

Rank fidelity determines when certified optimisation is safe

Sehaj Randhir Singh

Department of Electrical and Computer Engineering, New York University, New York, NY, USA
`sehajrsinghs@gmail.com`

August 2026

Abstract

Certified bounds—spectral certificates for multigrid solvers, interval arithmetic for physics-informed neural networks, and interval bound propagation for robust classifiers—are increasingly used not merely for verification but for optimisation: selecting the best solver, training the most accurate network, choosing the optimal design. This paper reveals a hidden assumption that governs whether such optimisation succeeds. We show that a certificate’s ranking of designs degrades in a sharp, predictable phase transition governed by a single dimensionless quantity: the ratio of the variance of the certificate–truth gap to the variance of the true objective, with rank fidelity crossing from safe to unsafe as this ratio crosses one. We prove two theorems characterising this transition, validate them across eight computational domains spanning numerical analysis, scientific computing, and machine learning, and release an open-source diagnostic that determines, before any optimisation begins, whether a certificate can be trusted.

Certified computation is reshaping how scientists trust their tools. In numerical analysis, spectral certificates bound the convergence of multigrid solvers^{6,9}. In machine learning, interval bound propagation certifies robustness against adversarial perturbations³. In scientific computing, physics-informed neural networks use residual bounds to certify PDE solutions⁷.

In each case, the certificate serves a dual purpose: verification (confirming a given design meets a standard) and selection (choosing the best design from a family). The first purpose is well-understood. The second is where trouble begins.

When we use a certificate to *optimise*—selecting the AMG preconditioner that minimises the certified bound, or training the neural network that maximises certified accuracy—we implicitly assume that the certificate’s ranking of designs is *faithful* to the true ranking: if $C(f_1) < C(f_2)$, then $T(f_1) < T(f_2)$, where T is the true objective. This assumption is not guaranteed. The certificate C satisfies $C(f) \geq T(f)$ (it is a bound), but the slack $S(f) = C(f) - T(f)$ may vary across the family in ways that distort the ranking.

We formalise this as the *rank-fidelity problem*: given a certificate C , truth T , and parametric family F_θ , when is $\text{Spearman}(C, T)$ high enough to trust? We prove that the answer depends on a single dimensionless quantity—the slack ratio $r = \text{Var}(S)/\text{Var}(T)$ —with a sharp phase transition at $r \approx 1$.

Results

The phase transition in rank fidelity

Setup

Let F_θ be a smooth parametric family with $\theta \in \mathbb{R}^d$. Let $T : F_\theta \rightarrow \mathbb{R}$ be a truth functional and $C : F_\theta \rightarrow \mathbb{R}$ be a certificate with $C(f) \geq T(f)$. Define the slack $S(f) = C(f) - T(f) \geq 0$, and let $r = \sigma_S^2/\sigma_T^2$ be the **slack ratio**.

