an orthonormal eigendecomposition with $q_v = v$ the chosen eigenvector of $\lambda_{\min}$, and expand the unit vector $\hat{v} = \sum_i c_i q_i$, so that $\sum_i c_i^2 = 1$, $c_v^2 = (\hat{v}^\top v)^2$ is the squared alignment, and $\beta^2 = \sum_{i\in\text{stiff}} c_i^2$ is the leakage, since $P_{\text{stiff}}\hat{v} = \sum_{i\in\text{stiff}} c_i q_i$. The remaining weight, carried by the benign directions other than $v$, is

$$\sum_{i\in\mathcal{S},\, i\neq v} c_i^2 = 1 - c_v^2 - \beta^2.$$

In this basis the curvature along $\hat{v}$ splits into three sums,

$$\hat{v}^\top \nabla^2\Phi\, \hat{v} = \sum_i \lambda_i c_i^2 = \lambda_{\min} c_v^2 + \sum_{i\in\mathcal{S},\, i\neq v} \lambda_i c_i^2 + \sum_{i\in\text{stiff}} \lambda_i c_i^2,$$

and each is bounded above using only $c_i^2 \geq 0$ and the definition of the two groups, $\lambda_i \leq \lambda_+$ on $\mathcal{S}$ and $\lambda_i \leq \lambda_{\max}$ everywhere:

$$\sum_{i\in\mathcal{S},\, i\neq v} \lambda_i c_i^2 \leq \lambda_+(1 - c_v^2 - \beta^2), \qquad \sum_{i\in\text{stiff}} \lambda_i c_i^2 \leq \lambda_{\max}\beta^2.$$

The first bound needs no assumption on the sign of $\lambda_+$, since each term is bounded by $\lambda_+ c_i^2$ individually. Therefore

$$\hat{v}^\top \nabla^2\Phi\, \hat{v} \;\leq\; \lambda_{\min} c_v^2 + \lambda_+(1 - c_v^2 - \beta^2) + \lambda_{\max}\beta^2, \tag{5}$$

and $\hat{v}^\top\nabla^2\Phi\hat{v} < 0$ as soon as the right-hand side is negative. Since $\lambda_{\min} = -|\lambda_{\min}|$, that is the stated condition $|\lambda_{\min}|c_v^2 > \lambda_+(1 - c_v^2 - \beta^2) + \lambda_{\max}\beta^2$. When $\lambda_+ = 0$ the middle term disappears and the condition reads $|\lambda_{\min}|c_v^2 > \lambda_{\max}\beta^2$. □

**Proof of Proposition 2.**

*Proof. (a)* Taylor's theorem with remainder along $u$, applied componentwise (or with the integral form of the remainder; the bound is the same), gives $f(\bar{x} \pm hu) = f \pm h\partial_u f + \frac{h^2}{2}\partial_u^2 f \pm \frac{h^3}{6}\partial_u^3 f + \frac{h^4}{24}\partial_u^4 f(\xi_\pm)$, so the noiseless second difference is $\partial_u^2 f + \frac{h^2}{24}\big(\partial_u^4 f(\xi_+) + \partial_u^4 f(\xi_-)\big)$, and contracting with $r^*$ bounds the bias by $\|r^*\|\, \frac{h^2}{24} \cdot 2M_4 = \|r^*\|\, M_4 h^2/12$ (Cauchy–Schwarz). The noise enters as $r^{*\top}(\varepsilon_+ - 2\varepsilon_0 + \varepsilon_-)/h^2$ with three independent draws; each $r^{*\top}\varepsilon$ is $\mathcal{N}(0, \sigma^2\|r^*\|^2)$, so the variance is $(1+4+1)\sigma^2\|r^*\|^2/h^4$. Minimising $Ah^2 + B/h^2$ with $A = \|r^*\|\, M_4/12$, $B = \sqrt{6}\,\|r^*\|\,\sigma$ gives $h_*^4 = B/A$ and the minimum $2\sqrt{AB} = 2\|r^*\|\sqrt{\sqrt{6}M_4\sigma/12}$. For the Jacobian column, the same expansion to third order gives the bias $\leq M_3 h^2/6$; the noise term $(\varepsilon_+ - \varepsilon_-)/2h$ has variance $2\sigma^2/4h^2$; minimising $M_3h^2/6 + \sigma/(\sqrt{2}h)$ gives $h_*^3 = 3\sigma/(\sqrt{2}M_3)$ and a minimum proportional to $M_3^{1/3}\sigma^{2/3}$. *(b)* The estimate is $c_u + b_h + s_h Z$ with $|b_h| \leq B_h := \|r^*\|\, M_4 h^2/12$, $s_h := \sqrt{6}\,\|r^*\|\,\sigma/h^2$ and $Z \sim \mathcal{N}(0,1)$, so $\Pr[\operatorname{sign}\hat{c}_u \neq \operatorname{sign} c_u] \leq \Phi(-(|c_u| - B_h)/s_h) \leq \delta$ whenever $|c_u| > B_h + z_{1-\delta}s_h$. Minimising the right-hand side over $h$ as in (a), with $B$ replaced by $z_{1-\delta}B$, gives $h_\delta^4 = z_{1-\delta}B/A$, i.e. $h_\delta = z_{1-\delta}^{1/4}h_*$, and the minimum $z_{1-\delta}^{1/2}c_*\,\|r^*\|\,(M_4\sigma)^{1/2}$; divide by $\|r^*\|$. The proposition is an idealised calculation: it contracts the noisy stencil with the exact residual, whereas a solver has only a noisy one. With $q$ independent calls averaged at each stencil point, $s_h$ is divided by $\sqrt{q}$ and the threshold by $q^{1/4}$. □

## D Additional theory

**Coalescence is where the Jacobian degenerates.** The columns of $J$ at $x = (a_1, \theta_1, a_2, \theta_2)$ are $\partial_{a_1} f = \varphi(\theta_1)$, $\partial_{\theta_1} f = a_1\varphi'(\theta_1)$, $\partial_{a_2} f = \varphi(\theta_2)$, and $\partial_{\theta_2} f = a_2\varphi'(\theta_2)$. At a point with $\theta_1 \neq \theta_2$ and both amplitudes nonzero these four vectors are linearly independent for the shape families used here (an assumption on the family, not a consequence of distinctness: $\varphi(\theta) = (1, \theta, \theta^2, 0, 0)$ never exceeds rank three), so $J$ has full column rank $n{=}4$ and the inverse problem is locally well-posed. On the *coalescence* locus $\theta_1 = \theta_2 = \theta^*$ they collapse to $\operatorname{span}\{\varphi(\theta^*), \varphi'(\theta^*)\}$: the rank

Table 3: Notation and dimensions, with the running two-component example $f(x) = a_1\varphi(\theta_1) + a_2\varphi(\theta_2)$. The *example* column is the biexponential case used throughout this section, $n$=4, $m$=40; the EPG relaxometry model of §F.1 is $n$=4, $m$=32. Data points outnumber parameters ($m > n$): the overdetermined regime.

| symbol | meaning | shape | example |
|---|---|---|---|
| $x = (a_1, \theta_1, a_2, \theta_2)$ | parameters (amplitudes, shapes) | $\mathbb{R}^n$ | $\mathbb{R}^4$ |
| $y,\ f(x),\ r = f(x) - y$ | data, prediction, residual | $\mathbb{R}^m$ | $\mathbb{R}^{40}$ |
| $\varphi(\theta)$ | single-component shape | $\mathbb{R}^m$ | $\mathbb{R}^{40}$ |
| $J = \partial f / \partial x$ | Jacobian | $\mathbb{R}^{m\times n}$ | $\mathbb{R}^{40\times 4}$ |
| $J^\top J$ | Gauss–Newton Hessian | $\mathbb{R}^{n\times n}$ | $\mathbb{R}^{4\times 4}$ |
| $\nabla^2 f_i$ | Hessian of output $i$ ($m$ of them) | $\mathbb{R}^{n\times n}$ | $\mathbb{R}^{4\times 4}$ |
| $\nabla^2\Phi = J^\top J + C$ | exact Hessian of $\Phi$ | $\mathbb{R}^{n\times n}$ | $\mathbb{R}^{4\times 4}$ |
| $\ker(J)$ at coalescence | Gauss–Newton's blind subspace | $\subset \mathbb{R}^n$ | 2-dim |
| $u_1, u_2$ | swap / split null modes | $\mathbb{R}^n$ | $\mathbb{R}^4$ |

drops to 2 and $\ker(J)$ becomes two-dimensional. *The reduced-model initialisation lands exactly here.* This rank deficiency is the geometric signature of *local non-identifiability*—at the coalescence point, two identical components cannot be separated by the data, so a two-dimensional space of directions leaves $f$ unchanged to first order (the amplitude swap exactly, the split only to first order—its second-order effect is the mechanism studied here). Away from coalescence, if the four Jacobian columns $\varphi(\theta_1), \varphi'(\theta_1), \varphi(\theta_2), \varphi'(\theta_2)$ are linearly independent then $J$ has full column rank and the model is locally identifiable (global identifiability needs shape-specific hypotheses we do not use); at coalescence their span drops to $\{\varphi, \varphi'\}$. This local non-identifiability is precisely what hands Gauss–Newton a curvature-free subspace to stall in. Equivalently, the two-component model manifold *pinches* onto the one-component manifold: a fold at which the image collapses from 4-D to 2-D (Transtrum et al., 2011; Fukumizu & Amari, 2000).

**Two null directions, two kinds of degeneracy.** The null space is spanned by $u_1 = (1, 0, -1, 0)$ (amplitude *swap*) and $u_2 = (0, a_2, 0, -a_1)$ (shape *split*), but they degenerate differently. Moving along $u_1$ transfers amplitude between two *identical* components and leaves $f$ exactly invariant—a genuine relabelling symmetry—so it is flat to all orders ($\nabla^2\Phi$ eigenvalue exactly 0). Moving along $u_2$ is degenerate only to *first* order: it changes $f$ at second order, and that second-order term is precisely the curvature $C = \sum_i r_i \nabla^2 f_i$ that Prop. 1 shows can be negative. The escape therefore lives on $u_2$, never on $u_1$.

**Degeneracy is necessary, not sufficient.** Rank deficiency only guarantees that the blind subspace $\ker(J)$ exists (where $J^\top J$ contributes no curvature; Lemma 2). A strict saddle additionally requires a *nonzero residual* $r^* \in \ker(J^\top)$ (from model misspecification or noise) and the sign condition $\varphi'' \cdot r^* < 0$ (Prop. 1): degeneracy builds the blind spot, the residual curves it downward. Coalescence is moreover the *principal* but not the only degeneracy locus—$J$ also loses rank where an amplitude vanishes ($a_i$=0 zeros the $\theta_i$ column), the configuration in which a component *disappears* rather than merges.

**Lemma 2** (Saddle criterion; standard)**.** *The following are textbook facts, collected because everything below rests on them. A strict saddle of $\Phi$ requires:*

- *(a) A nonzero residual in the model's left null space: $r(\bar{x}) \in \ker(J^\top) \setminus \{0\}$ where $\ker(J^\top) = \mathrm{range}(J)^\perp$.*
- *(b) A direction of negative correction in the Hessian: $u^\top C\, u < 0$ for some $u$, where at a critical point ($\nabla\Phi(\bar{x}) = J^\top r = 0$) the Hessian decomposes as $\nabla^2\Phi u = J^\top(Ju) + Cu$. On $\ker(J)$, the Gauss–Newton term vanishes and $\nabla^2\Phi u = Cu$.*
- *(c) Consequently overdetermined problems ($m > n$)* can *admit such saddles ($\dim\ker(J^\top) = m - \mathrm{rank}(J) \geq 1$), whereas a critical point at which $J$ has full row rank (only if $m \leq n$) is a zero-residual global minimum.*

**Remark 1** (When is the escape easiest to find?)**.** *The second form of the Hessian, $\nabla^2\Phi u = J^\top(Ju) + Cu$, shows why a negative direction is easiest to find on $\ker(J)$: the Gauss–Newton term*

*$J^\top(Ju)$ vanishes there, leaving only $Cu$. At a rank-deficient critical point (as in the coalescence construction of Prop. 1), the* $\ker(J)$ *subspace is large enough (*$\dim\ker(J) = n - \operatorname{rank}(J) \geq 2$*) that, when the point is a strict saddle ($\gamma < 0$ in Prop. 1), an escape direction lies there. This is why overdetermined problems ($m > n$) can hide strict saddles: there is room for both a nonzero residual (in the left null space,* $\dim\ker(J^\top) = m - \operatorname{rank}(J) \geq 1$*) and a negative direction in the null space.*

*Verified:* For the biexponential model $\varphi(\theta) = e^{-kt}$ (shape parameter $\theta = k$, $n = 4$ parameters $(a_1, k_1, a_2, k_2)$, $m = 40$ data points), at the coalescence point where both exponentials have the same rate $k_1 = k_2 = k^*$ and equal amplitudes $a_1 = a_2 = A^*/2$ (the standard two-exponential seeding from the best single-exponential fit): the residual of the one-exponential fit has $\left\|J^\top r\right\|_\infty = 1.3 \times 10^{-6}$ (stationary), $\|r\| = 0.25 \neq 0$ (nonzero residual), and $\operatorname{rank}(J) = 2$ (of $m = 40$), confirming that the coalescence point is indeed rank-deficient in the null space.

*Verified* (biexponential, $\varphi = e^{-kt}$, $g'' = t^2 e^{-k^* t}$): closed-form $\kappa = -0.25540$ vs. the numeric Rayleigh quotient $u_2^\top \nabla^2\Phi u_2 / \|u_2\|^2 = -0.25540$; $\operatorname{eig}(\nabla^2\Phi) = (-0.255,\, 0,\, 1.31,\, 19.18)$; $|\langle v_{\min}, u_2\rangle| = 1.000$ and $|\langle v_{\min}, u_1\rangle| = 0.000$. The sign criterion explains the measured deepening of the saddle with component separation (Fig. 9).