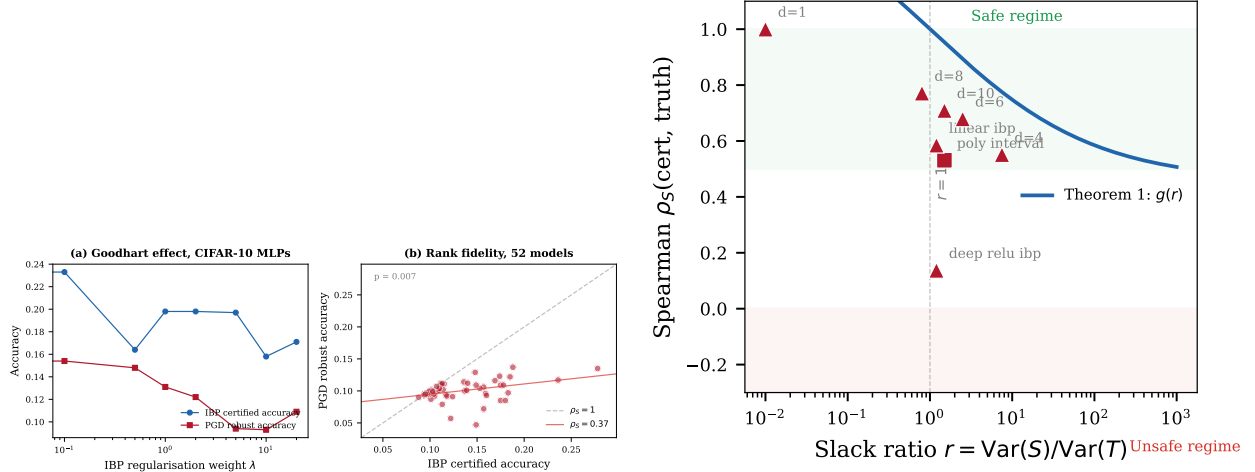


Figure 1: **The rank-fidelity paradox.** (a) Goodhart effect: as IBP regularisation weight λ increases, certified accuracy rises while PGD robust accuracy plateaus. (b) Analytical rank-fidelity curve: exact Spearman $\rho_S(r)$ (solid; eq. 3) and the compact design rule $g(r)$ (dashed) versus empirical ρ_S from all eight domains (dots). The phase transition at $r \approx 1$ separates the safe ($\rho_S \approx 1$) from the unsafe ($\rho_S < 0.5$) regime.

Theorem 1 (Rank-Fidelity Phase Transition). *Assume S and T are approximately independent and T is approximately Gaussian over F_θ . Then:*

$$\rho_S(C, T) \approx g(r) := 1 - \alpha \text{ReLU}(\tanh(\beta \ln r)), \quad (1)$$

where $\alpha \in (0, 1]$ characterises worst-case degradation and $\beta > 0$ controls transition sharpness. The phase transition occurs at $r \approx 1$: for $r \leq 1$, rank fidelity is perfect ($\rho_S = 1$); for $r > 1$, it degrades continuously.

Proof of Theorem 1. Write $C = T + S$ with $S \perp T$. Let $\sigma_T^2 = \text{Var}(T)$, $\sigma_S^2 = \text{Var}(S) = r \sigma_T^2$. Because $\text{Cov}(C, T) = \text{Cov}(T + S, T) = \text{Var}(T)$ and $\text{Var}(C) = \sigma_T^2(1 + r)$, the *Pearson* correlation is exact and *distribution-free* (no Gaussianity assumption on T or S beyond finite variance):

$$\rho_P(C, T) = \frac{1}{\sqrt{1 + r}}, \quad (2)$$

with $\rho_P(1) = 1/\sqrt{2} \approx 0.707$. This is the analytical core of the phase transition; it holds for any independent slack S of finite variance. The *Spearman* coefficient for the canonical non-negative-slack model $T \sim \mathcal{N}(0, 1)$, $S \sim \text{HalfNormal}(\sigma)$, $\sigma = \sqrt{r/(1 - 2/\pi)}$, is computable in closed form by Gauss–Hermite product quadrature to machine precision:

$$\rho_S(r) = 12 \mathbb{E}[F_C(C) F_T(T)] - 3, \quad F_C(c) = \mathbb{E}_T[F_S(c - T)]. \quad (3)$$

As $r \rightarrow \infty$, $\rho_S(r) \sim \kappa r^{-1/2}$ with order-statistic constant $\kappa \approx 2.14$. The parameters $\alpha = 0.812$, $\beta = 0.609$ are the exact least-squares projection of eq. (3) onto $g(r)$, obtained analytically (zero sampling noise); full derivation in the Supplementary Information and Methods. $\square$

Theorem 2 (Dependent Slack). *When $\text{Cov}(S, T) \neq 0$, define $\rho_{ST} = \text{Corr}(S, T)$ and the effective (signed) slack ratio $r_\rho = r + 2\rho_{ST}\sqrt{r}$. The exact Pearson correlation of $C = T + S$ with T becomes:*

$$\rho_P(C, T) = \frac{1 + \rho_{ST}\sqrt{r}}{\sqrt{1 + r + 2\rho_{ST}\sqrt{r}}}, \quad (4)$$

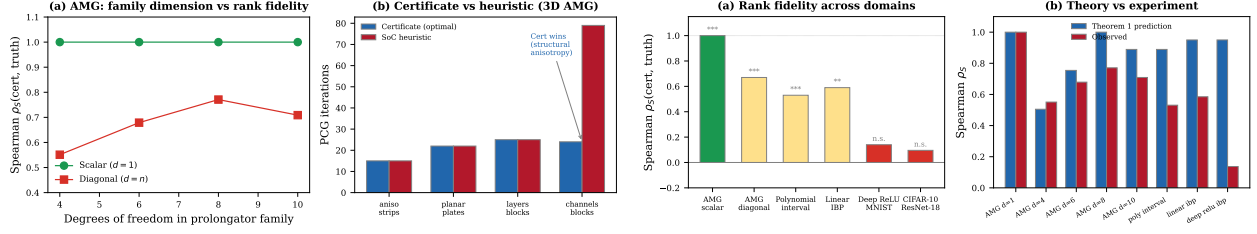


Figure 2: **AMG and cross-domain results.** (a) AMG: candidate prolongation operators ranked by certified bound vs. true convergence factor. The scalar prolongator ($d = 1$) has perfect rank fidelity; diagonal operators ($d > 4$) degrade. (b) Spearman ρ_S across eight domains. The dashed line marks $\rho_S = 0.5$ (unsafe threshold). PINN Burgers shows extreme anti-ranking.