**Remark 2** (Amplitude dependence)**.** *The prefactor $\kappa \propto a_1 a_2$ arises purely from the two-component structure: the split mode $u_2 = (0, a_2, 0, -a_1)$ carries the amplitudes, and each component's shape curvature is scaled by its own amplitude ($\partial^2_{\theta_i} f = a_i \varphi''$), so the directional second derivative contracts to $D(u_2, u_2) = a_2^2(a_1\varphi'') + a_1^2(a_2\varphi'') = a_1 a_2 A^* \varphi'' \in \mathbb{R}^m$, and only after pairing with the residual, $u_2^\top C\, u_2 = a_1 a_2 A^*(\varphi'' \cdot r^*)$, is it a scalar. Consequently the saddle* vanishes *($\kappa \to 0$) as either amplitude $\to 0$—the separate amplitude-degeneracy locus $a_i{=}0$ identified above, where a component disappears rather than merges —and, for fixed total $A^*$, $\kappa$ is* deepest at the balanced split $a_1 = a_2 = A^*/2$*, the split initialisation studied here. That split is also, by Prop. 1(c), the one point where $\kappa = \lambda_{\min}$ exactly, so the depth statement is about the true smallest eigenvalue there and about the bound $\lambda_{\min} \leq \kappa$ elsewhere.*

**Remark 3** (Vector shape / higher dimension)**.** *If $\theta \in \mathbb{R}^p$ (e.g. Gaussians with $\theta = (\mu, \sigma)$, $p = 2$, $n = 6$) the same construction gives one amplitude-swap flat mode plus $p$ shape-split modes $u_2^{(\ell)} = (0, a_2 e_\ell, 0, -a_1 e_\ell)$ where $e_\ell$ is the $\ell$-th canonical basis vector, so* $\ker(J)$ *has dimension $p + 1$. The scalar sign test of Prop. 1(c) is then replaced by a* matrix *test. Define the $p \times p$ symmetric matrix*

$$M_{k\ell} \;:=\; \left(\partial^2_{\theta_k\theta_\ell}\varphi(\theta^*)\right) \cdot r^*, \qquad k, \ell = 1, \ldots, p. \tag{6}$$

*Repeating the computation of Prop. 1 coordinatewise gives $C(u_2^{(k)}, u_2^{(\ell)}) = a_1 a_2 A^* M_{k\ell}$, $C(u_1, \cdot) = 0$, and, in full coordinates, $C = \operatorname{blockdiag}(0, a_1 M, 0, a_2 M)$—the exact analogue of Eq.* (3)*. (As in Prop. 1, the zero blocks use $\partial^2_{a_i\theta_k} f = \partial_{\theta_k}\varphi \perp r^*$ by stationarity of the reduced fit, and $\dim\ker(J) = p + 1$ assumes $\varphi, \partial_{\theta_1}\varphi, \ldots, \partial_{\theta_p}\varphi$ linearly independent.) Hence* the point is a strict saddle if and only if $\lambda_{\min}(M) < 0$*, i.e. iff $M$ is not positive semidefinite: if $w^\top M w < 0$ then $u = (0, a_2 w, 0, -a_1 w) \in \ker(J)$ and $u^\top \nabla^2\Phi u = u^\top C u = a_1 a_2 A^*\, w^\top M w < 0$, so $\lambda_{\min}(\nabla^2\Phi) < 0$; conversely $M \succeq 0$ gives $C \succeq 0$ (as $a_i > 0$) and $\nabla^2\Phi = J^\top J + C \succeq 0$. A negative diagonal entry $\partial^2_{\theta_\ell}\varphi \cdot r^* < 0$ is sufficient but* not *necessary: $M$ can be indefinite with every diagonal entry nonnegative, in which case the escape direction mixes the shape coordinates and no single-coordinate split is downhill. This accounts for the $n{=}6$ geometry in §5. For a composed model $f = \psi(s)$ with $s = \sum_k a_k \varphi(\theta_k)$ (the transmission family, $\psi = \exp(-\cdot)$), a split direction $u$ has $Ds[u] = 0$, so $D^2 f[u, u] = \operatorname{diag}(\psi'(s))\, D^2 s[u, u]$ and the split-direction sufficient condition holds with the effective residual $\operatorname{diag}(\psi'(s))\, r^*$ in place of $r^*$ (the converse needs the reduced local minimum); the additive algebra does not transfer unchanged.*

*Verified:* identity (4) holds to relative error $1.6 \times 10^{-8}$; over 2000 sampled admissible residuals, $\max|M_{\mu\mu}| = 3.6 \times 10^{-9}$. Corollary 1 is what makes the "100% of two-Gaussian targets" of Fig. 9 a structural fact rather than a lucky sample: the coalescence saddle there cannot be avoided by any choice of misspecification or noise that keeps $M_{\mu\sigma} \neq 0$. That exceptional set has codimension one in the residual but is not exotic. If the target and the sampling grid are both symmetric about $\mu^*$ (two equal peaks placed symmetrically, say), then $r^*$ is even in $t - \mu^*$ while $\partial^2_{\mu\sigma}\varphi$ is odd, so $M_{\mu\sigma} = 0$ by parity; with $M_{\mu\mu} = 0$ as well, the split along $\mu$ is then flat to second order and strictness rests on the sign of $M_{\sigma\sigma}$. Noise breaks the symmetry almost surely, but leaves $|M_{\mu\sigma}|$, and hence the depth of the saddle, small.

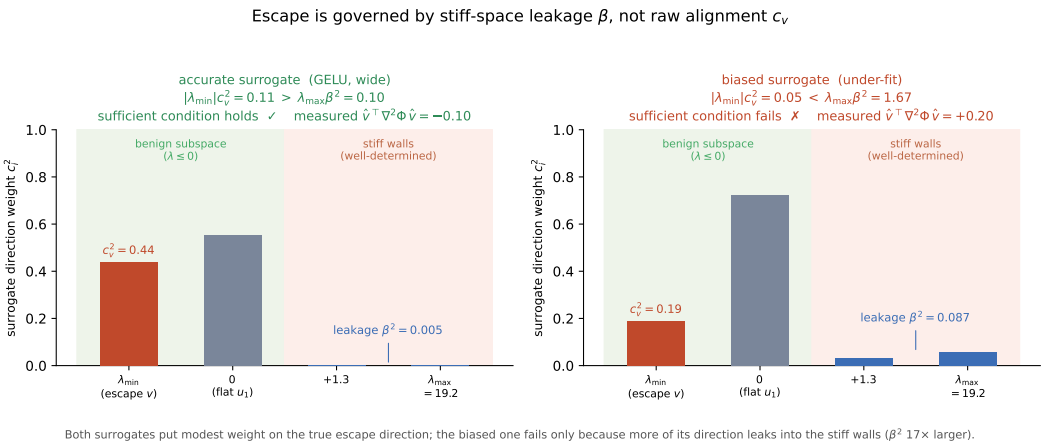


Figure 5: **Why fidelity means low leakage, not high alignment** (Lemma 1, verified spectrum). Both surrogate directions place only modest weight $c_v^2$ on the true escape mode $v$. The accurate surrogate *(left)* still yields descent because almost none of its direction leaks into the stiff, well-determined walls ($\beta^2 = 0.005$); the biased surrogate *(right)* fails only because its leakage is 17× larger—not because it is less aligned.

**Remark 4** (How to control leakage)**.** *In our sweep (monolithic biexponential surrogates, one training recipe) leakage is lowest with smooth activation functions (e.g., GELU vs. ReLU). The mechanism is that smooth networks respond gradually in high-curvature directions rather than saturating; a piecewise-linear monolithic network has no curvature to offer at all. This is quantified in §4.1 and Table 4: smoothness and width each roughly halve the curvature error (C err* $0.86 \to 0.40$ *for GELU,* $0.86 \to 0.44$ *at* $4\times$ *width) and together reduce it fourfold (*$0.21$ *for GELU + width), whereas* $8\times$ *more training data at the same training effort leaves it unchanged (*$0.86 \to 0.87$*)—a finding under this recipe, not a theorem about data.*

*Numerical check of Lemma 1 (biexponential, two trained surrogates):* `gelu256` ($c_v = 0.66$, $\beta^2 = 0.005$) satisfies the bound and gives $\hat{v}^\top \nabla^2\Phi \hat{v} = -0.10$ (descent); `base` ($c_v = 0.44$, $\beta^2 = 0.087$) violates it ($+0.20$). This is seen in Fig. 5: both surrogates have modest alignment on $v$, but only the one with low leakage achieves descent.

**Remark 5** (What the noise does to each route to curvature)**.** *Three consequences.* (i) *A finite-difference Hessian assembled entrywise with the stencil of §4 is* $\nabla^2\Phi + E$ *with* $\|E\|$ *growing with* $n$ *at the rate of (a); by Lemma 1 and the Davis–Kahan bound, the eigenvector it returns can leak into the stiff subspace by up to* $\beta \lesssim \|E\| / (\lambda_{stiff} - \lambda_{\min})$*, so the escape can be lost well before the sign test of (b) fails—an upper bound and a mechanism, not a proof of failure; §5 measures it. Enlarging* $h$ *trades noise for bias, and at a fixed stencil the optimal trade still scales as* $\sigma^{1/2}$*; only repeated calls (*$q^{-1/4}$*) or a smoother estimator improve on it, at a cost the tables count.* (ii) *A secant model built from noisy evaluations over a trust region of radius* $\Delta$ *has the same* $\sigma/\Delta$ *amplification in its Jacobian at one sample per call; averaging and restarts recover some of it at their evaluation cost (Table 17).* (iii) *A surrogate fitted by least squares to* $N$ *noisy values has derivatives that depend on the noise only through the fitted weights, with no* $1/h$ *or* $1/\Delta$ *factor; its Jacobian and curvature errors are set by approximation and leakage (Lemma 1), and in §5 they are flat across the noise levels tested. There is a second floor, for any method that compares single evaluations: the solver only sees* $\tilde{\Phi} = \frac{1}{2}\|r + \varepsilon\|^2$*, whose standard deviation is* $\sigma(\|r\|^2 + m\sigma^2/2)^{1/2} \approx \sigma\|r\|$ *away from* $r = 0$*, so two points are distinguishable from one evaluation each only when* $|\Phi(x) - \Phi(x')| \gtrsim \sigma(\|r(x)\| + \|r(x')\|)$*. Near the optimum, where* $\|r^*\| \approx \sigma_{obs}\sqrt{m}$ *for observation noise* $\sigma_{obs}$*, this is a relative floor of about* $4\sigma/(\sigma_{obs}\sqrt{m})$*: for the* $K$*=3 Voigt targets of §5 (*$m$*=57,* $\sigma_{obs}$*=0.01) it is* $3\%$*,* $10\%$ *and* $34\%$ *at evaluation noise* $0.1\%$*,* $0.3\%$ *and* $1\%$ *of the signal's rms. A success criterion tighter than the floor measures the floor, not the escape; we report both.*

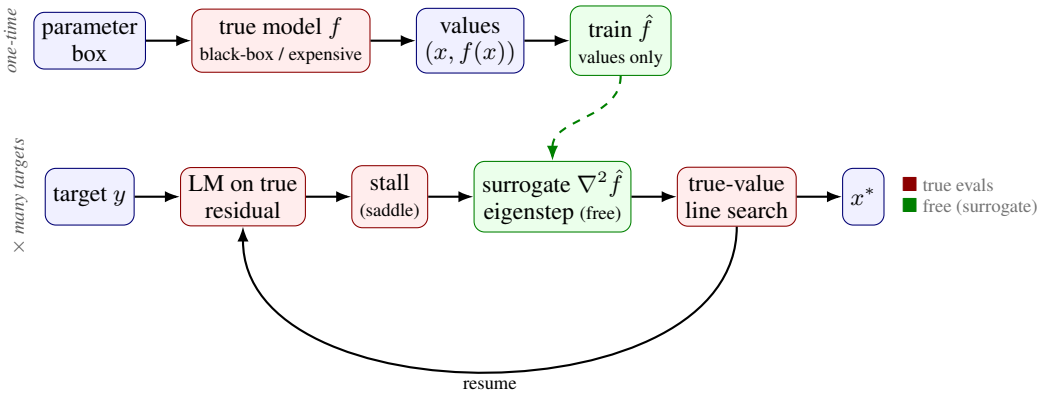


Figure 6: **Pipeline.** The surrogate is trained *once* on values of the true forward model (top), then reused (dashed) across every per-target solve (bottom). Only red steps spend true forward-model evaluations; the surrogate Hessian eigenstep that escapes the saddle is free. Step acceptance uses the true residual, so with exact evaluations an inaccurate surrogate can waste line-search evaluations but cannot raise the true objective; it can stop the solve early.

# E Method details

## E.1 Solver details

*LM core.* Trust-region LM with damping $\mu_0 = 10^{-3}$, gain-ratio acceptance ($\rho > 10^{-4}$ and a lower true loss), $\mu$ updated by the usual cubic rule, at most 20 damping trials per iteration and 200 iterations; stops on `gtol` ($\left\|\hat{J}^\top r\right\|_\infty < 10^{-7}$), `stalled` (no accepted step) or the budget. The same core with a central-difference Jacobian is the FD-LM baseline. *Eigenstep.* Two-sided line search along the unit eigenvector $\hat{v}$ of $\lambda_{\min}(\hat{H})$, with $\hat{H}$ formed and diagonalised in the physical coordinates and the step $x \pm \alpha\hat{v}$ taken in the same coordinates (no box rescaling of the direction; only the finite-difference steps of the baselines are scaled by the box half-width), $\alpha \in \{0.02, 0.05, 0.1, 0.2, 0.4\}$, points clamped to the box; the step is taken if any point lowers the true objective, else the round ends. With the secant correction a longer schedule (up to 1.6) is used, because a corrected LM is as blind as exact Gauss–Newton right after an eigenstep and the escape must be carried by repeated curvature steps. *Secant correction.* $E$ models $J - \hat{J}$ and is Broyden-updated from every LM trial step $\delta$, $E \leftarrow E + \big((r' - r) - (\hat{f}(x') - \hat{f}(x)) - E\delta\big)\delta^\top / \|\delta\|^2$, safeguarded by rescaling to $\|E\| \le \frac{1}{2}\left\|\hat{J}\right\|$ and reset to zero after three consecutive steps with gain ratio below 0.25. *Restarts.* Jitter of the split pair's shape parameters by 2, 5, 10, 20% of the box half-width in turn (uniform sign), from the initial point; the $\chi^2$ policy restarts while the final squared residual norm $\|r\|^2$ exceeds $2\,m(\sigma_{\text{obs}}^2 + \sigma^2)$ (at most 12 restarts), the jitter policy stops after two consecutive restarts without a 1% improvement (at most 12). *Surrogate training.* GELU MLP, $256 \times 4$, Adam; the coordinate network takes $(\operatorname{asinh}((t-\mu)/\sigma), \operatorname{asinh}(t-\mu), \log\sigma, \log\gamma)$ normalised to $O(1)$ over the box and predicts the dimensionless height $\sigma V$; 4000 single-peak spectra on the widest axis, 6000 steps of batch 8192, cosine learning-rate schedule from $2 \times 10^{-3}$. Its Jacobian and curvature are assembled blockwise by reverse-mode autodiff on the single-peak network, so they need no true-model calls and the network's input dimension does not grow with $K$ (the assembly's arithmetic and memory do; §F.2).

# F Additional experimental results

**Curvature fidelity (Table 4, Fig. 7).** Across a multi-variant sweep we measure the surrogate's derivatives against the true model's (autodiff, median relative error over 24 points), and three facts emerge. *(i) An order-of-derivative hierarchy:* the value fit is easiest, the Jacobian is accurate to a few percent even for mediocre surrogates ($\lesssim 10\%$), and the Hessian is 5–10× harder still—from

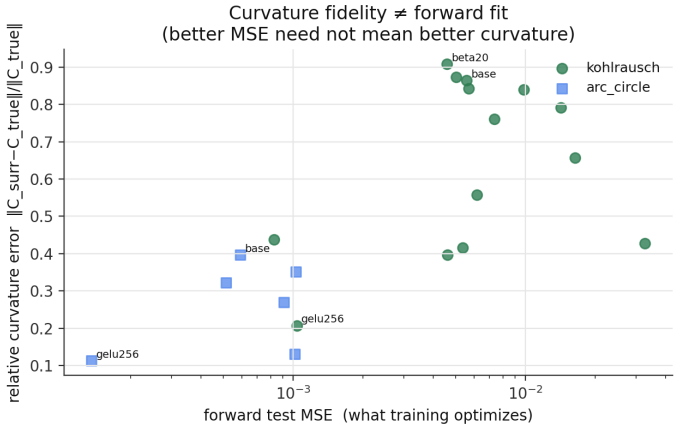


Figure 7: Curvature fidelity is not forward fit: relative curvature error is set by activation smoothness and width, and the best-forward-MSE variant is not the best-curvature one. Two sweep families: `kohlrausch`, the biexponential family of §5 (Table 4), and `arc_circle`, a short-arc circle fit $(c_x, c_y, R)$ to 25 noisy points on the ellipse $(2\cos\vartheta, \sin\vartheta)$, $\vartheta \in [-25°, 25°]$, with geometric residuals $\|p_j - c\| - R$ ($n$=3, $m$=25): a misspecified metrology problem with a weakly determined centre, included as a second, non-additive model.