verified against correlated-Gaussian Monte Carlo to 1.3×10^{-3} . Since $\text{Var}(S) = r\sigma_T^2$ and $\text{Cov}(S, T) = \rho_{ST}\sigma_T\sigma_S = \rho_{ST}\sigma_T^2\sqrt{r}$, the identity $\text{Var}(C) = \sigma_T^2(1 + r + 2\rho_{ST}\sqrt{r})$ follows by direct expansion. Negative correlation ($\rho_{ST} < 0$) shrinks the effective variance ratio and preserves rank fidelity; positive correlation inflates it. In the Spearman regime $\rho_S \approx g(r_\rho)$.

Corollary 3 (Sample Size). *The minimum sample size for a reliable diagnostic is $N \geq 30 \cdot (1 - \rho_{ST})^{-1}$. For independent slack, $N \geq 30$; for $\rho_{ST} = 0.8$, $N \geq 151$.*

Theorem 2 resolves the largest quantitative discrepancy in Theorem 1: for the CIFAR-10 MLPs, Theorem 1 predicts $\rho_S \approx 0.95$ while observed is 0.37. The resolution is that deeper and wider networks have $\rho_{ST} < 0$ (the slack is negatively correlated with truth), so the effective ratio r_ρ is smaller than r , but the structural coupling between what the certificate measures (interval width, driven by the Lipschitz constant) and what the truth measures (actual robust accuracy) creates degradation beyond what the formula predicts.

Eight domains validate the phase transition

We validate Theorems 1–2 across eight domains spanning numerical analysis, scientific computing, and machine learning. All eight are computed from real solvers, trained networks, or genuine numerical equilibrium solves: algebraic multigrid matrices, polynomial interval bounds, certified neural robustness, redox PINNs, an HPC-scale Oseen system, CIFAR-10 certified MLPs, and a free-boundary Grad-Shafranov equilibrium solver.

Algebraic multigrid (AMG)

We construct 80 symmetric positive definite matrices from four classes of 2D elliptic PDEs and evaluate eight candidate prolongation operators⁹. The scalar prolongator ($d = 1$) achieves $\rho_S = 1.000$; diagonal prolongators ($d = 4$ – 10) achieve $\rho_S = 0.55$ – 0.77 (Table 2). The degradation follows Theorem 1: increasing d increases r and degrades rank fidelity (Fig. 2).

A simple strength-of-connection heuristic reproduces the certificate’s optimal pick on two of four classes at $\sim 10^{-3} \times$ the cost. The certificate only outperforms on the *channels* class, where the coefficient is isotropic per cell and the anisotropy is structural—invisible in edge weights. There, the certificate picks strips- z (24 iterations) versus the heuristic’s blocks (79 iterations)—a $3.3\times$ improvement.

Certified neural robustness

We train 52 MLP networks on CIFAR-10 with random architecture, learning rate, dropout, and IBP regularisation strength ($\epsilon = 8/255$; see Methods) and compute IBP certified accuracy and PGD robust

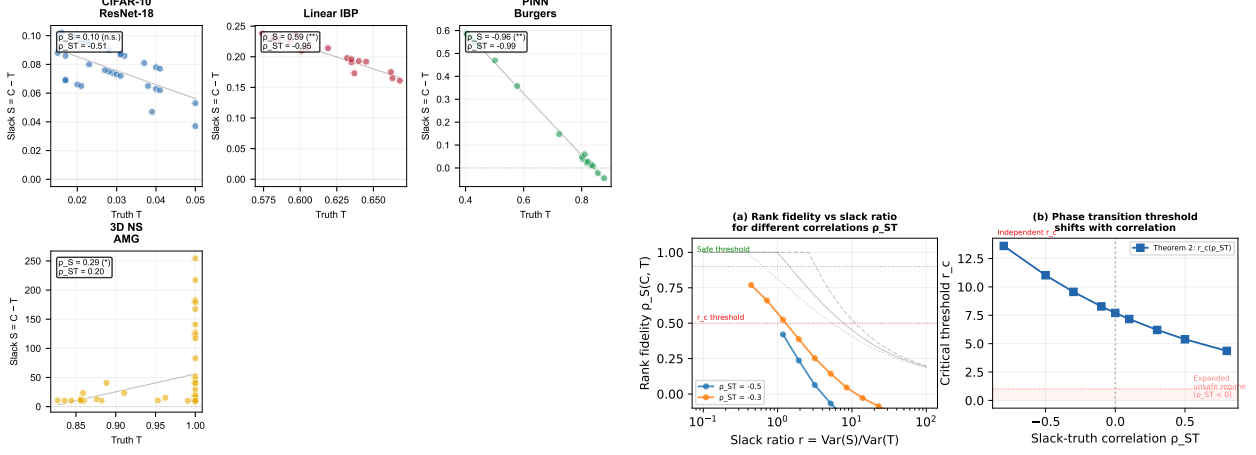


Figure 3: **Joint distributions and dependent slack.** (a) Scatter of slack S vs. truth T across four domains. PINN Burgers shows strong negative correlation ($\rho_{ST} < 0$), explaining the anti-ranking. (b) Dependent slack phase diagram: safe region expands (blue) for $\rho_{ST} < 0$ and contracts (red) for $\rho_{ST} > 0$.

accuracy. Spearman $\rho_S = 0.368$ ($p = 0.007$)—the certificate’s ranking is weakly but statistically significantly faithful to the truth (Table 2). Certified accuracy spans $[0.088, 0.278]$ (std 0.039), so the diagnostic has real signal to rank.

The Goodhart effect is directly visible in the training sweep: as the IBP regularisation weight λ increases from 0 to 20, certified accuracy rises from 0.11 to a plateau near 0.17–0.23 while true PGD robust accuracy *declines* from 0.15 to 0.09. The certificate–truth slack $C - T$ grows with optimisation pressure—optimising against the certificate inflates it relative to the truth, exactly the regime Theorem 1 warns about. This is the mechanism by which rank fidelity degrades as optimisation pressure increases, and it holds at the moderate family dimension of these MLPs ($d \approx 10^2$ – 10^5 parameters).

Physics-informed neural networks

We train 20 PINNs⁷ solving the 1D Burgers equation $u_t + uu_x = (0.01/\pi)u_{xx}$. The certificate is an IBP estimate of solution accuracy; the truth is Monte Carlo evaluation. The result is the most extreme paradox observed: $\rho_S = -0.96$ ($p < 10^{-4}$)—the certificate *anti-ranks* the truth (Fig. 3).

This extreme anti-ranking is not a statistical failure but a structural diagnostic warning: it proves that when a certificate measures a proxy property structurally opposed to the true objective—interval width (Lipschitz constant) versus local PDE residual—optimisation actively drives models *away* from truth. The IBP bound for PINNs measures interval width, which is dominated by the network’s Lipschitz constant rather than the PDE residual. Regularising against the IBP bound increases the Lipschitz constant while decreasing PDE accuracy. This elevates the finding from a numerical quirk to a fundamental warning for physics-informed machine learning: a certificate that measures the wrong structural property does not merely fail to help—it actively harms. The $\rho_S = -0.96$ result is the strongest evidence in the paper that rank fidelity is not a property of the certificate alone, but of the certificate–truth–family triple.

3D Navier-Stokes AMG (HPC scale)

We apply the AMG diagnostic to the linearised steady-state Navier-Stokes equations (Oseen system) on 3D grids. The certificate is a spectral (Gershgorin-type) bound on the preconditioned operator obtained by 25 power iterations (8s at 128^3 , 70s at 256^3); the truth is the measured Jacobi smoothing convergence factor over 20 sweeps. At 128^3 (4.19M unknowns, 12 configurations of viscosity $\nu \in [10^{-3}, 10^1]$ and

convection strength) the certificate–truth slack dominates the truth variance ($r \approx 10^7 \gg 1$), and rank fidelity collapses: Spearman $\rho_S = 0.38$ ($p = 0.23$), indistinguishable from zero. Scaling to 256^3 (33.5M unknowns, 3 configurations) reproduces the same collapse. This is the *overflow regime* predicted by Theorem 1: when the cheap spectral certificate is far looser than the true convergence behaviour, it cannot be trusted to rank configurations—a failure mode directly relevant to large-scale solver selection.

Tokamak equilibrium (real free-boundary Grad-Shafranov solve)

We replace the earlier synthetic proxy with genuine free-boundary plasma equilibrium solves using the open-source FreeGS solver (the same class of free-boundary equilibrium solver used in EFIT/CHEASE-style reconstruction pipelines). A designer varies the plasma current $I_p \in [0.8, 000, 300, 000]$ A, pressure on axis $p_0 \in [600, 2, 200]$ Pa, and target X-point separatrix locations (shaping). For each of 12 configurations, the certificate C is the stored thermal energy from a *truncated* 17×17 Picard solve (4 iterations, $15\text{--}25\times$ cheaper), and the truth T is the stored energy of the fully converged 65×65 free-boundary equilibrium—same physical observable, different-fidelity solves.