Table 4: Surrogate derivative fidelity on the biexponential family of §5 (stretched-exponential targets; `kohlrausch` in the code) : median relative error over 24 fixed points, true derivatives by autodiff. Softplus $64 \times 4$ unless stated; *Sobolev-2* is value-only training on a fixed 1024-point set plus a penalty matching the network's second differences $(\hat{f}(x+hv) - 2\hat{f}(x) + \hat{f}(x-hv))/h^2$, $h = 0.05$, random unit $v$, to the same differences of the true model—curvature supervision built from values alone. *jac err* $= \|\hat{J} - J\|/\|J\|$; *Hess err* $= \|\nabla^2\hat{f} - \nabla^2 f\|/\|\nabla^2 f\|$; $C$ *err* $= \|\hat{C} - C\|/\|C\|$ for the correction $C = \sum_i r_i \nabla^2 f_i$ the solver actually consumes. Value fit, Jacobian, and Hessian form a strict hierarchy—each order of differentiation is harder—and curvature fidelity is *decoupled* from forward MSE: within the softplus family sharpening $\beta$ lowers MSE yet *worsens* curvature ($\beta$=20 has the best in-family MSE and the worst curvature, 0.91), and the lowest-MSE variant overall (a wide net, $8.3 \times 10^{-4}$) is still not the best-curvature one. Smoothness and width are the levers, not data.

| surrogate variant | fwd MSE | jac err | Hess err | $C$ err |
|---|---|---|---|---|
| ReLU (piecewise-linear) | 5.9e−3 | 0.152 | 1.00 | 1.00 |
| softplus $\beta$=10 (base) | 5.6e−3 | 0.095 | 0.84 | 0.86 |
| softplus $\beta$=20 (sharpest) | 4.6e−3 | 0.110 | 1.01 | 0.91 |
| tanh | 5.4e−3 | 0.113 | 0.75 | 0.42 |
| GELU | 4.6e−3 | 0.074 | 0.52 | 0.40 |
| width ×4 (best MSE) | **8.3e−4** | 0.046 | 0.51 | 0.44 |
| ensemble (3) | 6.2e−3 | 0.085 | 0.69 | 0.56 |
| data 4096 (vs. 512) | 5.0e−3 | 0.094 | 0.97 | 0.87 |
| Sobolev-2 (value-only) | 9.8e−3 | 0.110 | 0.95 | 0.84 |
| **GELU + width** | **1.0e−3** | **0.031** | **0.23** | **0.21** |

$\sim 20\%$ for the best variant up to $\sim 100\%$ (the size of the true Hessian itself) for the worst. *(ii) Curvature fidelity is decoupled from forward MSE:* within the softplus family (which has sharpness parameter $\beta$), larger $\beta$ improves the value fit but *worsens* curvature—$\beta$=20 (sharp) has the best in-family MSE and the *worst* curvature error (0.91), while $\beta$=1 (smooth) has weaker forward fit but better curvature—and, across all variants, the lowest-MSE surrogate (a wide net, $8.3 \times 10^{-4}$) is still not the best-curvature one. Optimising for value accuracy selects sharp, near-piecewise-linear fits whose second derivative is worst. *(iii) The levers are smoothness and width, not data:* GELU and width each help substantially and combine to 0.11–0.21 curvature error ($\sim$ 4–8× better than the softplus baseline), while an 8× larger training set barely moves any column; the second sweep

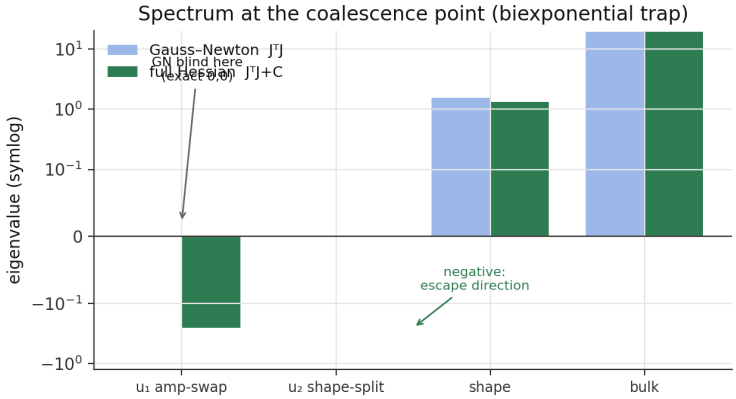


Figure 8: Spectrum at the biexponential coalescence trap. Gauss–Newton ($J^\top J$) has two exact zeros—it is blind in the coalescence subspace—while the full Hessian resolves one negative (escape) direction and one flat mode.

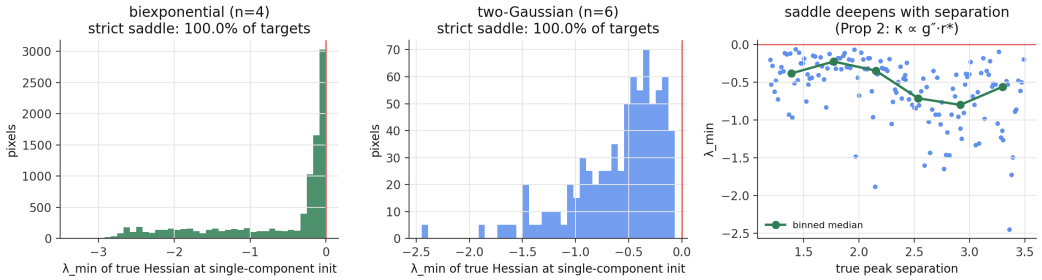


Figure 9: Left, centre: $\lambda_{\min}$ of the true Hessian at the single-component initialisation is negative for $100\%$ of targets in both families—for the Gaussian family this is now proved, not merely observed (Cor. 1). Right: the saddle deepens with component separation, as predicted by $\kappa \propto \varphi'' \cdot r^*$ (Prop. 1).

family (`arc_circle`, Fig. 7) agrees (there GELU+width reaches jac $0.018$, Hess $0.14$, $C$ $0.11$). Crucially, SG-LM does not need the Hessian to high accuracy—it consumes only the *sign and rough subspace* of one negative-curvature direction (Lemma 1)—so a $\sim 0.2$ relative Hessian error still yields a median escape-eigenvector alignment of $0.99$. The same fidelity carries over to the EPG simulator, where the value-only surrogate reproduces the correction matrix to within $\sim 10\%$ and the escape direction to alignment $0.99$.

**Anatomy of one saddle (Fig. 8).** At the biexponential trap $\mathrm{eig}(J^\top J) = (0, 0, 1.54, 19.2)$—GN blind in a 2-D subspace—while the full Hessian resolves it into $(-0.26, 0, 1.31, 19.2)$. A single eigenstep plus plain LM restart reaches the global optimum in $16$ iterations.

**Saddles are endemic (Fig. 9).** Our targets are two-component signals under model misspecification and noise—the regime in which one fits two components at all—so the single-fit residual $r^*$ is nonzero and, by Prop. 1(c), the point is a strict saddle whenever $\varphi'' \cdot r^* < 0$ for the scalar-shape biexponential family, and whenever $\lambda_{\min}(M) < 0$ for the vector-shape two-Gaussian family (Rem. 3). Over $10^4$ target/seed/noise combinations the relevant condition holds in *every* case (both families, all noise levels: $0$ of $10^4$ non-saddles). For the Gaussian family this is not a property of the sample: Cor. 1 shows $\lambda_{\min}(M) \le 0$ for *every* admissible residual, with equality only on the codimension-one set $M_{\mu\sigma} = 0$, which noisy targets miss almost surely. The values-only surrogate detects the negative direction in all but a $\le 0.2\%$ fraction, median eigenvector alignment $0.99$. Saddle depth grows with component separation, as Prop. 1 predicts. (When the reduced model itself is correctly specified and the data noiseless, $r^* \to 0$ and the point degenerates to a minimum—the boundary of the phenomenon.)

Table 5: **The same saddle in a grown network.** One-hidden-layer $\tanh$ networks grown from $K^*-1$ to $K^*$ units by splitting the unit with the most negative splitting index, every derivative exact (autograd); random teachers whose $(K^*-1)$-unit fit is at least $2\times$ worse than the $K^*$-unit optimum (instance counts in parentheses). *esc.*: % of instances on which the final loss is below $0.9\times$ the split-point loss; *opt.*: % within $1.05\times$ the best known $K^*$-unit fit; *grads*: median gradient evaluations ($^*$GD and Adam, started from a $1\%$ jitter of the split pair on the first $40$ instances: median steps to escape).

| method | $d$=1, 2 → 3, $n$=10 (109) | | | $d$=2, 3 → 4, $n$=17 (200) | | | $d$=2, 5 → 6, $n$=25 (197) | | |
|---|---|---|---|---|---|---|---|---|---|
| | esc. | opt. | grads | esc. | opt. | grads | esc. | opt. | grads |
| LM ($J^\top J$) | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| natural gradient | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| K-FAC | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| Hessian eigenstep + LM | 97 | 58 | 61 | 100 | 82 | 24 | 98 | 64 | 43 |
| splitting direction + LM | 97 | 58 | 62 | 100 | 82 | 24 | 98 | 63 | 42 |
| LM from 1% jitter | 98 | 58 | 64 | 100 | 82 | 24 | 98 | 63 | 42 |
| GD from 1% jitter* (40) | 5 | 0 | 15030 | 55 | 5 | 5074 | 38 | 0 | 8667 |
| Adam from 1% jitter* (40) | 98 | 38 | 864 | 88 | 50 | 275 | 92 | 52 | 618 |

### F.1 A MORE REALISTIC FORWARD MODEL: EPG RELAXOMETRY

The two families above are analytic models; to test the mechanism on a more realistic, recursion-defined forward model of the kind used clinically, we replace the exponential shape with the Extended Phase Graph (EPG) recursion for stimulated-echo-corrected multi-component $T_2$ relaxometry (myelin-water imaging): a state evolution over configuration orders that models the stimulated echoes produced by imperfect ($\alpha < 180°$) refocusing (Weigel, 2015; Prasloski et al., 2012; Malik et al., 2018). The per-echo signal is the output of that recursion, *not* $e^{-t/T_2}$ (at $\alpha = 180°$ it reduces to the exponential to machine precision, a check we use to validate the implementation). Concretely, writing $\theta = \log T_2$, the single-component shape is the echo train

$$\varphi(\theta;\alpha) = \text{EPG-CPMG}\big(e^{\theta}, \alpha;\, T_1, \text{TE}, 32\big) \in \mathbb{R}^{32}, \tag{7}$$

and the two-component model is the same amplitude-linear template as §3,

$$f(x) = a_1\,\varphi(\theta_1;\alpha) + a_2\,\varphi(\theta_2;\alpha), \qquad x = (a_1,\theta_1,a_2,\theta_2) \in \mathbb{R}^4, \quad m = 32, \tag{8}$$

now with $\varphi$ the EPG recursion rather than $e^{-t/T_2}$. Two-component fits ($n$=4, $m$=32 echoes; TE = 10 ms, $T_1$=1000 ms, effective refocusing $\alpha \in$ 150–180°) are initialised, as in practice, from the single-component fit. To be clear about scope: EPG is itself cheap and differentiable in principle, and we access it evaluation-only as before; this experiment therefore stresses the *realism of the failure mode and remedy* on a clinically used forward model, not the expensive- or non-differentiable-model cost regime, which we continue to emulate rather than meet in hardware.

*The theory transfers verbatim.* Working in $\log T_2$ coordinates, the coalescence point is stationary to $2 \times 10^{-13}$, $J^\top J$ is blind in the 2-dimensional split subspace, and the full Hessian resolves a strict negative escape mode (spectrum $(-0.006, 0, 0.07, 3.24)$) whose depth grows with component separation—Lemma 2 and Prop. 1 on the simulator, not a surrogate for it. Here we use the *basis* surrogate form (§4): a values-only GELU network learns only the single-component EPG echo train $\hat\varphi(\log T_2, \alpha) : \mathbb{R}^2 \to \mathbb{R}^{32}$, and the two-component forward model is assembled by the exact amplitude-linear combination $\hat f(x) = a_1\hat\varphi(\theta_1) + a_2\hat\varphi(\theta_2)$. This low-dimensional target is the cheaper one to train (training cost is accounted in §F.2; the EPG recursion itself is cheap here and stands in for the evaluation-only access regime), and it keeps all surrogate error on the shape's curvature—the quantity the escape step consumes—rather than on the (exactly known) amplitude structure. The surrogate ($N_{\text{train}} = 4000$ true evaluations; held-out MSE $1.1 \times 10^{-5}$) supplies a genuine negative-curvature descent direction at $19/19$ strict-saddle voxels (median eigenvector alignment $0.99$), confirming Lemma 1.

*A new wrinkle the analytic families hid.* Unlike the exponential and Gaussian toys, the *parametric* multi-component $T_2$ fit is intrinsically multi-modal—which is precisely why clinical pipelines fit a regularised $T_2$ spectrum by NNLS (Nonnegative Least Squares: constrain amplitudes $a_i \geq 0$) rather than a free few-component model. Escaping the coalescence saddle is therefore *necessary but not sufficient*: basin selection remains. This is naturally handled within the same amortization

principle by letting the surrogate perform the exploration itself—running the restart search entirely on the surrogate (zero true evaluations), deduplicating candidate basins, and polishing only the best two on the true model. On a 2000-voxel digital phantom (spatially varying myelin-water fraction, $B_1$ inhomogeneity, $1\%$ noise) this surrogate-guided solver **comes within** $1.4$ **points of multistart's reliability** ($98.5\%$ vs. $99.9\%$ within $1.05\times$) at 310 vs. $1442$ true evaluations per solve ($\mathbf{4.65}\times$ cheaper), and **recovers the myelin-water fraction equally well in the median** (median error $0.030$ vs. $0.030$)—so the win is a cost reduction at equal median accuracy, on a digital phantom; the tails and the $\approx 30$ vs. 2 failing voxels in 2000 are the difference that remains.

### F.2 Training amortization

The per-solve budgets above count only true evaluations *during the solve*; they exclude the one-time cost of training the surrogate. That cost is real—in a black-box or expensive-model setting, every training sample is a true forward-model call—and the surrogate is justified only when it is amortized across enough targets to repay it. We make the accounting explicit on the EPG relaxometry demonstration (§F.1), the setting where it matters most.

The values-only EPG surrogate costs $N_{\text{train}} = 4000$ true forward-model evaluations. With per-solve costs of $1442$ (multistart) and 310 (surrogate-guided), it breaks even after

$$N^* \;=\; \frac{N_{\text{train}}}{(\text{evals/solve, baseline}) - (\text{evals/solve, ours})} \;=\; \frac{4000}{1442 - 310} \;=\; 3.53 \;\approx\; 4 \text{ solves},$$

so a single relaxometry volume ($10^4$–$10^6$ voxels) amortizes training $\sim 10^4$–$10^5$ times over (Fig. 10). The bar is low *by design*: because escape needs only the sign and rough subspace of the curvature (Lemma 1), a few-thousand-sample surrogate already saturates both reliability and parameter recovery, so $N^*$ is a handful of solves rather than hundreds; enlarging the corpus is unnecessary. Conversely, when the number of targets is below $N^*$, or the forward model is both cheap and differentiable, the surrogate is not warranted—one takes the second-order term $C$ from the true model directly, by second-order autodifferentiation or finite differences, and still needs a negative-curvature step to use it.

**Wall-clock break-even.** Evaluation count is the right currency only for expensive models. Timed on the same machine, the surrogate's autodiff costs $0.6$ s per solve at $n{=}8$ and $13$ s at $n{=}32$ (assembling the $K$-component Jacobian and curvature), while the benchmark forward model costs $0.35$ ms per evaluation; SG-LM is faster in wall-clock only once an evaluation costs more than the surrogate time divided by the evaluations saved: against jitter multistart, the $2p^2$ probe and DFO-LS with restarts (Table 11) that is $0.6/(2721{-}166) = 0.23$ ms, $0.48$ ms and $1.1$ ms at $n{=}8$, and $13/(6895{-}228) = 1.9$ ms, $3.3$ ms and $5.1$ ms at $n{=}32$, before the baselines' own overheads (single core, double precision, no batching), which a PDE or Monte-Carlo solve exceeds and an analytic model—including the EPG recursion we use—does not.

**Scope-delimiting negatives.** (i) *Offline* (surrogate-only) solving: Gauss–Newton and full-Newton on the surrogate converge to the *same* surrogate minimiser—optimiser order cannot beat surrogate bias once both converge; the win requires true-residual anchoring. (ii) A *monolithic* surrogate of the full map degrades from the $n{=}4$ family to the $n{=}6$ one ($94\%$ to $54\%$, Table 7; two families, so dimension and model change together) because its Jacobian bias grows, misdirecting the LM steps—the leakage-$\beta$ mechanism of Lemma 1; learning the single-component shape instead (§4) removes the dependence on $n$, and the secant correction reduces the residual bias.