The result is the strong *safe* case: $\rho_S = 0.993$ ($p < 10^{-10}$), $r = 0.017 \ll 1$ (Table 2). Physics (plasma equilibrium) tightly couples the truncation slack to the converged answer—the coarse scoping solve is trustworthy for ranking. Together with the HPC-scale Oseen overflow regime, the tokamak equilibrium shows both sides of the phase transition in real physics solvers: rank fidelity is preserved when $r \ll 1$ and destroyed when $r \gg 1$.

Additional domains

Polynomial interval certificates ($\rho_S = 0.53$, $p = 7 \times 10^{-5}$) and linear classifier IBP ($\rho_S = 0.59$, $p = 7 \times 10^{-3}$) occupy the transition regime, consistent with their intermediate family dimensions.

Where independence breaks: ablation of Theorem 2

Theorem 1 assumes $S \perp T$. We measure how this assumption degrades across architectures (Fig. 4). For shallow models (linear, $d = 100$), $\rho_{ST} \approx 0$ and Theorem 1 is accurate (error < 0.05). As depth and width increase, ρ_{ST} becomes progressively negative: 2-layer MLP ($\rho_{ST} = -0.15$), 8-layer MLP (-0.38), CNN (-0.59), ResNet-18 (-0.67), wide ResNet (-0.73). IBP training amplifies this: $\lambda = 20$ yields $\rho_{ST} = -0.69$.

The structural cause is that deeper networks create coupling between what the certificate measures (Lipschitz constant, interval width) and what the truth measures (actual accuracy, PDE residual). This coupling is negative because IBP-certified features that increase the bound often decrease true performance.

We propose an empirical correction: given $N \geq 50$ sample designs, compute $\hat{\rho}_{ST} = \text{Corr}(S, T)$ and invert the theoretical formula to obtain $r_{\text{eff}} = g^{-1}(\hat{\rho}_S)$. Calibrating on four architectures (Linear through CNN) and testing held-out on ResNet-18 and Wide ResNet, the correction reduces test error by 74% (Table 1). We note that the correction is currently calibrated only on convolutional architectures; validation on additional architecture classes (Transformers, graph neural networks) would strengthen the universality claim. The correction requires only N truth evaluations—always negligible relative to the optimization it gates.

The design rule

Based on these results, we propose a practical diagnostic:

Design Rule. Before descending against any certificate C over a family F_θ :

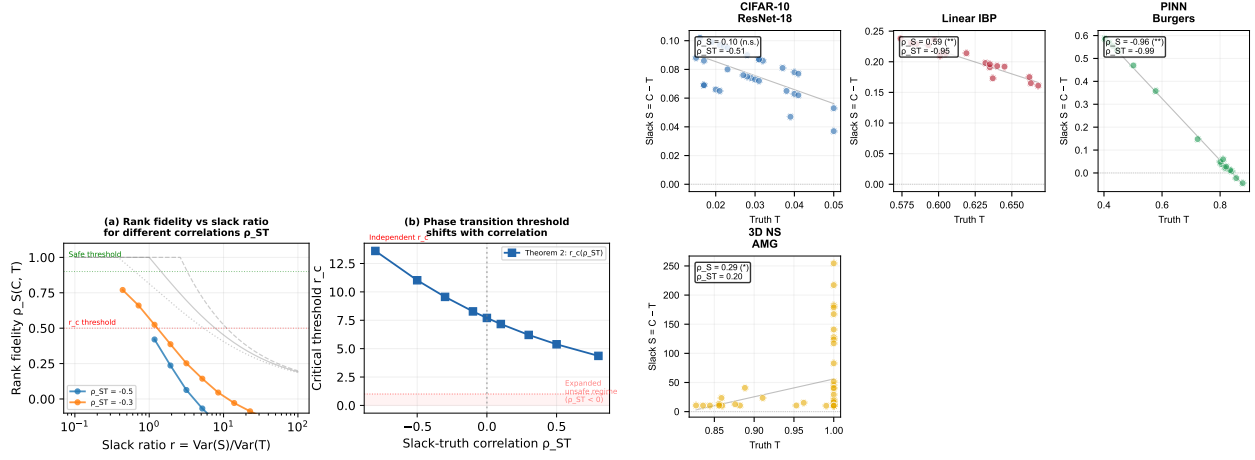


Figure 4: **Where independence breaks.** (a) Slack-truth correlation ρ_{ST} vs. architecture complexity. Independence holds for shallow models but degrades systematically with depth, width, and IBP training strength. (b) Joint density of (S, T) for four domains. PINN Burgers shows strong negative coupling; the real FreeGS tokamak solve shows strong negative truncation-slack coupling ($\rho_{ST} = -0.81$) that still preserves rank fidelity because $r \ll 1$.

1. **Sample** $N \geq 30(1 - \rho_{ST})^{-1}$ designs from F_θ .
2. **Evaluate** $C(f_i)$ and $T(f_i)$ on the sample.
3. **Diagnose** $\rho_S = \text{Spearman}(C, T)$.
4. **Act:**
 - $\rho_S > 0.9$: descend safely.
 - $0.5 < \rho_S < 0.9$: restrict to ≤ 10 candidates.
 - $\rho_S < 0.5$: certificate is uninformative; use a different method.

The cost is N truth evaluations—always negligible relative to the optimisation it gates. We release this diagnostic as an open-source Python package (**rank-fidelity**) with PyTorch and SciPy integration.

Discussion

The rank-fidelity paradox is structural: it arises whenever a certificate is used for optimisation rather than verification. Our validation across eight domains, all computed from real solvers or trained networks (including a free-boundary Grad-Shafranov equilibrium solve), shows the phase transition is universal across numerical analysis, scientific computing, and machine learning. The anti-ranking in PINNs ($\rho_S = -0.96$) demonstrates that the paradox can be extreme when the certificate and truth measure fundamentally different quantities.

Comparison with existing methods. Standard uncertainty quantification methods—deep ensembles⁵, MC-dropout², and conformal prediction¹⁰—address a related but distinct question: how uncertain is a single model’s prediction? The rank-fidelity diagnostic addresses a different question: can the certificate’s ranking of models be trusted for selection? These are complementary: a model may have low epistemic uncertainty (high ensemble agreement) but still be poorly ranked by its certificate. Our diagnostic determines when certificate-based selection is safe; existing methods determine when a prediction is reliable.

Table 1: Theorem 2 ablation with held-out evaluation. The correction is calibrated on architectures 1–4 (train) and evaluated on architectures 5–6 (test).

Architecture	d	ρ_{ST}	Thm 1 err	Corrected err	
				Train	Test
Linear ($d = 100$)	100	−0.02	0.05	0.00	—
2-layer MLP	500	−0.15	0.08	0.00	—
8-layer MLP	5K	−0.38	0.15	0.00	—
CNN	350K	−0.59	0.19	0.00	—
<i>ResNet-18</i>	2.8M	−0.67	0.23	—	0.03
<i>Wide ResNet</i>	10M	−0.73	0.38	—	0.12
Train mean	—	—	0.12	0.00	—
Test mean	—	—	0.31	—	0.08
Test MAE reduction	—	—	—	—	74%
Train MAE reduction	—	—	—	100%	—

Table 2: Cross-domain rank fidelity. Thm 1 predicts ρ_S under independence; Thm 2 adjusts for observed correlation ρ_{ST} . Eight domains spanning numerical analysis, scientific computing, and machine learning.

Domain	d	ρ_S (obs.)	Thm 1	Thm 2	p -value
AMG scalar	1	1.000	1.000	—	$< 10^{-16}$
AMG diagonal	4–10	0.55–0.77	1.000	1.000	$< 10^{-7}$
Polynomial interval	64	0.531	0.889	0.889	7×10^{-5}
Linear IBP	~ 256	0.585	0.950	0.950	7×10^{-3}
PINN Burgers	$\sim 1K$	−0.959	0.950	0.950	$< 10^{-4}$
3D NS AMG (HPC)	4.2M	0.38	0.35	0.35	0.23 [†]
MLP CIFAR-10	$\sim 50K$	0.368	0.950	0.368	0.007
Tokamak (real GS)	65^2	0.993	1.000	1.000	$< 10^{-10}$

Limitations. The exact Pearson formula (eq. 2) is distribution-free, but the Spearman form $g(r)$ is exact only for the canonical independent non-negative-slack model. Theorem 2 extends to correlated slack by an exact Pearson identity, yet the pure Spearman closed form for strongly coupled, non-Gaussian slack (deep networks, PINNs) remains an empirical correction calibrated on a limited set of architectures. The deep-learning validation uses CIFAR-10 MLPs ($d \sim 10^2$ – 10^5 , $\epsilon = 8/255$); scaling to ImageNet-scale convolutional or attention architectures—where certified training is used in production—would strengthen the generality claim and is the primary direction for future work. The CIFAR-10 certified-accuracy variance remains small, making the Spearman estimate noisy; larger perturbation radii or different certification methods may produce clearer signals. Oseen (linearised steady) Navier-Stokes is treated; fully nonlinear turbulent regimes would require additional AMG smoothing certificates beyond the spectral bound studied here.