### F.3 Scaling to higher dimension

**Two opposing forces.** Growing the parameter dimension $n$ acts on the surrogate's value through two competing channels. The first favours the surrogate and is unconditional: the *cost of curvature by classical means scales badly*. A finite-difference Hessian of $\Phi$ costs $2n^2{+}2n$ true evaluations per probe (the FD-curvature line of Table 7 pays this), a secant model $n{+}1$ per initial interpolation set (points are reused thereafter; the tables count actual calls), and even a matrix-free Lanczos pass on FD Hessian–vector products costs $O(nk)$ true evaluations for $k$ Ritz values; multistart's reliability is likewise bought with a start count that must grow to cover a larger space. The surrogate returns curvature by autodifferentiating the network—$O(1)$ *true* evaluations regardless of $n$—so the

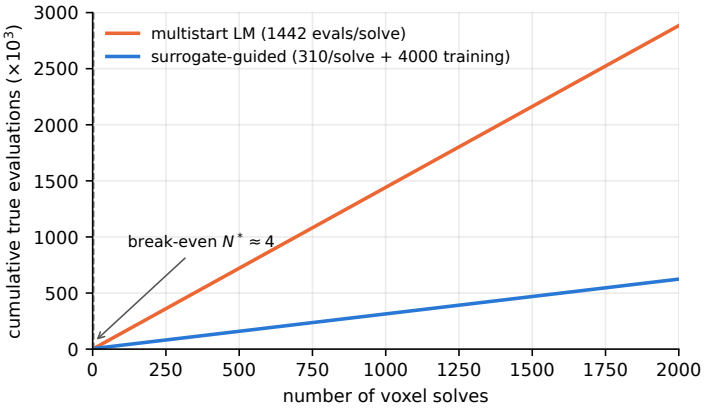


Figure 10: Amortization on EPG relaxometry (2000 voxels). Cumulative true forward-model evaluations vs. number of solves: the surrogate-guided solver pays a one-time $N_{\text{train}} = 4000$ training offset then adds little per solve, while multistart pays a steep cost every solve. Crossover at $N^* \approx 4$.

per-solve gap between surrogate-guided escape and any true-model second-order, secant or restart baseline widens with $n$. This is the mechanism measured in Fig. 2, and the direct-sum family below is its simplest instance.

The second channel is problem-dependent and cuts the other way: *a higher-dimensional instance is not automatically a harder, or better, test of escape.* Whether the coalescence trap deepens as one adds parameters is governed by Prop. 1: the escape curvature scales as $\kappa \propto \varphi'' \cdot r^*$, i.e. with the residual left by merging two components, which in turn scales with their *separation*. In a parametric fit over a bounded physical band that separation cannot stay large as the component count grows.

**Remark 6** (How vector magnitude and separation trade off in $\kappa \propto \varphi'' \cdot r^*$)**.** *Both $\varphi''$ and $r^*$ are $m$-dimensional vectors (one per data point). The dot product $\varphi'' \cdot r^*$ depends on two competing effects: (i)* ***Vector magnitude:*** *If more components mean more data ($m$ grows), both $\varphi''$ and $r^*$ become longer vectors, potentially making the dot product larger. (ii)* ***Separation effect:*** *If the two components get closer (crowding in a bounded domain), then $r^*$ shrinks in magnitude—the fit of a single component to two similar components is better. In the* direct-sum *family (Table 6, $b$ independent blocks, $n$=4$b$), separation is fixed and $m$=32$b$ grows, but the added data belong to other blocks: $\nabla^2\Phi$ is block diagonal, so the Rayleigh quotient along any split mode is the single-block $\kappa$, independent of $b$. What grows with $n$ is the* number *of negative-curvature directions ($b$ split modes, every one of which must be escaped), not their depth: the trap is held fixed per block and multiplied. In contrast, for a single bounded-domain model where you add $K$ components without adding data, separation shrinks and $\kappa$ weakens. The two families thus provide complementary tests: direct-sum isolates multiplicity at fixed depth, while single-domain isolates the crowding effect.*

**Remark 7** (Why the saddle shallows at higher $n$: crowding weakens component separation)**.** *The residual $r^*$ measures misfit when a single exponential or Gaussian replaces two. If the two components are well-separated (e.g., $T_{2,1} = 25\,ms$ and $T_{2,2} = 85\,ms$), $r^*$ is large and the saddle is deep. But in a bounded domain (say $T_2 \in [5, 150]\,ms$ for relaxometry), adding more components forces them to pack tighter. The nearest pair among 10 components is closer than the nearest pair among 2 (a random pair is not), so the separation of the unresolved pair shrinks and $r^*$ shrinks with it, making $\kappa$ shallower. Thus the saddle tends to shallow as $n$ grows, making the escape step easier to take, and harder to detect under noise—up to a point. Beyond that point (large $K$; Fig. 11), generic ill-conditioning from overlapping pairs dominates, and the saddle becomes a minor failure mode compared to near-degeneracy.*

**EPG $K$-pool: the trap can *shrink* with dimension (Fig. 11).** The sharpest illustration is multi-pool $T_2$ relaxometry itself: all $K$ pools sum into one shared echo train (Eq. 7),

$$f(x) = \sum_{k=1}^{K} a_k\, \varphi(\theta_k; \alpha) \in \mathbb{R}^{32}, \qquad x = (a_1, \theta_1, \ldots, a_K, \theta_K) \in \mathbb{R}^{2K} \;\; (n{=}2K,\; m{=}32). \tag{9}$$

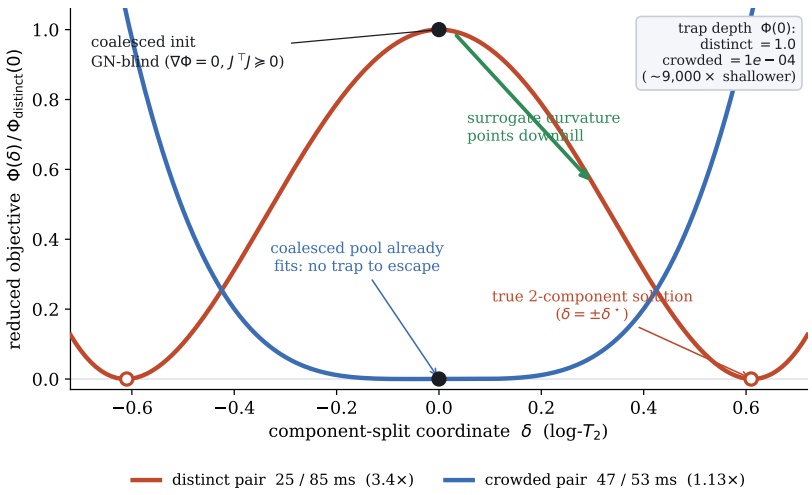


Figure 11: Why higher dimension need not mean a harder trap. True EPG objective along the component-split coordinate $\delta$ (log-$T_2$), amplitudes fixed at the true split. For a *distinct* pair (25/85 ms) the coalesced initialisation ($\delta$=0) is a strict saddle—Gauss–Newton is blind there ($\nabla\Phi$=0, $J^\top J \succeq 0$) and the surrogate curvature points to the true two-component solution at $\delta$=$\pm\delta^*$. For a *crowded* pair (47/53 ms, as forced by packing many pools into a bounded $T_2$ band) the same point is already essentially optimal: the trap is $\sim 9000\times$ shallower. Consistent with $\kappa \propto \varphi'' \cdot r^*$ (Prop. 1).

Fitting these $K$ pools from the standard coalescence initialisation—the two closest pools merged—the saddle a solver actually falls into is the *closest-pair* coalescence. Because relaxation times occupy a bounded band ($\sim$3–300 ms), packing more pools in forces the minimum spacing down: in our phantom the closest-pair $T_2$ ratio falls from $13.8\times$ at $K$=2 to $3.4\times$ ($K$=3), $1.68\times$ ($K$=5), and $1.11\times$ ($K$=10). By Prop. 1 the depth collapses in step: Fig. 11 evaluates the true EPG objective along the split direction and finds the coalesced point $\sim 9000\times$ closer to the optimum for a $1.13\times$ pair than for a $3.4\times$ pair —a strict saddle in the distinct case, effectively no trap in the crowded one. We therefore expect the difficulty to be *non-monotonic* in $K$: while the pairs stay distinct the trap is the strict saddle of Prop. 1; we conjecture, from the landscape of Fig. 11 and not from a solver sweep over $K$, which we do not report, that plain LM *recovers* at large $K$ ($K \approx$ 7–10) as the closest pair becomes near-degenerate and a regularised trust-region solver resolves it unaided. At large $K$ the failure mode has changed character—from a discrete coalescence saddle into the generic near-degeneracy (ill-posedness of multi-exponential inversion) that the escape step is not meant to address. The lesson: "higher $n$" must be built so the hard sub-structure persists rather than being diluted by crowding.

**Remark 8** (Scaling to higher dimension: preserve the saddle structure)**.** *When scaling a benchmark from $n = 4$ to larger $n$, the natural temptation is to add more parameters arbitrarily. Instead, the direct-sum family stacks independent blocks (here, EPG pulses) so that each contributes an identical coalescence saddle, keeping the geometry clean and interpretable. A naive $n$-exponential fit without this structure would degrade into generic ill-posedness (overlapping pairs, near-degeneracies from crowding), where the saddle-escape mechanism is no longer the primary failure mode and the result becomes uninformative about the phenomenon being studied.*

**A controlled scaling law: direct sums of coalescence blocks (Table 6).** To measure scaling with the trap held fixed, we build the family the embedding principle points to: a direct sum of $b$ *independent* two-component coalescence blocks, each the deep 25/85 ms EPG pair ($\theta^{(1)}$=log 25, $\theta^{(2)}$=log 85) with its own amplitudes and refocusing angle $\alpha_i$,

$$f(x) = \bigoplus_{i=1}^{b} \Big[ a_{i1}\,\varphi(\theta_{i1};\alpha_i) + a_{i2}\,\varphi(\theta_{i2};\alpha_i) \Big] \in \mathbb{R}^{32b}, \qquad x \in \mathbb{R}^{4b} \;\; (n{=}4b,\; m{=}32b), \tag{10}$$

Table 6: Direct-sum coalescence family ($b$ independent $25/85$ ms EPG blocks, $n{=}4b$), 100 spectra per $n$, unified true-evaluation accounting. Each cell: success rate / median true evaluations; success = true $\|r\|^2$ within $1.10\times$ a reference optimum, the threshold this earlier family was scored with (the families of §5 use $1.05\times$; the two are not compared). Plain Gauss–Newton LM is blind in all $b$ split modes; robust TRF escapes only a minority at every $n$ (19–32%); surrogate-guided matches multistart's reliability at a steady $\sim 9\times$ lower cost ($8.4$–$9.2\times$), the absolute saving widening with $n$. (The surrogate here is the monolithic single-block form and the classical comparators are only LM and TRF, so it is a secondary result.)

| $n$ | plain GN-LM | `scipy`-TRF | multistart | **surrogate-guided** |
|---|---|---|---|---|
| 4 | 2% / 27 | 19% / 10 | 100% / 626 | **99% / 70** |
| 8 | 0% / 51 | 24% / 18 | 100% / 1570 | **100% / 171** |
| 12 | 0% / 75 | 26% / 26 | 100% / 2411 | **100% / 268** |
| 16 | 0% / 99 | 30% / 34 | 100% / 3374 | **100% / 393** |
| 20 | 0% / 164 | 30% / 42 | 100% / 4394 | **100% / 504** |
| 24 | 0% / 196 | 29% / 50 | 100% / 5310 | **100% / 633** |
| 32 | 0% / 260 | 32% / 66 | 100% / 7981 | **100% / 909** |

where $\bigoplus$ is vertical concatenation. Each block is an exact, $n$-independent strict saddle, so applying Prop. 1 blockwise (the embedding principle) the Gauss–Newton-blind subspace $\ker(J)$ has dimension $2b$ — $b$ exactly flat swap modes and $b$ split modes, the latter carrying the negative curvature — and both *grow linearly with* $n$, while every block is resolved by the same $K$-independent single-pool surrogate. This also removes the model-family confound of the $n{=}4 \to 6$ comparison, where the model changed with $n$; here only the block count does. Table 6 sweeps to $n{=}32$: plain Gauss–Newton LM is trapped at $\leq 2\%$ at every $n$ (blind in all $b$ split modes); a robust trust-region solver (`scipy` TRF) escapes only 19–32%; multistart stays reliable but its cost climbs steeply ($626 \to 7981$ true evaluations, $n{=}4 \to 32$); and surrogate-guided escape holds $\geq 99\%$ at roughly one robust-solve's cost, $\sim 9\times$ below multistart at every $n$ with the absolute saving widening. The $K$-Voigt families of §5 are the stronger version of this experiment: coupled rather than block-separable, against every classical escape rather than two, with the ratio growing with $n$ rather than steady.

**What this family represents.** The direct-sum construction is the canonical shape of *block-separable* amortized inverse problems, in which a low-dimensional nonlinearity recurs many times with independent parameters. Multi-region or multi-voxel relaxometry fitted jointly, batched spectral deconvolution over disjoint windows, panel and mixed-effects models with a per-unit nonlinear term, and any setting where dimension grows by *replicating* a hard sub-problem rather than by coupling, all share this structure: the hard nonlinearity is low-dimensional and repeated, so one $K$-independent surrogate resolves every instance while the blind subspace grows with the number of repeats. The honest caveat is that the measurement is block-separable and the surrogate solver exploits that known structure—but the trapped baselines (Gauss–Newton, TRF) have identical structural access and still fail, so the advantage is the surrogate's per-block negative-curvature direction, supplied for free, not privileged information.

### F.4 Supplementary tables: protocol, the grid-start setting, bounds, seeds, box, ablation

**Full protocol.** Four forward-model families, all accessed evaluation-only by every method. **Biexponential** $a_1 e^{-k_1 t} + a_2 e^{-k_2 t}$ ($n{=}4$, $m{=}40$ points on $[0.05, 5]$; box $a_i \in [0, 1.2]$, $k_1 \in [0.05, 2.5]$, $k_2 \in [0.5, 10]$; FLIM/relaxometry): targets are stretched exponentials $\exp(-(t/\tau)^\beta)$ with $\tau \in [0.6, 1.8]$, $\beta \in [0.45, 0.9]$ (a misspecified model) plus Gaussian noise of standard deviation $0.01$; 200 targets. **Two-Gaussian** $\sum_{i=1}^{2} a_i \exp(-((t-\mu_i)/\sigma_i)^2)$ ($n{=}6$, $m{=}50$ points on $[0, 10]$; box $a_i \in [0.2, 1.5]$, $\mu_i \in [2, 8]$, $\sigma_i \in [0.4, 2]$; spectroscopy): two peaks of amplitude $[0.4, 1.3]$, width $[0.5, 1.3]$ and separation $[1.2, 3.5]$ about a centre in $[3.5, 6.5]$, plus noise $0.01$; 150 targets. **$K$-Voigt** $\sum_{k=1}^{K} a_k V(t - \mu_k; \sigma_k, \gamma_k)$ ($n{=}4K$, $K = 2, 3, 4, 6, 8$, i.e. $n = 8$–$32$), unit-area Voigt profiles on an axis of $K{-}1$ slots of width 14 sampled every $0.5$ ($m = 28(K-1)+1$, so $m/n$ lies between 3.6 and 6.2): each target has $K{-}2$ resolved "spectator" peaks and one *unresolved pair* separated by 2–4 units, about one line width, with amplitudes in $[3, 6]$, $\sigma \in [0.8, 1.6]$, $\gamma \in [0.3, 1]$, and Lorentzian tails

coupling all slots; noise $0.01$; $100$ targets per $K$. **Transmission** $\exp(-\sum_k a_k V(\cdot))$ (Beer–Lambert absorption, the same peaks at half amplitude): the same targets and axis, but non-separable, so no amplitude can be projected out. Every parameter has a box (amplitudes $[0, 14]$, centres on the axis, $\sigma \in [0.4, 4]$, $\gamma \in [0.03, 3]$ for Voigt) shared by all components.