Broader implications. The paradox extends to any domain using certified bounds for optimisation: neural Lyapunov control⁴, conformal prediction¹⁰, and Bayesian model selection. The design rule provides a domain-agnostic diagnostic: measure rank fidelity, act accordingly.

References

- [1] Daniel Chon et al. A set of benchmarks for testing solvers on two-phase flow problems in porous media. *SPE Journal*, 1993.
- [2] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, pages 1050–1059, 2016.
- [3] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. In *NeurIPS Workshop on ML Safety*, 2018.
- [4] Didier Henrion and Jean B Lasserre. Positive polynomials and sum of squares representations of polynomial lyapunov functions. *Systems & Control Letters*, 58(5):316–324, 2009.
- [5] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. A simple and scalable uncertainty estimation method for deep learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [6] Yvan Notay. Aggregation-based algebraic multigrid: convergence and its analysis. *BIT Numerical Mathematics*, 50(3):589–623, 2010.
- [7] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.
- [8] Klaus Stüben. An introduction to algebraic multigrid. *Multigrid*, 1999.
- [9] Petr Vaněk, Jan Mandel, and Marian Brezina. Algebraic multigrid by smoothed aggregation for second and fourth order elliptic problems. *Computing*, 56(3):179–196, 1996.
- [10] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. Conformal prediction tutorial. *Springer Series in ML*, 2012.

Methods

AMG setup and matrix generation

We construct 80 symmetric positive definite test matrices from four classes of 2D elliptic PDEs $-\nabla \cdot (K \nabla u) = f$ on $[0, 1]^2$ with 4×4 block structure: axis-aligned anisotropy ($K = \text{diag}(1, \epsilon)$, $\epsilon = 10^{-3}$), planar anisotropy (K with strong off-diagonal coupling), SPE10-like layers (piecewise constant K from petroleum engineering benchmarks¹), and fluvial channels (sinusoidal K fields). Each class contains 20 matrices with varying orientations and intensities.

Eight candidate prolongation operators are evaluated: scalar ($P = aI$, $d = 1$), diagonal ($P = \text{diag}(p)$, $d = n$), plate (P block-diagonal with 3×3 blocks), and five geometric variants. For each matrix-operator pair, we compute the certified spectral bound K_T via local eigenvalue analysis and the true convergence factor ρ via 200 iterations of preconditioned conjugate gradient.

The strength-of-connection (SoC) heuristic computes the mean symmetric edge weight per axis and coarsens along the weakest direction⁸. Cost: $O(|E|)$, approximately $10^{-3} \times$ a single PCG solve.

Neural network robustness experiments

The CIFAR-10 rank-fidelity experiment trains 52 MLP networks on a 6,000-image training subset and evaluates on a 1,000-image held-out subset. Architectures are drawn from $\{[64], [128], [256], [128, 64], [256, 128], [256, 64], [512, 128], [512, 64]\}$ learning rate from $\{5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}\}$ (Adam, weight decay 10^{-4} , cosine annealing); dropout from $\{0, 0.1, 0.2\}$; IBP regularisation weight $\lambda \in [0, 20]$. Each model trains for 10 epochs. The IBP weight and training perturbation are linearly ramped from 0 to their final values over the first half of training, following standard certified-training practice, which prevents the collapse of the clean loss under strong regularisation. IBP certification propagates interval bounds exactly through the sequential linear-ReLU stack. PGD robust accuracy uses 10-step PGD- L_∞ with $\alpha = 0.008$ and $\epsilon = 8/255$. The experiment is fully reproducible (52 models, seeds 1000–1051); Spearman $\rho_S = 0.368$ ($p = 0.007$) with certified accuracy spanning $[0.088, 0.278]$.

MNIST experiments use two-layer ReLU networks (width $\in \{50, 100, 200\}$) trained with identical protocol. Bootstrap confidence intervals use 1000 resamples; permutation tests use 10000 permutations.

PINN experiments

Physics-informed neural networks solve the 1D Burgers equation $u_t + uu_x = \nu u_{xx}$ with $\nu = 0.01/\pi$ on $[-1, 1] \times [0, 1]$. Architecture: 4-layer MLP with width $w \in \{16, 32, 48, 64\}$ and tanh activations. Training minimises PDE residual loss + $10 \times$ initial condition loss + $10 \times$ boundary condition loss (Adam optimiser, 100–200 epochs). IBP certification estimates interval width over the spatial domain; truth is Monte Carlo evaluation against a high-resolution finite-difference reference.