*Initialisation.* The converged reduced fit, split: for $n$=4 and 6 the single-component optimum, computed in double precision with exact derivatives (uncounted, since the reduced fit is a given in the workflow), duplicated with half the amplitude each; for $K$-Voigt the converged $(K-1)$-peak fit with the peak of largest local residual energy split into two coincident halves. Where the reduced optimum is interior to the box this is the exact critical point of §3: median $\left\|J^\top r\right\|_\infty$ at the start is $10^{-8}$ ($n$=4, 6) and $10^{-8}$–$10^{-5}$ (Voigt). On $41\%$ ($n$=4), $32\%$ ($n$=6) and 33–$58\%$ (Voigt) of targets the reduced fit sits on a bound—for the Voigt pair almost always the Lorentzian width at its lower bound, because a blended pair is flat-topped—and the point is stationary only in the projected sense. Table 7 uses the interior targets, which are what the theory is about; the $K$-Voigt tables and figures use all $100$ targets per $K$, and the at-bound targets are reported separately in §F.4. *Scoring.* Every call to the forward model is counted (rejected trial steps, stencil points, restarts and reference points included; a central-difference Jacobian is $2n$ calls, our symmetric finite-difference Hessian $2n^2+2n$, an initial DFO-LS model $n+1$); *success* means the final true $\|r\|^2$ is within $1.05\times$ the best value found by any method or by an uncounted reference (six-start FD-LM at $n \le 6$; the exact fit from the true parameters for Voigt), a pooled best, not a certified optimum. With $100$ targets per cell, success rates are finite-sample estimates and the methods are compared on the same targets; we do not read differences of a few points between methods as established. Budgets: $4000$ evaluations at $n \le 6$, $1000n$ for Voigt, nominal—the counter is not a hard stop, a library-internal Jacobian or stencil started inside the budget can finish outside it, and every recorded count is the count actually spent (VarPro-TRF reaches $2\times$ the budget on two targets, the lazy trust-region baselines exceed it by up to $7\%$ on 3–$4\%$ of runs, and the DFO-LS averaging runs of Table 17 are the largest case, see there); under evaluation noise most failing methods stop well within budget on their own noise-adapted criteria (one sample per call except in Table 17; DFO-LS in its noisy-objective mode). *Surrogates.* One GELU MLP ($256 \times 4$) of the single-component shape per family, values only: $\hat\varphi(k) \in \mathbb{R}^{40}$ and $\hat\varphi(\mu, \sigma) \in \mathbb{R}^{50}$ for the analytic families (trained with $512$ fresh samples per epoch for $800$ and $2000$ epochs, the recipe of Table 4—i.e. $4 \times 10^5$ and $10^6$ single-component evaluations, uncounted: for these cheap analytic families training is not amortised in the tables, and at $n$=4 DFO-LS is the better method outright (Table 7); the amortisation argument is made for the Voigt surrogate and for EPG, §F.2), and one coordinate network $\hat\varphi(t - \mu, \log\sigma, \log\gamma)$ for both Voigt families, trained once, on $4000$ single-peak spectra (true evaluations) on the $K$=8 axis, and reused at every $K$. This assumes the single-component shape can be evaluated on its own—a $K$=1 call of the forward model, or for the transmission family the log of one—which is the oracle the surrogate is trained from; under evaluation noise the training noise is applied to the component values, so the noisy transmission experiments assume separate, component-level training access: the solver observes noise after the exponential, the surrogate was trained on noise before it, and the two observation models are not equivalent (the transformed noise is signal-dependent and biased). The additive-family noise results carry no such assumption. Its fidelity at the coalescence points ($K$=3; the network is the same at every $K$): forward error $0.4\%$, Jacobian $2.6\%$, correction $C$ $3.5\%$, escape-eigenvector alignment $1.000$. *Baselines* (the Jacobian-based ones with the same central-difference Jacobian; DFO-LS and Py-BOBYQA build their own interpolation models; §2): FD-LM (the paper's trust-region LM), scipy's trust-region reflective solver (Branch et al., 1999; Virtanen et al., 2020), VarPro with the same LM core and with TRF on the reduced residual (Golub & Pereyra, 2003), an NL2SOL-style Dennis–Gay–Welsch secant method with an exact trust-region step (Dennis et al., 1981), trust-region Newton and ARC (Cartis et al., 2011) with finite-difference Hessians (re-differenced every fifth accepted step), the finite-difference curvature probe (LM, then a $2n^2+2n$ Hessian and the same eigenstep line search as SG-LM whenever LM stops), the theory-aware probe ($2p^2$ evaluations for the split matrix $M$), DFO-LS with and without its restarts (Cartis et al., 2019) (box mapped to the unit cube; $\rho_{\text{beg}} = 0.1$), Py-BOBYQA (Powell, 2009) (a general-purpose model-based solver on $\Phi$ rather than on the residual; dominated by DFO-LS everywhere—$93\%$ at 397 mean evaluations at $n \le 6$, $49\%$ at $4228$ across the Voigt family—and not tabulated), adaptive multistart LM (random box starts until two consecutive starts fail to improve by $1\%$) and jitter multistart LM (restarts from the coalescence point with the pair's shape parameters perturbed by 2–$20\%$ of the box, same stopping rule). We do not report the perturbed-GD and Adam baselines of our earlier experiments: untuned they fail, and tuned they cannot compete on evaluation count, since each gradient costs $2n$.

**The grid-start setting, and its two artefacts (Table 8).** Our original tables were produced in single precision from a grid-search single-component fit that is stationary only to $\left\|J^\top r\right\|_\infty \approx 10^{-2}$ ($n$=4) and $10^{-1}$ ($n$=6). Two things follow. FD-LM leaves the point on 91% of $n$=4 targets there, and on 74% in double precision from the same start: the escape is the non-stationarity of the start and the single-precision rounding of the central difference, not curvature, and it disappears on the exact point (Table 7). And scipy-TRF's "0% in 14 evaluations" is a precision artefact: its default two-point step is below single-precision resolution, so it differentiates rounding noise and stops; in double precision from the same start it reaches 100% ($n$=4) and 94% ($n$=6). The 4000-evaluation multistart of that version restarts until its budget is spent, so the 42× it implied was a ratio against a cap; adaptive multistart stops at 475–930 median evaluations in single precision and 486–3036 in double.

Table 7: Blindness on the exact critical point, all methods and variants Success % / median / mean true evaluations, interior targets, double precision.

| | biexponential, $n$=4 (118 targets) | | | two-Gaussian, $n$=6 (102 targets) | | |
|---|---|---|---|---|---|---|
| method | succ. % | median | mean | succ. % | median | mean |
| *Gauss–Newton with exact / small-step derivatives (blind)* | | | | | | |
| FD-LM | 0 | 9 | 9 | 0 | 13 | 13 |
| scipy-TRF | 0 | 10 | 11 | 1 | 14 | 15 |
| VarPro-LM | 0 | 6 | 6 | 1 | 10 | 12 |
| VarPro-TRF | 40 | 38 | 40 | 53 | 74 | 75 |
| NL2SOL (DGW secant) | 0 | 9 | 9 | 0 | 13 | 38 |
| FD curvature probe (original) | 0 | 9 | 9 | 0 | 13 | 13 |
| *second-order and secant-model escapes (true evaluations)* | | | | | | |
| TR-Newton, FD Hessian | 53 | 3049 | 2453 | 93 | 694 | 1039 |
| ARC, FD Hessian | 51 | 3217 | 2611 | 93 | 695 | 1022 |
| FD curvature probe ($2n^2+2n$) | 87 | 228 | 614 | 94 | 352 | 593 |
| theory-aware FD probe ($2p^2$) | 76 | 1960 | 1193 | 94 | 192 | 437 |
| DFO-LS | 100 | 66 | 71 | 83 | 115 | 117 |
| DFO-LS + restarts | 100 | 294 | 374 | 95 | 348 | 391 |
| adaptive multistart LM | 100 | 372 | 385 | 100 | 530 | 1621 |
| *surrogate curvature (this paper)* | | | | | | |
| SG-LM (monolithic surrogate) | 94 | 64 | 109 | 54 | 83 | 109 |
| SG-LM (component surrogate) | 85 | 98 | 165 | 86 | 69 | 80 |
| SG-LM + secant | 84 | 90 | 146 | 86 | 85 | 95 |
| SG-LM + $\chi^2$ restarts | 93 | 99 | 183 | 95 | 70 | 339 |
| SG-LM + secant + $\chi^2$ restarts | 92 | 90 | 149 | 96 | 87 | 357 |

Table 8: The grid-start setting: single or double precision, grid start (not stationary), all targets. Success % / median true evaluations.

| | biexponential, $n$=4 (200) | | | | two-Gaussian, $n$=6 (150) | | | |
|---|---|---|---|---|---|---|---|---|
| | float32, grid | | float64, grid | | float32, grid | | float64, grid | |
| method | % | med. | % | med. | % | med. | % | med. |
| FD-LM | 91 | 173 | 74 | 172 | 1 | 168 | 58 | 298 |
| scipy-TRF | 0 | 14 | 100 | 97 | 0 | 7 | 94 | 147 |
| VarPro-LM | 86 | 104 | 87 | 82 | 30 | 134 | 84 | 204 |
| NL2SOL (DGW secant) | 88 | 180 | 71 | 326 | 66 | 290 | 66 | 279 |
| FD curvature probe (original) | 92 | 245 | 74 | 172 | 61 | 683 | 58 | 304 |
| FD curvature probe ($2n^2+2n$) | 92 | 213 | 78 | 212 | 66 | 542 | 67 | 474 |
| TR-Newton, FD Hessian | 84 | 955 | 84 | 1007 | 93 | 796 | 93 | 694 |
| adaptive multistart LM | 100 | 475 | 100 | 486 | 100 | 930 | 100 | 3036 |
| multistart | 100 | 4000 | 100 | 4002 | 100 | 4003 | 100 | 4004 |
| SG-LM (monolithic surrogate) | 98 | 44 | 97 | 52 | 55 | 82 | 52 | 92 |
| SG-LM+restart (original) | 100 | 90 | 100 | 110 | 99 | 497 | 99 | 460 |
| SG-LM (component surrogate) | 88 | 56 | 71 | 67 | 61 | 86 | 49 | 88 |

Table 9: Interior versus at-bound reduced fits, double precision, analytic families. On at-bound targets the split point is stationary only in the projected sense; at $n$=4 every method then leaves it, at $n$=6 the projected LM solvers slide along the $\sigma$ bound and the $\chi^2$ restarts recover.

| | $n$=4 | | | | $n$=6 | | | |
|---|---|---|---|---|---|---|---|---|
| | interior (118) | | at a bound (82) | | interior (102) | | at a bound (48) | |
| method | % | med. | % | med. | % | med. | % | med. |
| FD-LM | 0 | 9 | 100 | 139 | 0 | 13 | 0 | 53 |
| scipy-TRF | 0 | 10 | 100 | 91 | 1 | 14 | 94 | 99 |
| VarPro-TRF | 40 | 38 | 100 | 62 | 53 | 74 | 96 | 74 |
| FD curvature probe ($2n^2$+$2n$) | 87 | 228 | 100 | 179 | 94 | 352 | 6 | 148 |
| theory-aware FD probe ($2p^2$) | 76 | 1960 | 100 | 139 | 94 | 192 | 27 | 72 |
| TR-Newton, FD Hessian | 53 | 3049 | 100 | 1016 | 93 | 694 | 92 | 630 |
| DFO-LS | 100 | 66 | 100 | 65 | 83 | 115 | 94 | 90 |
| DFO-LS + restarts | 100 | 294 | 100 | 304 | 95 | 348 | 94 | 252 |
| adaptive multistart LM | 100 | 372 | 100 | 387 | 100 | 530 | 100 | 488 |
| SG-LM (component surrogate) | 85 | 98 | 100 | 58 | 86 | 69 | 10 | 41 |
| SG-LM + secant + $\chi^2$ restarts | 92 | 90 | 100 | 50 | 96 | 87 | 100 | 258 |

Table 10: $K$-Voigt, interior versus at-bound reduced fits: in $95\%$ of the at-bound cases the coordinate on the bound is the merged pair's $\log\gamma$ at its lower limit (a blended pair is flat-topped, so the reduced fit drives the Lorentzian width to zero). Success % / median evaluations.

| | $K$=3 ($n$=12) | | | | $K$=8 ($n$=32) | | | |
|---|---|---|---|---|---|---|---|---|
| | interior (56) | | $\log\gamma$ at bound (44) | | interior (46) | | $\log\gamma$ at bound (54) | |
| method | % | med. | % | med. | % | med. | % | med. |
| theory-aware FD probe ($2p^2$) | 95 | 480 | 98 | 2082 | 98 | 1332 | 91 | 4726 |
| FD curvature probe ($2n^2$+$2n$) | 98 | 1086 | 100 | 2664 | 100 | 5538 | 91 | 8782 |
| DFO-LS + restarts | 100 | 857 | 100 | 788 | 100 | 2098 | 100 | 2234 |
| jitter multistart LM | 100 | 1386 | 100 | 3834 | 100 | 3598 | 100 | 8070 |
| SG-LM | 91 | 88 | 89 | 99 | 91 | 98 | 80 | 112 |
| SG-LM + secant + $\chi^2$ restarts | 98 | 94 | 100 | 140 | 98 | 108 | 98 | 154 |
| SG-LM + jitter restarts | 93 | 202 | 91 | 264 | 98 | 204 | 94 | 282 |

Table 11: Scaling, additive Voigt family: success % and mean true evaluations per solve, all 100 targets per $n$ (Fig. 2, left).

| method | $n$=8 | | $n$=12 | | $n$=16 | | $n$=24 | | $n$=32 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | % | mean | % | mean | % | mean | % | mean | % | mean |
| FD-LM | 0 | 824 | 0 | 774 | 0 | 1366 | 0 | 2213 | 0 | 2064 |
| VarPro-LM | 0 | 103 | 0 | 891 | 0 | 1338 | 0 | 1202 | 0 | 2396 |
| VarPro-TRF | 19 | 75 | 24 | 147 | 28 | 171 | 22 | 222 | 30 | 340 |
| NL2SOL (DGW secant) | 0 | 1000 | 0 | 1079 | 0 | 1793 | 8 | 2672 | 13 | 2664 |
| FD curvature probe ($2n^2$+$2n$) | 96 | 1715 | 99 | 2184 | 95 | 3481 | 95 | 7020 | 95 | 8273 |
| theory-aware FD probe ($2p^2$) | 95 | 1423 | 96 | 1557 | 94 | 2459 | 91 | 4245 | 94 | 4095 |
| TR-Newton, FD Hessian | 91 | 2194 | 99 | 3232 | 97 | 5799 | 93 | 12760 | 94 | 21356 |
| jitter multistart LM | 94 | 2721 | 100 | 2810 | 99 | 4771 | 100 | 6839 | 100 | 6895 |
| DFO-LS | 91 | 220 | 90 | 268 | 88 | 401 | 79 | 622 | 69 | 815 |
| DFO-LS + restarts | 98 | 721 | 100 | 961 | 100 | 1347 | 100 | 2193 | 100 | 2781 |
| SG-LM | 84 | 105 | 90 | 101 | 80 | 104 | 85 | 106 | 85 | 116 |
| SG-LM + secant | 92 | 135 | 91 | 149 | 83 | 139 | 87 | 143 | 88 | 168 |
| SG-LM + secant + $\chi^2$ restarts | 99 | 166 | 99 | 198 | 97 | 226 | 99 | 284 | 98 | 228 |
| SG-LM + jitter restarts | 86 | 297 | 92 | 266 | 88 | 294 | 96 | 314 | 96 | 322 |

Table 12: Scaling, transmission family (non-separable; VarPro does not apply): success % and mean true evaluations (Fig. 2, right).

| method | $n$=8 | | $n$=12 | | $n$=16 | | $n$=24 | | $n$=32 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | % | mean | % | mean | % | mean | % | mean | % | mean |
| FD-LM | 0 | 513 | 0 | 416 | 0 | 829 | 0 | 1173 | 1 | 1075 |
| NL2SOL (DGW secant) | 0 | 1104 | 0 | 833 | 4 | 1616 | 7 | 2565 | 11 | 2599 |
| FD curvature probe ($2n^2+2n$) | 91 | 1497 | 99 | 2107 | 95 | 3923 | 95 | 5873 | 98 | 9036 |
| theory-aware FD probe ($2p^2$) | 85 | 1155 | 95 | 1489 | 83 | 2284 | 86 | 2981 | 89 | 3932 |
| TR-Newton, FD Hessian | 84 | 2874 | 96 | 4209 | 96 | 6764 | 94 | 13167 | 98 | 21437 |
| jitter multistart LM | 91 | 2958 | 100 | 3409 | 100 | 5495 | 99 | 7047 | 100 | 9228 |
| DFO-LS | 82 | 218 | 86 | 286 | 81 | 383 | 70 | 620 | 45 | 651 |
| DFO-LS + restarts | 97 | 757 | 99 | 1130 | 99 | 1529 | 99 | 2520 | 100 | 3904 |
| SG-LM | 79 | 100 | 89 | 118 | 82 | 119 | 84 | 114 | 85 | 112 |
| SG-LM + secant | 81 | 130 | 97 | 133 | 83 | 154 | 86 | 133 | 88 | 176 |
| SG-LM + secant + $\chi^2$ restarts | 93 | 187 | 99 | 145 | 92 | 188 | 93 | 168 | 92 | 187 |
| SG-LM + jitter restarts | 85 | 300 | 93 | 325 | 94 | 334 | 99 | 365 | 100 | 285 |

Table 13: Stochastic forward model, additive Voigt family: success within $1.05\times$ / within $2\times$ / fraction of targets on which the pair was separated ("split"), 100 targets per cell (Fig. 3).

| method | $s = 0.1\%$ | | | $s = 0.3\%$ | | | $s = 1\%$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $1.05\times$ | $2\times$ | split | $1.05\times$ | $2\times$ | split | $1.05\times$ | $2\times$ | split |
| $K$=3 ($n$=12) | | | | | | | | | |
| FD-LM | 3 | 5 | 12 | 0 | 1 | 11 | 0 | 1 | 20 |
| scipy-TRF | 0 | 6 | 31 | 0 | 1 | 26 | 0 | 1 | 30 |
| FD curvature probe ($2n^2+2n$) | 20 | 38 | 61 | 0 | 2 | 21 | 0 | 1 | 33 |
| theory-aware FD probe ($2p^2$) | 9 | 18 | 27 | 0 | 1 | 11 | 0 | 1 | 21 |
| jitter multistart LM | 52 | 69 | 94 | 1 | 21 | 89 | 0 | 0 | 71 |
| DFO-LS | 25 | 60 | 88 | 0 | 31 | 80 | 0 | 5 | 86 |
| DFO-LS + restarts | 46 | 95 | 99 | 0 | 48 | 97 | 0 | 5 | 93 |
| SG-LM | 73 | 83 | 88 | 46 | 72 | 81 | 3 | 49 | 69 |
| SG-LM + secant + $\chi^2$ restarts | 69 | 95 | 98 | 21 | 82 | 98 | 1 | 53 | 97 |
| SG-LM + jitter restarts | 85 | 87 | 94 | 65 | 83 | 92 | 12 | 68 | 89 |
| $K$=6 ($n$=24) | | | | | | | | | |
| FD-LM | 0 | 5 | 0 | 0 | 5 | 0 | 0 | 5 | 9 |
| scipy-TRF | 0 | 10 | 12 | 0 | 6 | 4 | 0 | 5 | 6 |
| FD curvature probe ($2n^2+2n$) | 0 | 5 | 8 | 0 | 6 | 34 | 0 | 5 | 27 |
| theory-aware FD probe ($2p^2$) | 0 | 5 | 5 | 0 | 6 | 14 | 0 | 5 | 19 |
| jitter multistart LM | 0 | 6 | 76 | 0 | 8 | 68 | 0 | 5 | 66 |
| DFO-LS | 1 | 18 | 47 | 0 | 9 | 52 | 0 | 5 | 54 |
| DFO-LS + restarts | 3 | 70 | 89 | 0 | 21 | 87 | 0 | 5 | 74 |
| SG-LM | 73 | 81 | 85 | 49 | 68 | 77 | 8 | 51 | 56 |
| SG-LM + secant + $\chi^2$ restarts | 71 | 97 | 98 | 23 | 90 | 91 | 1 | 64 | 92 |
| SG-LM + jitter restarts | 88 | 92 | 95 | 80 | 88 | 96 | 19 | 86 | 96 |

Table 14: Stochastic forward model, transmission family: same columns as Table 13.

| method | $s = 0.1\%$ | | | $s = 0.3\%$ | | | $s = 1\%$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1.05× | 2× | split | 1.05× | 2× | split | 1.05× | 2× | split |
| *$K$=3 ($n$=12)* | | | | | | | | | |
| FD-LM | 5 | 30 | 45 | 1 | 26 | 58 | 0 | 16 | 64 |
| scipy-TRF | 0 | 23 | 25 | 0 | 18 | 35 | 0 | 18 | 38 |
| FD curvature probe ($2n^2+2n$) | 22 | 70 | 85 | 1 | 32 | 71 | 0 | 15 | 75 |
| theory-aware FD probe ($2p^2$) | 9 | 38 | 59 | 1 | 26 | 64 | 0 | 16 | 70 |
| jitter multistart LM | 45 | 78 | 94 | 6 | 68 | 93 | 0 | 20 | 91 |
| DFO-LS | 3 | 66 | 88 | 0 | 37 | 91 | 0 | 17 | 85 |
| DFO-LS + restarts | 4 | 88 | 100 | 0 | 47 | 91 | 0 | 17 | 87 |
| SG-LM | 51 | 78 | 74 | 29 | 68 | 67 | 0 | 41 | 47 |
| SG-LM + secant + $\chi^2$ restarts | 37 | 93 | 89 | 6 | 77 | 84 | 0 | 51 | 72 |
| SG-LM + jitter restarts | 73 | 87 | 95 | 50 | 88 | 92 | 0 | 73 | 90 |
| *$K$=6 ($n$=24)* | | | | | | | | | |
| FD-LM | 0 | 30 | 5 | 0 | 30 | 8 | 0 | 29 | 28 |
| scipy-TRF | 0 | 29 | 9 | 0 | 29 | 5 | 0 | 29 | 6 |
| FD curvature probe ($2n^2+2n$) | 0 | 30 | 13 | 0 | 31 | 23 | 0 | 27 | 41 |
| theory-aware FD probe ($2p^2$) | 0 | 30 | 8 | 0 | 30 | 12 | 0 | 29 | 34 |
| jitter multistart LM | 0 | 42 | 65 | 0 | 35 | 54 | 0 | 29 | 58 |
| DFO-LS | 0 | 34 | 46 | 0 | 30 | 49 | 0 | 28 | 51 |
| DFO-LS + restarts | 0 | 78 | 91 | 0 | 32 | 66 | 0 | 28 | 58 |
| SG-LM | 49 | 75 | 62 | 22 | 59 | 46 | 0 | 42 | 30 |
| SG-LM + secant + $\chi^2$ restarts | 39 | 92 | 77 | 7 | 83 | 75 | 0 | 61 | 57 |
| SG-LM + jitter restarts | 79 | 91 | 96 | 63 | 89 | 93 | 1 | 86 | 85 |

Table 15: Surrogate training seed: success % of the SG-LM variants with three independently trained coordinate networks (the theory-aware probe as the seed-free reference).

| | $K$=3 ($n$=12), seed 1 / 2 / 3 | | | $K$=6 ($n$=24), seed 1 / 2 / 3 | | |
|---|---|---|---|---|---|---|
| method | % | % | % | % | % | % |
| theory-aware FD probe ($2p^2$) | 96 | 96 | 96 | 91 | 91 | 91 |
| SG-LM | 90 | 82 | 96 | 85 | 76 | 88 |
| SG-LM + secant | 91 | 95 | 100 | 87 | 87 | 90 |
| SG-LM + secant + probe fallback | 95 | 97 | 100 | 89 | 88 | 90 |
| SG-LM + jitter restarts | 92 | 91 | 100 | 96 | 92 | 99 |

Table 16: Widening the Lorentzian-width box from $\gamma \geq 0.03$ to $\gamma \geq 0.003$ (own surrogate): success % / median evaluations. Nothing moves, because the reduced fit follows the bound down.

| method | $n$=8 | | $n$=12 | | $n$=16 | | $n$=24 | | $n$=32 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | % | med. | % | med. | % | med. | % | med. | % | med. |
| FD curvature probe ($2n^2+2n$) | 98 | 1120 | 100 | 1266 | 98 | 2644 | 96 | 4822 | 96 | 6646 |
| theory-aware FD probe ($2p^2$) | 96 | 850 | 96 | 660 | 93 | 1574 | 91 | 2440 | 94 | 2440 |
| jitter multistart LM | 94 | 2166 | 100 | 1806 | 100 | 3786 | 100 | 5636 | 100 | 6592 |
| SG-LM | 89 | 103 | 90 | 103 | 88 | 104 | 83 | 106 | 84 | 117 |
| SG-LM + secant + probe fallback | 92 | 124 | 96 | 118 | 91 | 124 | 91 | 124 | 92 | 128 |
| SG-LM + jitter restarts | 93 | 276 | 95 | 240 | 94 | 253 | 94 | 266 | 94 | 270 |

**Does averaging rescue the classical methods under noise?** Every noisy result above uses one sample per call. Table 17 gives the strongest classical escapes every evaluation as the mean of $q$ independent noisy calls, all $q$ counted, finite-difference steps re-adapted to the effective noise $\sigma/\sqrt{q}$, same $1000n$ budget (additive family, $n$=24, $s$=0.1%, 100 targets). DFO-LS with restarts recovers from 3% to 51% at $q$=4 and 71% at $q$=16, at 3 000 and 12 000 mean calls, 12–47× SG-LM's 258

Table 17: **Noise averaging at equal budget.** Additive $K$-Voigt, $n$=24, evaluation noise $s$=0.1%: each evaluation is the mean of $q$ noisy calls, all counted; success within 1.05× and 2× the reference, pair separated (*split*), mean true calls per solve. SG-LM rows for reference ($q$=1).

| method | $q$ | 1.05× | 2× | split | mean calls |
|---|---|---|---|---|---|
| DFO-LS + restarts | 1 | 3 | 70 | 89 | 634 |
| DFO-LS + restarts | 4 | 51 | 83 | 92 | 3023 |
| DFO-LS + restarts | 16 | 71 | 82 | 85 | 12144 |
| jitter multistart LM | 1 | 0 | 6 | 76 | 1644 |
| jitter multistart LM | 4 | 0 | 6 | 83 | 7746 |
| jitter multistart LM | 16 | 0 | 14 | 85 | 22188 |
| theory-aware FD probe ($2p^2$) | 1 | 0 | 5 | 5 | 217 |
| theory-aware FD probe ($2p^2$) | 4 | 0 | 5 | 0 | 842 |
| theory-aware FD probe ($2p^2$) | 16 | 0 | 5 | 2 | 3180 |
| SG-LM + jitter restarts | 1 | 88 | 92 | 95 | 258 |
| SG-LM + secant + $\chi^2$ restarts | 1 | 71 | 97 | 98 | 143 |

Table 18: Success % from the split point against the gradient tolerance the reduced fit was converged to (`gtol`); $\|g\|_\infty$ is the median $\|J^\top r\|_\infty$ actually reached at the split point.

| `gtol` | biexponential, $n$=4 (35 targets) | | | | | two-Gaussian, $n$=6 (43 targets) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\|g\|_\infty$ | FD-LM | scipy-TRF | NL2SOL | DFO-LS | $\|g\|_\infty$ | FD-LM | scipy-TRF | NL2SOL | DFO-LS |
| $10^{-1}$ | $4\times10^{-3}$ | 57 | 100 | 51 | 100 | $2\times10^{-2}$ | 67 | 95 | 91 | 95 |
| $10^{-2}$ | $2\times10^{-3}$ | 57 | 100 | 51 | 100 | $1\times10^{-3}$ | 56 | 95 | 91 | 95 |
| $10^{-3}$ | $3\times10^{-4}$ | 57 | 100 | 46 | 100 | $2\times10^{-4}$ | 35 | 93 | 93 | 95 |
| $10^{-4}$ | $3\times10^{-5}$ | 57 | 100 | 46 | 100 | $2\times10^{-5}$ | 9 | 84 | 91 | 93 |
| $10^{-5}$ | $2\times10^{-6}$ | 46 | 100 | 46 | 100 | $1\times10^{-6}$ | 5 | 56 | 86 | 95 |
| $10^{-6}$ | $2\times10^{-7}$ | 34 | 86 | 40 | 100 | $2\times10^{-7}$ | 2 | 5 | 47 | 93 |
| $10^{-8}$ | $2\times10^{-8}$ | 0 | 0 | 0 | 100 | $4\times10^{-9}$ | 2 | 2 | 0 | 81 |
| $10^{-10}$ | $1\times10^{-8}$ | 0 | 0 | 0 | 100 | $4\times10^{-9}$ | 2 | 2 | 0 | 81 |
| $10^{-15}$ | $1\times10^{-8}$ | 0 | 0 | 0 | 100 | $4\times10^{-9}$ | 2 | 2 | 0 | 81 |

calls for 88% (an empirical comparison of sampling policies, not a test of Prop. 2, which concerns a central-difference estimator, not DFO-LS); jitter multistart separates the pair on 76% ($q$=1) to 83–85% ($q$=4, 16) of targets but its finite-difference LM never refines to 1.05×, and the $2p^2$ probe has 0% strict success at every $q$. Averaging is a legitimate noise-aware policy and it narrows the reliability gap; the tested $q$=4 and 16 policies remain less accurate and more evaluation-intensive than SG-LM in this setting (one family, one $n$, one noise level; the $q$=16 mean of 12 144 calls is below the 24 000 cap, so other stopping or sampling policies are not excluded). The averaging wrapper passed DFO-LS the remaining budget as a limit on objective calls, each of which spends $q$ raw calls, so ten of the 71 successes at $q$=16 finished between 25 000 and 35 500 raw calls, beyond the nominal cap; the reported means include them at the calls spent, and 61% is the success rate within the cap.

**How converged must the reduced fit be?** The theorem concerns the exact critical point; in practice the reduced fit is converged to a tolerance, and the split point sits off stationarity by that much (the duplicated shapes are still equal; the point is on the coalescence manifold but not at its critical point). Table 18 converges the reduced fit to scipy gradient tolerance `gtol`$\in \{10^{-1}, \ldots, 10^{-15}\}$ (`xtol`=`ftol`= $10^{-15}$), splits, and starts each method from there, interior targets only, in double precision. The accidental escapes of a loosely converged start disappear between $\|J^\top r\|_\infty \sim 10^{-6}$ and $10^{-7}$: scipy-TRF, whose one-sided differences are the most escape-prone, falls from 100% to 0–2%, FD-LM from 57–67% to 0–2%, the NL2SOL secant from 51–91% to 0%, while DFO-LS, whose model is a finite secant, is unaffected. Any production solver run to its default tolerance on the reduced fit (scipy: $10^{-8}$) lands in the blind regime; the original float32/grid table (Table 8) was the loose regime.

**Jacobian-vs-curvature ablation, in full.** Tables 20–22 give every cell of the grid behind Table 1: both families at every $n$, and the additive family at 0.1% and 0.3% evaluation noise, without restarts and with the family's headline restart policy. The column *jac%* is the share of each method's true

Table 19: **A non-neural component surrogate.** SG-LM with the GELU coordinate network and with a Gaussian RBF interpolant of the same 4 000 spectra (length-scale 3.5 spacings), additive $K$-Voigt family, success % / mean true evaluations; 100 targets per $n$ except $n$=32 (21 targets, both surrogates on the same ones).

| surrogate | method | $n$=8 | | $n$=12 | | $n$=16 | | $n$=32 | |
|---|---|---|---|---|---|---|---|---|---|
| | | % | mean | % | mean | % | mean | % | mean |
| GELU network | SG-LM, single attempt | 84 | 105 | 90 | 100 | 80 | 104 | 86 | 116 |
| GELU network | SG-LM + $\chi^2$ restarts | 94 | 195 | 97 | 148 | 97 | 177 | 90 | 122 |
| GELU network | surrogate-Jacobian LM, no curvature | 0 | 25 | 0 | 28 | 0 | 27 | 0 | 30 |
| Gaussian RBF | SG-LM, single attempt | 79 | 110 | 79 | 108 | 69 | 115 | 52 | 172 |
| Gaussian RBF | SG-LM + $\chi^2$ restarts | 93 | 366 | 91 | 173 | 89 | 205 | 62 | 209 |
| Gaussian RBF | SG-LM + jitter restarts | 82 | 358 | 85 | 320 | – | – | – | – |
| Gaussian RBF | surrogate-Jacobian LM, no curvature | 0 | 26 | 0 | 24 | 0 | 27 | 0 | 26 |

evaluations spent in finite-difference Jacobians. The *FD Hessian* escape with the surrogate Jacobian (row 7) is the method that would suit a practitioner who does not know which pair coalesced; its cost is the $2n^2+2n$ stencil, which is what the surrogate curvature saves. Under noise the FD-Jacobian rows fail with every curvature source, the probe rows fail without restarts, and only the surrogate curvature escapes on the first attempt. The *surrogate, none* row of the restart blocks is the symmetry-breaking control (run E): surrogate-Jacobian LM with the family's restart policy and no curvature step at all, so the first attempt sits on the saddle (0–1% success) and nearly every escape comes from a restart's jittered pair. The last block of each table is the single-perturbation control: the split pair's shape coordinates are moved once by 2% (or 5%) of the box half-width in a random direction—the first point of the jitter-restart schedule—and surrogate-Jacobian LM is run once with no eigenstep and no restart; SG-LM from the same point is the matched curvature row. Table 19 replaces the network by a Gaussian radial-basis interpolant of the same 4 000 single-peak spectra (`code/rbf_surrogate.py`: 3 000 centres from the line core and 800 from the tails, per-dimension length-scales 3.5× the median centre spacing, ridge least squares on $1.6 \times 10^5$ of the $7.9 \times 10^5$ sampled values; in-sample MSE $2.5 \times 10^{-8}$ against the network's $7.6 \times 10^{-8}$). At the $K$=3 split points its value error is 0.13% (network 0.43%), Jacobian error 1.4% (2.6%), $C$ error 5.8% (3.5%), escape-eigenvalue error 3.0% (2.4%), alignment 1.00; the same interpolant with length-scales 2.2× the spacing has value error 0.35%, better than the network, and $C$ error 70%, alignment 0.96. The RBF is slower per solve (its Hessian is a dense kernel sum), so the jitter-restart variant was run at $n$=8 and 12 only (an $n$=16 run stopped after 9 targets and is not shown), and $n$=32 on the first 21 targets (the run was cut at submission; the network rows at $n$=32 are restricted to the same 21). One accounting caveat for the finite-difference curvature baselines of Tables 11–14 and 20–22: when their fourth and last eigenstep is accepted, the logged final value is the one before that step (47 of 500 additive and 59 of 500 transmission targets for the surrogate-Jacobian + FD-Hessian arm; fewer elsewhere), so their recorded finals are upper bounds and their success rates lower bounds; the same held for SG-LM in the runs behind Tables 11–16, whose solver then allowed a fourth eigenstep.