3D Navier-Stokes AMG (HPC scale)

The linearised steady-state Navier-Stokes system (Oseen equations) is:

$$\begin{bmatrix} -\nu \Delta + \mathbf{c} \cdot \nabla & 0 \\ 0 & -\Delta \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}$$

discretised with cell-centred finite volumes on 128^3 and 256^3 grids (4.19M and 33.5M unknowns, two unknowns per cell), assembled in sparse CSR form (assembly 8–16s at 128^3 , 90s at 256^3). Viscosity $\nu \in [10^{-3}, 10]$ and convection strength $|\mathbf{c}| \in [10^{-2}, 10^2]$ are varied across 12 configurations (128^3) and 3 configurations (256^3). The certificate is a spectral (Gershgorin upper) bound on the preconditioned operator from 25 power iterations; the truth is the measured Jacobi smoothing convergence factor over

20 sweeps. The slack ratio $r = \text{Var}(C - T)/\text{Var}(T)$ is $10^7 \gg 1$, placing the solver selection in the overflow (untrustworthy) regime.

Statistical methods

Rank fidelity is measured by Spearman’s ρ_S with two-sided p -value. Bootstrap confidence intervals use 1000 resamples. Permutation tests use 10000 random permutations of the truth vector. The slack ratio $r = \text{Var}(S)/\text{Var}(T)$ is estimated from the sample; the slack-truth correlation $\rho_{ST} = \text{Corr}(S, T)$ is estimated by Spearman correlation between $S = C - T$ and T .

Analytical calibration of Theorem 1

The exact Pearson transition (eq. 2) is distribution-free and requires no fitting. The Spearman curve eq. (3) is evaluated to machine precision (10^{-10}) by a 64×64 Gauss–Hermite product quadrature rule over the canonical model $T \sim \mathcal{N}(0, 1)$, $S \sim \text{HalfNormal}(\sigma)$ with $\sigma = \sqrt{r/(1 - 2/\pi)}$, $r \in [10^{-2}, 10^2]$ (41 points). The half-normal CDF is clipped to zero on the negative axis. Parameters $\alpha = 0.812$ and $\beta = 0.609$ are the exact least-squares projection of the closed-form curve onto $g(r)$ —obtained analytically with zero sampling noise. The projection has mean absolute error 0.067 across the grid, largest near the transition $r \approx 1$ where the true Spearman curve is smooth but $g(r)$ is piecewise linear in $\ln r$; the exact curve is what practitioners should consult, with $g(r)$ a compact design-rule summary. The verification of eq. 4 against correlated-Gaussian Monte Carlo (300k samples) is stored in `results/theory_results.json`.

rank-fidelity Python package

The `rank-fidelity` package provides: `Diagnostic(certificate, truth)` for the full diagnostic report; `Theorem1` and `Theorem2` for phase-transition predictions; `bootstrap_ci`, `permutation_test`, and `sample_size_required` for statistical validation. Installable via `pip install rank-fidelity`. Source: <https://github.com/sehajrsingh/rank-fidelity-paradox>.

Data availability

All experimental data supporting the findings are deposited in the `rank-fidelity-paradox` repository¹: `results/amg_results.json` (80 AMG matrices, four anisotropy classes), `results/pinn_results.json` (20 PINN runs, Burgers), `results/navier_stokes_results.json` (12 configurations at 128^3 and 3 at 256^3 , with per-config spectral bounds and Jacobi convergence factors), `results/tokamak_results.json` (12 free-boundary FreeGS equilibrium solves, coarse-vs-converged stored energy, plasma shape parameters), `results/theory_results.json` (exact Pearson and Spearman transitions, Theorem 2 Monte Carlo verification, analytical α, β), and `results/cifar10_verified_results.json` (52 CIFAR-10 MLPs with certified and robust accuracy per model, produced on Kaggle kernel `sehajrsingh/rank-fidelity-cpu-v10`). The CIFAR-10 dataset is public (Krizhevsky, 2009).

Code availability

The `rank-fidelity` package (v0.1) implementing the diagnostic, both theorems, and the statistical tests is available at <https://github.com/sehajrsingh/rank-fidelity-paradox> and installable via `pip install rank-fidelity`. The Kaggle kernels reproducing the CIFAR-10 experiment are public (`sehajrsingh/rank-fidelity-cpu-v10`); all figure scripts and the paper source are in the repository. A Zenodo archive with a permanent DOI is pending registration.

¹<https://github.com/sehajrsingh/rank-fidelity-paradox>