### F.5 An external simulator: unresolved spectroscopic binaries with Korg

**Why this problem.** A test of practical significance and of generalisation beyond the paper's own forward models needs an amplitude-linear component structure with $p \geq 2$, a forward model that is external and expensive in the operational sense (seconds per call, no derivatives used), a natural nested-model initialisation of the richer model from the reduced fit, and enough targets to amortise training. Unresolved spectroscopic binaries supply all four: El-Badry et al. (2018) model the normalised spectrum of an unresolved binary as the light-weighted sum of two single-star spectra, note that a $q$=1 binary at zero velocity offset is indistinguishable from a single star, and initialise about ten optimisers over the mass ratio for that reason. Our problem is a controlled, relaxed version of that setting, not their pipeline: two stars with independent labels and free light weights, synthetic targets, no isochrone or continuum model.

**Setup.** Component $\varphi(\theta)$ is Korg's (Wheeler et al., 2023, 2024) rectified spectrum (APOGEE DR17 line list, MARCS atmospheres) on an APOGEE-like log-wavelength grid ($6 \times 10^{-6}$ dex, 4.14 km/s per pixel) over 15 700–16 000 Å with a 150 km/s margin, at $R$=22 500; the analysis win-

Table 20: Jacobian-vs-curvature ablation, additive $K$-Voigt family, 100 targets per $n$. Success % / mean evaluations / share of evaluations in FD Jacobians.

| Jac. | escape | $n$=8 | | | $n$=12 | | | $n$=16 | | | $n$=24 | | | $n$=32 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | % | mean | jac% | % | mean | jac% | % | mean | jac% | % | mean | jac% | % | mean | jac% |
| *no restarts* | | | | | | | | | | | | | | | | |
| FD | none | 0 | 814 | 88 | 0 | 795 | 91 | 0 | 1356 | 93 | 0 | 2220 | 95 | 0 | 2141 | 96 |
| FD | $2p^2$ probe | 95 | 1423 | 88 | 96 | 1557 | 91 | 94 | 2459 | 93 | 91 | 4245 | 96 | 94 | 4095 | 96 |
| FD | FD Hessian | 96 | 1715 | 76 | 99 | 2184 | 68 | 95 | 3481 | 67 | 95 | 7020 | 64 | 95 | 8273 | 50 |
| FD | surr. | 94 | 1304 | 89 | 94 | 1406 | 92 | 87 | 2281 | 94 | 89 | 3792 | 96 | 90 | 4252 | 97 |
| surr. | none | 0 | 25 | 0 | 0 | 28 | 0 | 0 | 27 | 0 | 0 | 26 | 0 | 0 | 28 | 0 |
| surr. | $2p^2$ probe | 76 | 112 | 0 | 84 | 111 | 0 | 69 | 112 | 0 | 74 | 114 | 0 | 72 | 113 | 0 |
| surr. | FD Hessian | 84 | 445 | 8 | 93 | 823 | 7 | 84 | 1417 | 5 | 87 | 2960 | 4 | 86 | 5146 | 3 |
| surr. | surr. (SG-LM) | 84 | 105 | 0 | 90 | 100 | 0 | 80 | 104 | 0 | 85 | 105 | 0 | 85 | 115 | 0 |
| *with $\chi^2$ restarts* | | | | | | | | | | | | | | | | |
| FD | surr. | 97 | 1617 | 89 | 100 | 1645 | 92 | 98 | 3138 | 94 | 99 | 4575 | 96 | 98 | 4701 | 97 |
| surr. | none | 93 | 149 | 0 | 96 | 141 | 0 | 93 | 129 | 0 | 90 | 124 | 0 | 78 | 100 | 0 |
| surr. | $2p^2$ probe | 93 | 167 | 0 | 98 | 172 | 0 | 95 | 171 | 0 | 97 | 167 | 0 | 87 | 141 | 0 |
| surr. | FD Hessian | 93 | 759 | 8 | 98 | 1102 | 7 | 96 | 1997 | 5 | 99 | 3743 | 4 | 96 | 6342 | 3 |
| surr. | surr. (SG-LM) | 94 | 195 | 0 | 97 | 148 | 0 | 97 | 177 | 0 | 99 | 166 | 0 | 96 | 151 | 0 |
| *single perturbed split, no restarts* | | | | | | | | | | | | | | | | |
| surr. | none (2% split perturbation) | 63 | 64 | 0 | 81 | 60 | 0 | 66 | 62 | 0 | 74 | 73 | 0 | 73 | 68 | 0 |
| surr. | none (5%) | 71 | 68 | 0 | 83 | 62 | 0 | 74 | 66 | 0 | 88 | 67 | 0 | 78 | 60 | 0 |
| surr. | surr. (SG-LM) (2%) | 86 | 97 | 0 | 92 | 83 | 0 | 86 | 95 | 0 | 92 | 110 | 0 | 91 | 104 | 0 |

Table 21: The same grid on the transmission family, jitter restarts in the lower block.

| Jac. | escape | $n$=8 | | | $n$=12 | | | $n$=16 | | | $n$=24 | | | $n$=32 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | % | mean | jac% | % | mean | jac% | % | mean | jac% | % | mean | jac% | % | mean | jac% |
| *no restarts* | | | | | | | | | | | | | | | | |
| FD | none | 0 | 522 | 86 | 0 | 420 | 90 | 0 | 822 | 93 | 0 | 1165 | 95 | 1 | 1118 | 96 |
| FD | $2p^2$ probe | 85 | 1155 | 87 | 95 | 1489 | 91 | 83 | 2284 | 93 | 86 | 2981 | 95 | 89 | 3932 | 97 |
| FD | FD Hessian | 91 | 1497 | 74 | 99 | 2107 | 67 | 95 | 3923 | 70 | 95 | 5873 | 59 | 98 | 9036 | 53 |
| FD | surr. | 88 | 1110 | 88 | 97 | 1433 | 92 | 89 | 2572 | 94 | 89 | 2968 | 96 | 94 | 4372 | 97 |
| surr. | none | 0 | 24 | 0 | 0 | 27 | 0 | 0 | 25 | 0 | 0 | 26 | 0 | 1 | 26 | 0 |
| surr. | $2p^2$ probe | 75 | 104 | 0 | 81 | 116 | 0 | 72 | 107 | 0 | 75 | 111 | 0 | 77 | 100 | 0 |
| surr. | FD Hessian | 78 | 439 | 8 | 91 | 850 | 7 | 83 | 1472 | 5 | 87 | 2990 | 4 | 89 | 5395 | 3 |
| surr. | surr. (SG-LM) | 79 | 99 | 0 | 89 | 118 | 0 | 82 | 118 | 0 | 84 | 113 | 0 | 85 | 111 | 0 |
| *with jitter restarts* | | | | | | | | | | | | | | | | |
| FD | surr. | 91 | 2994 | 90 | 100 | 3451 | 93 | 100 | 5879 | 95 | 99 | 7223 | 96 | 100 | 9545 | 97 |
| surr. | none | 80 | 226 | 0 | 85 | 227 | 0 | 76 | 216 | 0 | 85 | 228 | 0 | 97 | 228 | 0 |
| surr. | $2p^2$ probe | 78 | 243 | 0 | 83 | 252 | 0 | 74 | 236 | 0 | 80 | 248 | 0 | 94 | 239 | 0 |
| surr. | FD Hessian | 89 | 1207 | 8 | 96 | 2086 | 6 | 97 | 3682 | 5 | 99 | 7190 | 4 | 100 | 12132 | 3 |
| surr. | surr. (SG-LM) | 85 | 297 | 0 | 93 | 323 | 0 | 94 | 332 | 0 | 99 | 362 | 0 | 100 | 283 | 0 |
| *single perturbed split, no restarts* | | | | | | | | | | | | | | | | |
| surr. | none (2% split perturbation) | 68 | 69 | 0 | 75 | 69 | 0 | 71 | 61 | 0 | 75 | 69 | 0 | 79 | 57 | 0 |
| surr. | none (5%) | 73 | 62 | 0 | 80 | 66 | 0 | 76 | 65 | 0 | 78 | 64 | 0 | 86 | 54 | 0 |
| surr. | surr. (SG-LM) (2%) | 79 | 101 | 0 | 89 | 101 | 0 | 89 | 104 | 0 | 87 | 105 | 0 | 96 | 87 | 0 |

Table 22: The same grid under evaluation noise (additive family, $K = 3$ and 6, $s = 0.1\%$ and $0.3\%$; the surrogate is trained on equally noisy values; scored on the noiseless misfit).

| Jac. | escape | $n$=12, $s$=0.1% | | $n$=12, $s$=0.3% | | $n$=24, $s$=0.1% | | $n$=24, $s$=0.3% | |
|---|---|---|---|---|---|---|---|---|---|
| | | % | mean | % | mean | % | mean | % | mean |
| *no restarts* | | | | | | | | | |
| FD | none | 3 | 125 | 0 | 109 | 0 | 175 | 0 | 171 |
| FD | $2p^2$ probe | 9 | 200 | 0 | 122 | 0 | 217 | 0 | 231 |
| FD | FD Hessian | 20 | 935 | 0 | 613 | 0 | 1924 | 0 | 2467 |
| FD | surr. | 41 | 513 | 1 | 410 | 0 | 697 | 0 | 626 |
| surr. | none | 0 | 28 | 0 | 28 | 0 | 28 | 0 | 28 |
| surr. | $2p^2$ probe | 33 | 74 | 3 | 56 | 8 | 58 | 17 | 67 |
| surr. | FD Hessian | 46 | 608 | 3 | 469 | 10 | 1696 | 18 | 1871 |
| surr. | surr. (SG-LM) | 73 | 87 | 46 | 83 | 73 | 95 | 49 | 84 |
| *with $\chi^2$ restarts* | | | | | | | | | |
| FD | surr. | 53 | 956 | 3 | 3467 | 0 | 9870 | 0 | 8694 |
| surr. | none | 83 | 129 | 58 | 141 | 85 | 100 | 66 | 98 |
| surr. | $2p^2$ probe | 88 | 153 | 62 | 168 | 85 | 126 | 65 | 128 |
| surr. | FD Hessian | 87 | 1227 | 58 | 1501 | 85 | 3595 | 71 | 3667 |
| surr. | surr. (SG-LM) | 87 | 139 | 66 | 158 | 91 | 125 | 67 | 130 |
| *single perturbed split, no restarts* | | | | | | | | | |
| surr. | none (2% split perturbation) | 58 | 44 | 34 | 38 | 49 | 45 | 42 | 43 |
| surr. | none (5%) | 66 | 47 | 38 | 42 | 72 | 48 | 55 | 46 |
| surr. | surr. (SG-LM) (2%) | 75 | 63 | 51 | 68 | 70 | 70 | 59 | 68 |

dow after cropping the margin is 15 698–16 003 Å, $m$=1392 pixels. Box: $T_{\rm eff} \in [4000, 7000]$ K, $\log g \in [3.8, 5.0]$, $[\mathrm{M/H}] \in [-1, 0.5]$, $v_{\rm rad} \in [-100, 100]$ km/s, light weights $a \in [0, 1.2]$, all parameters box-normalised to $[0, 1]$ so that the eigenstep and jitter conventions of §E.1 apply unchanged (unlike the analytic families, where $\hat{H}$ and the step are in physical coordinates, here both are formed and taken in the normalised coordinates). Radial velocity is an exact Fourier phase shift of the rest-frame spectrum on the log grid (differentiable, so the assembled Jacobian and curvature include the velocity rows); the component surrogate is a $256\times4$ GELU network from $(T_{\rm eff}, \log g, [\mathrm{M/H}])$ to the rest-frame spectrum, trained on values at 2 000 Latin-hypercube points (held-out error $0.045\%$ of the spectrum, $1\%$ of the spread across the box). Targets: block A (60), near-twins with label offsets $\leq 5\%$ of the box, light ratio 0.35–0.65, $|\Delta v| \in [3, 12]$ km/s (below the 13 km/s resolution element; above it the single-star fit absorbs the doubled lines into pressure broadening and $68\%$ of reduced fits end on the $\log g$ bound; below 3 km/s, in our pilot, the split point already met the $1.05\times$ criterion at this noise); block B (40), $\Delta v = 0$ with 5–$15\%$ offsets. Observation noise $0.5\%$ of the continuum. Initialisation is the single-star fit of the true model (trust-region least squares with a finite-difference Jacobian, uncounted), split symmetrically; that fit is numerically, not exactly, stationary: its recorded gradient $\infty$-norm in box units is $3 \times 10^{-5}$ to $1.4 \times 10^{-2}$ over the 41 interior non-trivial targets (median $8 \times 10^{-4}$), so the split points are approximate critical points, which FD-LM nevertheless fails to leave on 48 of 52. The reference is a fit from the true labels, and success is $1.05\times$ the pooled best. Targets whose split point is already within $1.05\times$ of the best are excluded as trivial: 9 of block A, at the smallest offsets, and 39 of the 40 of block B—with $\Delta v = 0$, label offsets of 5–$15\%$ of the box leave the single-star fit within 0.1–$7\%$ of the two-star optimum at this noise, so that block is uninformative at the $1.05\times$ criterion and the atmospheric-only escape is untested at a tolerance that would resolve it. Methods and budgets are those of §5 (1500 assembled-model evaluations, each costing zero, one or two new syntheses under the cache, so a synthesis count above 1500 is not a budget violation), with a ten-start light-ratio/velocity multistart (FD-LM from the split point with the light ratio and velocity offset re-drawn per start) as an analogue of the survey's practice—their $q$ is a mass ratio that also moves the secondary's labels through an isochrone, so this is not their optimiser; the $\chi^2$ floor uses the known noise. The initial single-star fit is excluded for every method (its logged synthesis counter accumulated across a worker's targets and is not reported), so all savings are incremental from a supplied reduced fit. Every rest-frame synthesis is counted through a component cache cleared before each method (a finite difference in one component does not re-synthesise the other; a velocity move is free), and wall-clock is measured per solve, including the surrogate's autodiff and the Julia server round trip; 2.3 s median per synthesis on the cluster used.

**Curvature against Korg's finite differences.** Before the main run, 100 reduced fits of a pilot cohort (50 near-twins, 50 distinct pairs) were re-evaluated with fresh syntheses at $\theta^*$ and a $2p^2$ stencil (1 900 syntheses). The true split matrix $M = \sum_i r_i^* \nabla^2 \varphi_i$ (box units) has a negative eigenvalue on $100\%$; the surrogate's $M$ agrees in sign on $100\%$ and within $2.1\%$ in norm (median; $13\%$ at the 90th percentile), escape eigenvector aligned 1.00. Block-wise: the atmospheric $3\times3$ block's lowest eigenvalue is reproduced to 1–$2\%$ with direction aligned 0.99 and sign right on 98–$100\%$, but that block's remaining entries are small and its Frobenius error is $40\%$; the velocity entry is within $2\%$; the mixed block within $4\%$ on distinct pairs and $38\%$ on twins. The velocity curvature is 150–$300\times$ the atmospheric one, which is what makes the full-matrix figure flattering and the block-wise one the honest statement.

**Results.** Table 23 gives every method on the 52 non-trivial targets and by interior (41) and at-bound (11, all $\log g$ at its upper limit) reduced fits; Fig. 12 the cumulative wall-clock against targets solved with the surrogate's offline cost on the plot. FD-LM stops at the split point ($8\%$, the trivial-adjacent cases). The $2p^2$ probe and single-attempt SG-LM fail on the same eleven shallowest saddles ($|\Delta v| \leq 5.1$ km/s, trap depth 5–$18\%$ of the loss): the direction is right but the true line search finds no decrease at the $\alpha$ grid of §E.1; jittered restarts solve all of them (four starts each). The no-curvature control, surrogate-Jacobian LM with the same $\chi^2$ restarts, reaches $35\%$ (18 of 52): its first attempt succeeds on 3 targets against 41 for SG-LM's single attempt—the two differ only in the curvature step; the three are three of FD-LM's four successes and include the two least-converged reduced fits (gradient norm 0.014), so an imperfect reduced fit, not the surrogate, is the likely cause, though the mechanism is not established—and the $\chi^2$ test fires on 15, all of which the restarts then solve (it fires exactly where the split point's residual exceeds twice the noise floor, trap depth 2.3–$11\times$ the floor; on the other 37 the split point is within $2\times$ the floor and the test cannot

Table 23: Spectroscopic binaries, every method on the non-trivial targets and split by whether the single-star fit is interior or on a bound: success within $1.05\times$, pair separated, median syntheses, median measured minutes per solve.

| method | non-trivial targets | | | | interior reduced fit | | | | reduced fit at a bound | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | % | split | synth. | min | % | split | synth. | min | % | split | synth. | min |
| (min = median wall-clock minutes per solve) | | | | | | | | | | | | |
| FD-LM | 8 | 8 | 30 | 1.5 | 10 | 10 | 33 | 1.6 | 0 | 0 | 24 | 0.9 |
| theory-aware FD probe ($2p^2$) | 79 | 79 | 251 | 9.5 | 73 | 73 | 244 | 8.7 | 100 | 100 | 262 | 9.7 |
| jitter multistart LM | 100 | 100 | 546 | 21.3 | 100 | 100 | 585 | 22.1 | 100 | 100 | 530 | 18.7 |
| light-ratio/velocity multistart (10 starts) | 100 | 100 | 513 | 20.3 | 100 | 100 | 523 | 21.4 | 100 | 100 | 473 | 17.6 |
| DFO-LS + restarts | 100 | 100 | 1450 | 57.6 | 100 | 100 | 1565 | 58.3 | 100 | 100 | 1248 | 47.6 |
| SG-LM (single attempt) | 79 | 79 | 54 | 2.2 | 73 | 73 | 51 | 2.2 | 100 | 100 | 63 | 2.3 |
| SG-LM + $\chi^2$ restarts | 79 | 79 | 54 | 2.2 | 73 | 73 | 51 | 2.2 | 100 | 100 | 63 | 2.3 |
| SG-LM + jitter restarts | 100 | 100 | 154 | 6.1 | 100 | 100 | 153 | 6.0 | 100 | 100 | 162 | 6.3 |
| surrogate-Jacobian LM + $\chi^2$ (no curvature) | 35 | 35 | 12 | 0.6 | 29 | 29 | 10 | 0.6 | 55 | 55 | 59 | 2.6 |
| surrogate-Jacobian LM, one $2\%$ perturbation (no curvature) | 100 | 100 | 53 | 2.0 | 100 | 100 | 52 | 2.0 | 100 | 100 | 61 | 2.1 |
| surrogate-Jacobian LM + jitter restarts (no curvature) | 100 | 100 | 146 | 5.5 | 100 | 100 | 146 | 5.4 | 100 | 100 | 147 | 5.5 |
| SG-LM from the same $2\%$ perturbation | 100 | 100 | 53 | 2.0 | 100 | 100 | 52 | 2.0 | 100 | 100 | 61 | 2.2 |

fire). Its matched jitter-restart variant reaches $52$ of $52$ at $146$ syntheses (SG-LM $154$), and from a single $2\%$ perturbation of the split pair, with no restart and no eigenstep, surrogate-Jacobian LM reaches $52$ of $52$ at $53$ syntheses and $2.0$ min—identical, target by target, to SG-LM from that start, whose eigenstep never fires on the $52$ (it fires on $13$ of the trivial targets), and fewer than SG-LM's $54$ from the exact point. On this problem the curvature step was therefore needed only from the exact split point: for the tested $2\%$ perturbation the surrogate-Jacobian solver reached the criterion on all $52$ targets without it (once the pair is split, the gradient of $\Phi$ along the split is nonzero at first order in the perturbation, and LM descends along it). For SG-LM the $\chi^2$ policy never fires: its failures leave residuals within $15\%$ of the noise floor ($\chi^2/\mathrm{dof} \approx 1.0$–$1.15$ at $1392$ degrees of freedom), statistically acceptable, so the jitter policy is the one reported, as for the transmission family. Offline cost: $2\,000$ syntheses ($1.93$ h on one core, summed measured durations; the stage-1 mean of $3.5$ s per synthesis exceeds the solve-time median of $2.2$ s) plus $510$ s of training, $T_{\mathrm{off}} = 2.07$ h in all (training-set generation and fitting; the $1\,900$ pilot-verification syntheses are development, not deployment, cost). The wall-clock break-even is $N_* = T_{\mathrm{off}}/(\bar{T}_{\mathrm{base}} - \bar{T}_{\mathrm{SG}})$ with *mean* measured minutes per attempted target (means, not the medians of the tables, because synthesis time varies with the stellar labels): $\bar{T}_{\mathrm{SG}} = 8.4$, the two multistarts $25.7$ and $28.2$, DFO-LS $64.6$, giving $7.2$, $6.3$ and $2.2$ targets (Fig. 12); counted in syntheses, $2\,000/(513 - 154) \approx 6$. The surrogate's autodiff is $1.8$ s per solve against $108$ s of synthesis.

**What this does and does not show.** It shows the coalescence saddle in an external synthesis code at the point a practitioner's reduced fit produces, a value-trained surrogate whose escape curvature matches the code's own finite differences, and a 3–$9\times$ saving in syntheses and wall-clock over the classical escapes and a ten-start multistart at equal observed success, incremental from a supplied reduced fit, with the curvature step accounting for the single-attempt gap between $6\%$ and $79\%$ ($3$ against $41$ of $52$) from the exact point and a $2\%$ perturbed split sufficing without it ($52$ of $52$): the saving is the free Jacobian's. It does not yet attribute the saving to *learned* curvature: the dominant escape direction is the velocity split, whose curvature $M_{vv} = r^{*\top} S''(v) s(\eta^*)$ is available analytically once the reduced fit's rest-frame spectrum $s(\eta^*)$ is cached, so a control with the surrogate Jacobian and an analytic velocity-only eigenstep is the minimal attribution test, and a tied-velocity cohort (one shared $v$, escape only through the stellar labels) the test of the learned atmospheric curvature; neither was run before submission, nor were an RBF-driven SG-LM, a surrogate-only fit with true validation, or an LM on Korg's native automatic-differentiation Jacobian. The model is relaxed relative to the survey's (free light weights rather than continuum-weighted fluxes, no isochrone, no rotational broadening), the targets are synthetic, and MARCS interpolation makes the true map only piecewise smooth across grid cells.




Figure 12: Mean-cost lines, not a trace of ordered runs: cumulative wall-clock on one core against targets attempted, at the mean measured seconds per attempt on the non-trivial targets (failed attempts included); SG-LM's intercept is the surrogate's offline cost, training-set generation plus fitting. Success rates in the legend.

## G NETWORK DESIGN FOR CURVATURE FIDELITY

Given §4.1, the design question is concrete: build a network whose *second* derivative is trustworthy, accepting that this inverts the usual priorities of function approximation. This makes a handful of neural-network choices load-bearing.

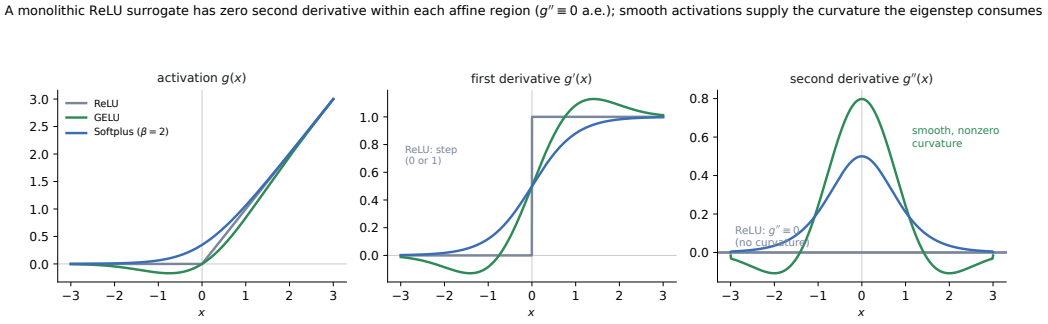


Figure 13: The activation determines whether a *monolithic* surrogate, piecewise affine in the differentiated coordinates, carries usable curvature: a ReLU network's second derivative is zero almost everywhere; smooth activations (GELU, Softplus) have a continuous, nonzero second derivative—the object the negative-curvature eigenstep consumes. For component-composed or feature-transformed networks see Rem. 9.

**The activation decides whether a monolithic surrogate has curvature at all.** An MLP with ReLU activations is a *piecewise-linear* function of its input (Nair & Hinton, 2010; Glorot et al., 2011; Montúfar et al., 2014): for a monolithic surrogate $\hat{f}$ of the map, on each linear region $\nabla^2 \hat{f} \equiv 0$, so the Newton correction $C = \sum_i r_i \nabla^2 \hat{f}_i$ vanishes identically and the surrogate Hessian collapses to $\hat{H} = \hat{J}^\top \hat{J} \succeq 0$ (Fig. 13, right); the statement is about that construction only (Rem. 9).

**Remark 9** (A monolithic ReLU surrogate carries no curvature)**.** *For a monolithic surrogate $\hat{f}$ that is piecewise affine in the differentiated coordinates, $\nabla^2 \hat{f} \equiv 0$ away from the kinks, so $\hat{H} = \hat{J}^\top \hat{J} \succeq 0$ and the eigenstep of the* SG-LM *box (Fig. 1) never fires:* SG-LM *reduces to plain LM, a structural deficiency rather than one of fidelity. We claim no such statement for the component surrogate $\hat{f} = \sum_k a_k \hat{\varphi}(\theta_k)$: with a ReLU $\hat{\varphi}$ its shape block $\partial^2_{\theta\theta} \hat{\varphi}$ vanishes, but the mixed amplitude–shape entries $r^{*\top} \partial_\theta \hat{\varphi} = r^{*\top}(\partial_\theta \hat{\varphi} - \partial_\theta \varphi)$ are O(Jacobian error) and generally nonzero, can give $\hat{H}$*

*a negative eigenvalue, and that direction can have negative true curvature; a network applied to nonlinear features ($\log\sigma$, $\log\gamma$ in our coordinate network) or composed with a nonlinear outer map (the transmission family) is not piecewise affine in the physical parameters at all. Smooth activations are therefore a design choice for faithful component curvature, supported by Table 4, not an impossibility result.*

Empirically this is exactly what we observe: a ReLU-surrogate variant of SG-LM (monolithic biexponential surrogate, the setting of Table 4) reduces exactly to plain LM—the eigenstep never fires. The fidelity sweep confirms it quantitatively (Table 4): ReLU's relative curvature error is exactly $1.00$—it captures *none* of the true curvature—even though its forward MSE ($5.9 \times 10^{-3}$) is competitive with the smooth variants, the sharpest illustration that value fit does not imply curvature. Smooth activations—GELU (Hendrycks & Gimpel, 2016), Softplus (Dugas et al., 2001), $\tanh$—have continuous nonzero second derivatives and do supply an escape direction.

**Why GELU, not merely "smooth".** Two requirements pull against each other: an activation smooth enough to yield trustworthy curvature, yet well-conditioned enough to train. The first is settled by the fidelity sweep already read off in §4.1 (Table 4, Fig. 7); one further in-family datum belongs here, because it is about the activation rather than about training objectives: within the softplus family the sharp $\beta$=20, which tends to ReLU as $\beta \to \infty$, has the best in-family MSE and the worst curvature ($0.91$). Sharpening buys value fit by spending exactly the smoothness the escape needs: as $\beta$ increases, $\nabla^2 \hat{f}$ becomes less smooth and closer to piecewise-constant (like ReLU's zero Hessian), so while MSE improves the curvature fidelity deteriorates. In this sweep the trade-off is direct: the activations with the best value fit are those with the weakest second derivatives, the object the eigenstep needs. Among smooth choices, $\tanh$ saturates and its second derivative decays in the tails; sharp Softplus recovers ReLU's trainability but loses curvature—and requires choosing $\beta$ to navigate the trade-off; GELU is smooth ($C^\infty$), non-saturating, and has no tunable scale parameter, so it sidesteps the sharpening dilemma entirely. GELU gave the best second-order fidelity in our multi-variant sweep, which is why it is our default. This mirrors practice in physics-informed networks and implicit neural representations, where smooth activations ($\tanh$, sinusoids) are chosen precisely because the loss consumes high-order derivatives (Raissi et al., 2019; Sitzmann et al., 2020; Rahaman et al., 2019).

**Trainability and numerical stability are the price of smoothness.** ReLU is the easiest activation to train—its gradient is exactly $0$ or $1$, it never saturates, and it cannot overflow—which is precisely why it is the default and why abandoning it is not free. Smooth activations reintroduce numerical fragility. Our Softplus surrogates initially produced `NaN` loss at the first epoch: the textbook form $\frac{1}{\beta}\log(1 + e^{\beta x})$ overflows to $\infty$ for $\beta x \gtrsim 88$ in `fp32`; the fix is the numerically stable primitive $\mathrm{softplus}(\beta x)/\beta$ (identical in exact arithmetic). Separately, unbounded shape parameters fed through $e^{-kt}$ overflow when the training box is too wide, which we remove with per-model input normalisation to $[-1, 1]$ and output standardisation. GELU sidesteps both failure modes (bounded-magnitude gradients, no bare exponential), a further practical reason to prefer it.

**Warm-starting: a dispensable workaround, not a contribution.** Because ReLU trains so easily, an appealing idea is to train a ReLU network first and then *swap* its activation for a smooth one and fine-tune—a form of warm-starting / knowledge transfer (Chen et al., 2016; Ash & Adams, 2020). We used exactly this in early experiments (ReLU $\to$ Softplus, then fine-tune at $10\times$ lower learning rate with tighter gradient clipping). It is *not* novel, and it is not free: transplanting ReLU weights into a Softplus network causes exploding gradients (hence the delicate fine-tuning schedule), and warm-started networks are known to generalise differently from freshly initialised ones (Ash & Adams, 2020). Crucially, it proved *unnecessary*: GELU networks train stably from scratch, so our reported surrogates use direct training and no warm start. Since the amortised training cost is already negligible ($N^* \approx 4$ solves, §F.2), the marginal savings warm-starting might offer do not justify its fragility. We report it as an option we tried and set aside, not as an innovation—and we did not find evidence that it improves the final surrogate.

**Relation to derivative-aware learning.** Our surrogate obtains curvature as a *by-product* of value-only training, and smoothness and width reduce its leakage $\beta$ only indirectly; §4.1 sets out why supervising derivatives directly is the route to attacking $\beta$ at its source, and what a derivative-free

second-difference variant would look like. The design point to add here is that the two are compatible: nothing in the SG-LM box (Fig. 1) assumes value-only training—it assumes only that $\nabla^2 \hat{f}$ exists and is cheap—so a Sobolev-trained surrogate is a drop-in replacement whose effect would be to shrink $\beta$, not to change the escape mechanism.